Consensus RNA secondary structure prediction by ranking *k*-length stems

Denise Y. F. Mak[1], **Gary Benson**[2]
[1]Graduate Program in Bioinformatics, Boston University, Boston, MA 02215 USA
[2]Dept. of Computer Science and Dept. of Biology, Boston University, Boston, MA 02215 USA

Abstract— *The accurate computational prediction of RNA secondary structures is a difficult task, but an important one, since RNA structure is usually more evolutionarily conserved than primary sequence. We describe a dynamic programming algorithm called FoldRRS (***Fold***ing of* **R***NA by* **R***anking of* **S***tems) that predicts a consensus secondary structure from a multiple sequence alignment. Our algorithm exploits the use of k-length stems ($k = 2$) to acquire base pairing probability and covariation information from individual sequences. We test sequences from the BRAliBase I data set [1] and the Rfam database [2]. Our results were compared against three algorithms, RNAalifold, Pfold, and KNetFold, that are similar in nature. FoldRRS exhibits an increase in accuracy over the other programs in data sets which contain longer and/or more numerous sequences.*

Keywords: RNA, secondary structure, *k*-length stems

1. Introduction

The functional role of RNA has expanded rapidly to include regulatory cellular functions such as regulating transcription, gene silencing and genome maintenance. The 2007 ENCODE pilot project discovered, surprisingly, that the entire human genome may yield higher than expected RNA transcription as many novel non-protein coding transcripts were identified [3]. Functional RNA families generally have a common secondary structure as the function of these molecules is inherently tied to secondary structure. It is known that the secondary structure is often more evolutionarily conserved than primary sequence. In fact, the primary sequences of RNA families usually do not have very high sequence identity which means that folding algorithms that use only sequence information are likely to perform poorly.

Some of the earliest single sequence RNA structure prediction algorithms use dynamic programming to maximize base pairs [4] or minimize structural energies [5]. Another solution generates a partition function to calculate the probabilities of every possible base pairing in the sequence [6].

The single sequence approach is generally not as accurate as comparative approaches because information can be gathered from multiple sequences that would otherwise be missing from a single sequence. A comparative approach gathers information from nucleotide base pairing interactions which typically form canonical Watson-Crick base pairs (AU, UA, GC, GC) or a wobble base pair (GU, UG). The tremendous evolutionary pressures selecting for structural elements strongly suggest that primary sequences contain covariation information. These include consistent and compensatory mutations which change the nucleotide base pair while still preserving the existing structure. A consistent mutation occurs when a single position changes producing a valid base pair combination (GU → GC) and a compensatory mutation occurs when both positions are mutated but maintain a valid base pair (GU → UA). There are a variety of ways to score covariation information, Lindgreen *et al.* [7] showed that the best covariation measure was RNAalifold's covariation score [8] and concluded that combining that with McCaskill's base pairing probability matrix score [6] could be a desirable approach.

One of the common comparative approaches starts with an alignment and retrieves information from the alignment to build a consensus secondary structure. It has been shown that with sequence alignments, for sequences with medium to high percent identity ($> 60\%$) [9], the sequences are diverse enough to provide covariation information and the comparative approach performs well. Prevalent algorithms following this method include RNAalifold [8], Pfold [10] and KNetFold [11]. RNAalifold extends Nussinov's dynamic programming algorithm [4] to include thermodynamic rules and a covariation measure. Pfold uses a stochastic context-free grammar (SCFG) to produce a prior probability distribution of RNA structures [10]. KNetFold uses a machine learning algorithm to analyze k-nearest neighbor classifiers to predict a base pair [11]. KNetFold also has the ability to predict pseudoknot structures.

In this paper, we describe a method to predict a consensus secondary structure from multiple sequences. Our method requires a good multiple sequence alignment. We exploit information about hairpin forming stems (also known as stacking region) which contain at least k base pairs [5]. We combine two pieces of information from k-length stems ($k = 2$) 1) probabilities of base pairings in these stems, and 2) information on consistent and compensatory mutations (mutations in two positions that preserve Watson-Crick base pairing) observed in the stems. The remainder of this paper is organized as follows. In section 2 we describe the method.

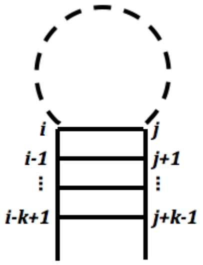

Fig. 1: An (i, j, k) stem: the base pairs $(i, j), (i-1, j+1), \cdots, (i-k+1, j+k-1)$ forming a k-length stem.

