Maximum expected accurate structural neighbors of an RNA secondary structure

Feng Lou
Laboratoire pour la Recherche en Informatique (LRI)
Université Paris-Sud XI
Orsay cedex, France
lou@lri.fr

Peter Clote
Biology Department
Boston College
Chestnut Hill, MA 02467, USA
clote@bc.edu

***Abstract*—Since RNA molecules regulate genes and control alternative splicing by *allostery*, that is, by switching between two distinct secondary structures, it is important to develop algorithms to predict RNA *conformational switches*. It has recently emerged that RNA secondary structure can be more accurately predicted by computing the *maximum expected accurate* (MEA) structure, rather than the *minimum free energy* (MFE) structure. The MEA structure $\mathcal{S}$ has maximum score $2\sum_{(i,j)\in\mathcal{S}} p_{i,j} + \sum_i$ unpaired q_i, where first sum is taken over all base pairs (i,j) belonging to $\mathcal{S}$, and the second sum is taken over all unpaired positions in $\mathcal{S}$, and where $p_{i,j}$ [resp. q_i] is the probability that i,j are paired [resp. i is unpaired] in the ensemble of low energy structures.**
Results: Given an arbitrary RNA secondary structure $\mathcal{S}_0$ for an RNA nucleotide sequence $a = a_1, \ldots, a_n$, we say that another secondary structure $\mathcal{S}$ of a is a k-neighbor of $\mathcal{S}_0$, if the base pair distance between $\mathcal{S}_0$ and $\mathcal{S}$ is k. Here we describe the algorithm RNAborMEA, which for an arbitrary initial structure $\mathcal{S}_0$ and for all values $0 \leq k \leq n$, computes the secondary structure $MEA(k)$, having *maximum expected accuracy* over all k-neighbors of $\mathcal{S}_0$. We apply our algorithm to the class of *purine riboswitches*.
Availability: Source code for RNAborMEA can be downloaded from http://bioinformatics.bc.edu/clotelab/RNAborMEA/.

***Keywords*-RNA secondary structure, maximum expected accurate structure, minimum free energy structure, riboswitch, conformational switch**

I. Introduction

RNA secondary structure conformational switches play an essential role in a number of biological processes, such as regulation of viral replication [3] and of viroid replication [4], regulation of R1 plasmid copy number in *E. coli* by *hok/sok* system [9], transcriptional and translational gene regulation in prokaryotes by riboswitches [15], regulation of alternative splicing in eukaryotes [7], and stress-responsive gene regulation in humans [20], etc. Due to their biological importance, several groups have developed algorithms that attempt to recognize conformational switches – in particular, *riboswitches* [1], [5], [2]. Most current approaches heavily depend on detecting the *aptamer*, located in the $5'$-portion of the riboswitch, that is responsible for high affinity binding of a particular ligand ($K_D \approx 5$ nM) that triggers the conformational change [16]. Computational tools that rely on *stochastic context free grammars*, such as Infernal and CMFinder, have been trained to recognize riboswitch aptamers; in particular, Infernal was used to create the Rfam database [12], which includes 14 families of riboswitch aptamers.

Upon ligand binding, the $3'$-portion of a riboswitch, called *expression platform*, undergoes a conformation change, forming a stem-loop that aborts transcription (thus effecting *transcriptional regulation*, as in guanine riboswitches) or that sequesters the Shine-Dalgarno sequence (thus effecting *translational regulation*, as in thiamine pyrophosphate (TPP) riboswitches). Due to evolutionary pressure for accurate ligand recognition, there is generally high sequence identity in the aptamer region; however, there is low sequence identity (data not shown) in the *expression platform*. Since current riboswitch detection algorithms do not attempt to predict the location of the expression platform, we have developed a tool, RNAborMEA, that yields information concerning alternative structures of a given RNA sequence. This tool can suggest the presence of a conformational switch; however, much more work must be done to actually produce a riboswitch gene finder, part of the difficulty due to the fact that riboswitch aptamers contain *pseudoknots* that cannot be captured by secondary structure.