The data used to test the algorithm are detailed in section 3 and we highlight the results in section 4. Finally, we discuss the results in section 5.

2. Method

2.1 FoldRRS Algorithm

The following steps outline the algorithm:

1) Calculate multiple sequence alignment
2) Generate base pairing probability matrices for each individual sequence
3) Adjust the matrices with gaps according to the multiple alignment
4) Scan each matrix for k-length stems with each entry having probability ≥ 0.001
5) Identify and score *common* stems
6) Sort the *common* stems by two-part process
7) Remove common stems with zero covariation score that have a stem score below a minimum threshold m
8) Combine common stems into larger stems
9) Use DP algorithm to fill structure matrix, S and energy matrix, E
10) Entry in E with the lowest minimal free energy represents the consensus secondary structure, stored in S

The first step to the algorithm yields a multiple sequence alignment. We use ClustalW [12] which was chosen for its ease of use and popularity, although any other multiple sequence alignment program could be used. A base pairing probability matrix for each sequence is generated using McCaskill's algorithm [6] implemented in the Vienna RNA package [13]. Gaps are added to each matrix corresponding to gaps in the multiple sequence alignment. Steps 1-3 are similar to [14]. The differences begin with the remaining steps and are described below.

We scan each base pairing probability matrix for k-length stems in which each individual entry has probability ≥ 0.001 as it is unlikely for structural base pairs to have lower probabilities [14]. A k-length stem is defined as k base pairs at positions $(i, j), (i-1, j+1), \cdots, (i-k+1, j+k-1)$ as illustrated in figure 1. We use the label (i, j, k) to describe a k-length stem with the closing base pair occurring between positions i and j. In what follows, we use $k = 2$. This value represents the minimum stem length.

We define a stem to be *common* if it was found in at least 2 of the sequences. Each common stem has a stem score, $T_{i,j,k}$, which is the sum of its probability and covariation scores:

$$T_{i,j,k} = P_{i,j,k} + C_{i,j,k}$$

Let x be the number of sequences, $b^A_{i,j}$ be the base pair probability for positions i and j in sequence A, and $V^A_{i,j}$ be a valid stem indicator which equals 1 if all the base pairs in a k-length stem, closing at i and j, have probability ≥ 0.001, and equals 0 otherwise. The probability score, $P_{i,j,k}$, is the average of the base pairing probabilities in the k-length stem summed over all the sequences where it was found. $P_{i,j,k}$ is zero if the stem is never found.

$$P_{i,j,k} = \sum_{A=1}^{x} \left(\frac{\sum_{s=0}^{k-1} b^A_{i-s,j+s}}{k} \right) \cdot V^A_{i,j}$$

The covariation score is calculated using RNAalifold's [8] covariation scoring method which gives weights for consistent and compensatory mutations. Let BP be the set of allowed base pairs: $BP = \{$AU, UA, CG, GC, GU, UG$\}$. The allowed base pairs are described in matrix Π:

$$\Pi^A_{i,j} = \begin{cases} 0 & \text{if base pair } (A[i], A[j]) \notin BP \\ 1 & \text{if base pair } (A[i], A[j]) \in BP \end{cases}$$

Invalid base pairs are described in matrix I which is used to measure a penalty score:

$$I^A_{i,j} = \begin{cases} 1 & \text{if base pair } (A[i], A[j]) \notin BP \\ 0 & \text{if base pair } (A[i], A[j]) \in BP \end{cases}$$

Let $\delta(A[i], A[j], B[i], B[j])$ be the Hamming distance between two base pairs in sequence A and B at positions i and j:

$$\delta = \begin{cases} 0 & \text{if the two base pairs are identical} \\ 0.5 & \text{if the base pairs differ in exactly} \\ & \text{one position (consistent mutation)} \\ 1 & \text{if the base pairs differ in both} \\ & \text{positions (compensatory mutation)} \end{cases}$$

BRAliBase I data					
Sequences	Mean Length	Mean Similarity		Num. Sequences	
		High	Med	High	Med
S. cerevisiae tRNA-PHE	73	84.4	60.0	11	11
E. coli RNase P	377	81.5	67.1	9	11
E. coli SSU rRNA	1542	90.7	80.0	11	11

Table 1: BRAliBase I data characteristics. The data set is labeled after the reference sequence.

Rfam data			
Sequences	Mean Length	Mean Similarity	Num. Sequences
U8 snRNA (RF00096)	111	63.7	6
Lysine riboswitch (RF00168)	179	60.6	47
HCV IRES region (RF00061)	243	73.9	79

Table 2: Rfam data characteristics. The data set is labeled after the Rfam family.

The covariation score is the addition of the number of mutations between all sequence pairs minus the number of invalid base pairs at positions i and j.

$$C_{i,j,k} = \sum_{s=0}^{k-1} \left(\frac{1}{\binom{x}{2}} \sum_{A<B} \delta(A[i-s], A[j+s], B[i-s], B[j+s]) \cdot \Pi^{A}_{i-s,j+s} \Pi^{B}_{i-s,j+s} - \frac{1}{x} \sum_{A=1}^{x} I^{A}_{i-s,j+s} \right)$$

The common stems are ranked in a two-step process. They are first separated into two groups, one having non-zero covariation scores and the other having covariation scores of zero. Each group is sorted from highest to lowest stem score. A minimum stem score, m, for common stems with a zero covariation score is set to

$$m = 0.25 \times \text{total number of sequences.}$$

The value of 0.25 was chosen empirically as it allows high stem scores with no covariation information to be included. The two groups are joined into a ranked list of common stems with the stems having non-zero covariation scores added first.

We combine common stems that have overlapping base pairs and stem scores that are adjacently ranked since no other common stem is ranked between the two. The joining of common stems into a larger stem can potentially provide stronger evidence for base pairing because the probability and covariation scores are recalculated for the longer stem. Longer stems are energetically favored.

The list of n common stems contains the base pairs that are most likely to be paired. We use a dynamic programming method to build the consensus secondary structure from this information.

2.2 Dynamic Programming

The dynamic programming algorithm step builds the consensus secondary structure by minimizing the minimum free energy (MFE) of the structure. The structure matrix, S, and the corresponding energy matrix, E, have n rows and n columns for the n *common* stems. The ranking of the common stems is important as it determines the order in which they are added to form the consensus secondary structure. Each common stem is added to the previously defined structure and the energy of the newly formed structure is re-calculated.

The following rules apply when deciding which base pairs to add to a structure.

1) Pseudoknots are discarded. Two base pairs (i, j) and (k, l) cannot have $i < k < j < l$.
2) A nucleotide can only be base paired with one other nucleotide.
3) The loop distance between bases must be at least 3 nucleotides long.
4) A higher ranked stem wins when there is a conflict as described by the first three rules.

The entry S_{ij} represents the j^{th} ranked common stem being added in i^{th} order where $i \leq j$. A common stem cannot be added higher than its ranking. For example, the second ranked common stem can *only* be added first (the first ranked stem was discarded) or second (after the addition of the first ranked common stem).

The initial values of matrix S are the secondary structures of adding a single common stem j and stored in S_{1j}. The structure energies are stored respectively in the energy matrix, E_{1j}. The DP step of the algorithm is

$$E_{ij} = \text{minimum energy}_{d=i-1}^{j-1} \left\{ S_{(i-1)d} + stem_j \right\}$$

where $stem_j$ is common stem j and the minimum energy E_{ij} is an average energy over all sequences for structure S_{ij}. The entries in matrix S follow the same path as matrix E. The lowest entry in E_{ij} represents the consensus secondary structure stored in S_{ij}.

Data	Alignment	Num. of common stems	
	Length	before combining	after combining
S. cerevisiae tRNA-PHE (H)	73	41	31
S. cerevisiae tRNA-PHE (M)	75	18	13
E. coli RNaseP (H)	385	168	127
E. coli RNaseP (M)	458	145	125
E. coli SSU rRNA (H)	1554	562	495
E. coli SSU rRNA (M)	1604	595	547
U8 snRNA (RF00096)	146	56	45
Lysine riboswitch (RF00168)	274	50	48
IRES region of HCV (RF00061)	413	43	36

Table 3: Number of common stems before and after combining adjacently ranked overlapping stems.

3. Data

3.1 BRAliBase I sequences

We use three of the four BRAliBase I data sets [1] to measure the accuracy of our program (table 1). We did not test *E. coli* LSU rRNA sequences as RNAalifold and Pfold's webserver limits were reached. These sequences have been established as a benchmark for RNA secondary structure prediction algorithms. The data comprise a diverse group of sequences and each contains two multiple sequence alignments generated by ClustalW [12] representing high and medium sequence identity. The high similarity group has a mean pairwise sequence identity of $80 - 90\%$ and the medium similarity group has a mean pairwise sequence identity of $60 - 80\%$. The first sequence in the alignment is labeled the reference sequence, B_1, and an experimentally validated structure, S_1, is also provided. The other sequences in the alignment, $B_2 \cdots B_n$, belong to the same family and exhibit structural similarities which our program aims to identify.

RNA folding programs take sequences and predict a consensus structure from them. But, in order to measure the accuracy of the prediciton, a reference consensus structure is needed. However, when a consensus structure is unavailable, a method called *consensus reconstruction* is used [1] to create one from the structure of the reference sequence.

3.2 Rfam sequences

We also test sequences taken from the Rfam database release 9.0 [2] (table 2). We use the Rfam seed alignments instead of the full alignments because seed alignments were hand-curated and contain known representative members of the families. The full alignments contain many more sequences added through computational measures.

3.3 HCV IRES Region

We further investigate the IRES region from the Hepatitis C virus taken from the Los Alamos hepatitis C sequence database [15]. We retrieved 194 full IRES region sequences (1-342 bases because the last pseudoknot within the IRES region between the coding start site is ignored) and after removing ambiguous sequences, 173 remain. The sequences have an alignment length of 356, mean length of 342 and mean pairwise sequence identity of 94.8%. An experimentally confirmed secondary structure [16] is used to evaluate our program.

4. Results

4.1 Time Complexity

The time complexity of FoldRRS beginning at step 4 of the algorithm is $O(xl^2)$ for the scanning of k-length stems for each of x sequences of alignment length l and assuming that k is held constant. The dynamic programming step takes time $O(xtn^2)$ where n is the number of common stems, and t is the time required to calculate the average energy of a structure. From table 3, n is approximately equal to $l/2$ for small l (< 100) and for larger l, n is approximately equal to $l/3$. The time complexity of ClustalW is $O(x^4 + l^2)$ and $O(xl^3)$ for generating individual base pairing matrices. Overall, the time complexity of FoldRRS is $O(xtl^2 + x^4 + xl^3)$.

4.2 Prediction Measure

In order to measure the accuracy of our prediction, we need to be able to compare the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). We use the sensitivity (X), selectivity (Y) and Matthews Correlation Coefficient (*MCC*) measurement values as defined in [1] to compare our results. The sensitivity and selectivity values are defined as:

$$sensitivity = \frac{TP}{TP + FN} \quad selectivity = \frac{TP}{TP + (FP - \xi)}$$

The false positive term is divided into three categories, *inconsistent, contradiction* or *compatible*. *Inconsistent* base pairs conflict with a base pair in the reference structure. *Contradicting* base pairs cause non-nested (pseudoknot) situations with respect to the reference struture. *Compatible* base pairs are those that are not present in the reference structure but produce no conflict and the ξ term allows these base pairs to be removed from the false positive count.

E. coli RNaseP $l = 377$ $n = 11$(M) or 9(H)					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold (H)	79	112	61 (77.2)	16 (79.2)	0.781
RNAalifold (M)	102	93	69 (67.6)	22 (75.8)	0.715
Pfold (H)	79	64	50 (63.3)	8 (86.2)	0.738
Pfold (M)	102	104	92 (90.2)	5 (94.8)	0.925
KNetFold (H)	79	91	55 (69.6)	25 (68.8)	0.691
KNetFold (M)	102	104	77 (75.5)	23 (77.0)	0.761
FoldRRS (H)	**79**	**93**	**56 (70.9)**	**9 (86.2)**	**0.781**
FoldRRS (M)	**102**	**92**	**88 (86.3)**	**0 (100.0)**	**0.929**

Table 4: Results for *E. coli* RNase P. l is mean sequence length, n is number of sequences. H and M indicate, respectively, high and medium similarity sets

E. coli SSU rRNA $l = 1542$ $n = 11$					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold (H)	453	483	290 (64.0)	177 (62.1)	0.630
RNAalifold (M)	444	452	387 (87.2)	36 (91.5)	0.893
KNetFold (H)	453	492	292 (64.5)	174 (62.7)	0.635
KNetFold (M)	444	464	328 (73.9)	103 (76.1)	0.750
FoldRRS (H)	**453**	**369**	**286 (63.1)**	**64 (81.7)**	**0.718**
FoldRRS (M)	**444**	**373**	**325 (73.2)**	**29 (91.8)**	**0.820**

Table 5: Results for *E. coli* SSU rRNA. Pfold could not be tested as the maximum sequence length is 500 on the webserver.

U8 snRNA (RF00096) $l = 111$ $n = 6$					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold	38	23	12 (31.6)	7 (63.2)	0.442
Pfold	38	20	16 (42.1)	4 (80.0)	0.577
KNetFold	38	46	25 (65.8)	14 (64.1)	0.645
FoldRRS	**38**	**42**	**27 (71.1)**	**11 (71.1)**	**0.707**

Table 6: Results for U8 snRNA.

Lysine riboswitch (RF00168) $l = 179$ $n = 47$					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold	53	38	38 (71.7)	0 (100.0)	0.846
KNetFold	53	53	52 (98.1)	0 (100.0)	0.990
FoldRRS	**53**	**44**	**44 (83.0)**	**0 (100.0)**	**0.910**

Table 7: Results for Lysine riboswitch. Pfold could not be tested as the maximum number of sequences allowed is 40 on the webserver.

The Matthews Correlation Coefficient combines both sensitivity and selectivity and is defined as:

$$MCC = \frac{TP \times TN - (FP - \xi) \times FN}{\sqrt{(TP + FP - \xi)(TP + FN)(TN + FP - \xi)(TN + FN)}}$$

The BRAliBase webpage contains perl scripts to compute these values.

4.3 Limitations

Of the three programs to which we compare our results, the publicly available Pfold webserver has a limit of 40 sequences and an alignment length limit of 500. Because of this, we could not test Pfold with *E. coli* SSU rRNA and with the IRES region of HCV.

We did not change RNAalifold, Pfold or KNetFold's default program parameter settings although this could have potentially improved performance.

5. Discussion

Results are shown in tables 4–9. These tables give the number of true and false positives found by each programs as well as the Matthews Correlation Coefficient which combines these into a common measure. Because in many cases, one program does better on true positives and worse on false positives, we focus on the MCC measure.

For the short *S. Cerevisiae* tRNA sequences, all programs find the complete structures and so these results will not be discussed further.

IRES region of HCV (RF00061) $l = 243$ $n = 79$					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold	77	28	26 (33.8)	0 (100.0)	0.580
KNetFold	77	55	39 (50.6)	14 (73.6)	0.609
FoldRRS	**77**	**35**	**33 (42.9)**	**0 (100.0)**	**0.653**

Table 8: Results for HCV's IRES region. Pfold could not be tested as the maximum number of sequences allowed is 40 on the webserver.

IRES region of HCV from Los Alamos HCV sequence database $l = 342$ $n = 173$					
Algorithm	BPs in Ref.	BPs in Subj.	TPs (sens.)	FPs (select.)	MCC
RNAalifold	99	95	51 (51.5)	40 (56.0)	0.535
KNetFold	99	102	48 (48.5)	49 (49.5)	0.488
FoldRRS	**99**	**86**	**51 (51.5)**	**30 (63.0)**	**0.568**

Table 9: Results for the HCV IRES region taken from the Los Alamos HCV sequence database. Pfold could not be tested as the maximum number of sequences allowed is 40 on the webserver.

For the remaining 8 sets of sequences, FoldRRS has the best MCC score in 6 sets. These include the *E. coli* RNaseP medium and high similarity sets (Table 4), the *E. coli* SSU rRNA high similarity set (Table 5), the U8 snRNA set (Table 6), the Lysine riboswitch set (Table 7), the HCV IRES set from Rfam (Table 8), and the HCV IRES set from Los Alamos (Table 9).

In particular, in the SSU rRNA high similarity set (Table 5), one of the sets with the longest sequence, FoldRRS yields nearly the same number of true positive base pairs as the other programs but picks only 64 false positives which is *one-third* the number picked by the other programs. This tendency for FoldRRS to be more conservative in picking false positives is also apparent in the other long sequence sets (Tables 4, 5 – medium similarity, 8, 9). Two of these sets also contain the highest number of sequences and would be most likely to harbor the most covariation information. These set characteristics suggest that our filter for common stems and our stem rankings are effective with longer and/or more numerous sequences.

For the remaining 2 sets of sequences, FoldRRS has the second best MCC score, being outperformed by RNAalifold on the *E. coli* SSU rRNA medium similarity set (Table 5), and by KNetFold on the Lysine riboswitch set (Table 7).

The number of common stems found by FoldRRS with $k = 2$ are highlighted in table 3. There was a slight decrease in the number of common stems after combining them, which indicates common stems having stem lengths ≥ 3. This would seem to indicate that an initial scan of varying stem lengths might improve the prediction as our algorithm only combines adjacently ranked common stems with stem length of 2.

6. Conclusion

We have developed a competitive new secondary structure prediction algorithm that selects likely k-length stems which are vital components of RNA structure. We identify these stems by combining covariation information provided by a sequence alignment with base pairing probabilities. We show in a collection of 8 data sets which vary in length and number of sequences and degree of pairwise sequence similarity, that FoldRRS outperforms other similar RNA structure prediction programs (in 6 sets) or comes in second (in 2 sets) and that the program does consistently well in data sets which contain longer and/or more numerous sequences.

7. Acknowledgments

We wish to thank Paul Gardner for help on using the scripts and sequences and Andrea Branch for insightful discussions.

References

[1] P. P. Gardner and R. Giegerich, “A comprehensive comparison of comparative rna structure prediction approaches,” *BMC Bioinformatics*, vol. 5, p. 140, Sept. 2004, pMID: 15458580.

[2] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, “Rfam: annotating non-coding rnas in complete genomes,” *Nucl. Acids Res.*, vol. 33, pp. D121–124, 2005. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/33/suppl_1/D121

[3] The ENCODE Project Consortium, “Identification and analysis of functional elements in 1% of the human genome by the encode pilot project,” *Nature*, vol. 447, pp. 799–816, June 2007. [Online]. Available: http://dx.doi.org/10.1038/nature05874

[4] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, “Algorithms for loop matchings,” *SIAM Journal on Applied Mathematics*, vol. 35, pp. 68–82, July 1978. [Online]. Available: http://www.jstor.org/stable/2101031

[5] M. Zuker and P. Stiegler, “Optimal computer folding of large rna sequences using thermodynamics and auxiliary information,” *Nucleic Acids Research*, vol. 9, pp. 133–48, 1981, pMID: 6163133.

[6] J. S. McCaskill, “The equilibrium partition function and base pair binding probabilities for rna secondary structure,” *Biopolymers*, vol. 29, pp. 1105–19, 1990, pMID: 1695107.

[7] S. Lindgreen, P. P. Gardner, and A. Krogh, “Measuring covariation in rna alignments: physical realism improves information measures,” *Bioinformatics*, p. btl514, Oct. 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl514v1

[8] I. L. Hofacker, M. Fekete, and P. F. Stadler, “Secondary structure prediction for aligned rna sequences,” *Journal of Molecular Biology*, vol. 319, pp. 1059–66, June 2002, pMID: 12079347.

[9] P. P. Gardner, A. Wilm, and S. Washietl, “A benchmark of multiple sequence alignment programs upon structural rnas,” *Nucl. Acids Res.*, vol. 33, pp. 2433–2439, Apr. 2005. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/33/8/2433

[10] B. Knudsen and J. Hein, “Pfold: Rna secondary structure prediction using stochastic context-free grammars,” *Nucl. Acids Res.*, vol. 31, pp. 3423–3428, July 2003. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/31/13/3423

[11] E. Bindewald and B. A. Shapiro, “Rna secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers,” *RNA*, vol. 12, pp. 342–352, Mar. 2006. [Online]. Available: http://rnajournal.cshlp.org/cgi/content/abstract/12/3/342

[12] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins, “Clustal w and clustal x version 2.0,” *Bioinformatics*, vol. 23, pp. 2947–2948, Nov. 2007. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/21/2947

[13] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, “Fast folding and comparison of rna secondary structures,” *Monatshefte für Chemie / Chemical Monthly*, vol. 125, pp. 167–188, Feb. 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00818163

[14] I. L. Hofacker and P. F. Stadler, “Automatic detection of conserved base pairing patterns in rna virus genomes,” *Computers & Chemistry*, vol. 23, pp. 401–14, June 1999, pMID: 10404627.

[15] A. Agrawal, J. Szinger, R. Funkhouser, and B. Korber, “Los alamos hepatitis c immunology database.” *Applied Bioinformatics*, Jan 2005. [Online]. Available: http://bioinformatics.adisonline.com/pt/re/abi/abstract.00822942-200504040-00002.htm

[16] M. Honda, M. R. Beard, L. H. Ping, and S. M. Lemon, “A phylogenetically conserved stem-loop structure at the 5' border of the internal ribosome entry site of hepatitis c virus is required for cap-independent viral translation,” *J Virol*, vol. 73, no. 2, pp. 1165–74, Feb 1999.