In previous work [10], [11], we described a novel program RNAbor to predict RNA conformational switches. For a given secondary structure $\mathcal{S}$ of a given RNA sequence $\mathbf{s}$, the secondary structure $\mathcal{T}$ of $\mathbf{s}$ is said to be a k-neighbor of $\mathcal{S}$, if the base pair distance between $\mathcal{S}$ and $\mathcal{T}$ is k. (Base pair distance is defined later.) Given an arbitrary initial structure $\mathcal{S}_0$, for all values $0 \leq k \leq n$, the program RNAbor [10], computes the secondary structure $MFE(k)$, having minimum free energy over all k-neighbors of $\mathcal{S}_0$. In this paper, we extend our work by computing for all values $0 \leq k \leq n$, the secondary structure $MEA(k)$, having maximum expected accuracy (MEA) over all k-neighbors of $\mathcal{S}_0$.

In [8], Do et al. introduced[1] the notion of *maximum expected accurate* (MEA) secondary structure, determined as follows: *(i)* compute base pairing probabilities $p(i,j)$ using a trained stochastic context free grammar; *(ii)* compute probabilities $q(i) = 1 - \sum_{i<j} p(i,j) - \sum_{j<i} p(j,i)$ that position i does not pair; *(iii)* using a dynamic programming algorithm similar to that of Nussinov and Jacobson [19], determine that secondary structure $\mathcal{S}$ having maximum score $\sum_{(i,j)\in\mathcal{S}} 2\alpha \cdot p(i,j) + \sum_{i \text{unpaired}} \beta q_i$, where the first sum is over paired positions (i,j) of $\mathcal{S}$ and the second sum is over positions i located in loop regions of $\mathcal{S}$, and where $\alpha, \beta > 0$ are parameters with default values 1. Subsequently Kiryu et al. [13] computed the MEA structure by replacing the stochastic context free grammar computation of base pairs in *(i)* by using McCaskill's algorithm [17], which computes the Boltzmann base pairing probabilities

$$p(i,j) = \frac{\sum_{(i,j)\in\mathcal{S}} exp(-E(S)/RT)}{\sum_{\mathcal{S}} exp(-E(S)/RT)} \quad (1)$$

Here $E(S)$ is the free energy of secondary structure S, with respect to the Turner energy model [22], R is the universal gas constant, and T is absolute temperature. Thus $p(i,j)$ is the sum of the Boltzmann factors of all secondary structures that contain the fixed base pair (i,j), divided by the partition function, which latter is the sum of Boltzmann factors of all secondary structures. In fact, Kiryu et al. [13] describe an algorithm to compute the MEA structure common to all RNAs in a given alignment. Later, Lu et al. [14] rediscovered Kiryu's method; in addition, Lu et al. computed suboptimal MEA structures by implementing an analogue [23].

In this paper, we extend the MEA technique to compute the maximum expected accurate *k-neighbor* of a given RNA secondary structure $\mathcal{S}_0$; i.e. that secondary structure which has maximum expected accuracy over all structures that differ from $\mathcal{S}_0$ by exactly k base pairs.

II. Preliminaries

Recall the definition of RNA secondary structure.

Definition 1: A secondary structure S on RNA sequence $s_1, \ldots, s_n$ is defined to be a set of ordered pairs (i,j), such that $1 \le i < j \le n$ and the following are satisfied.

1) *Watson-Crick or GU wobble pairs:* If (i,j) belongs to S, then pair (a_i, a_j) must be one of the following canonical basepairs: (A,U), (U,A), (G,C), (C,G), (G,U), (U,G).
2) *Threshold requirement:* If (i,j) belongs to S, then $j - i > \theta$, where θ, generally taken to be equal to 3, is

[1]Miyazawa [18] first introduced the concept of *maximum expected accuracy* in the context of sequence alignment of two amino acid sequences $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$. Miyazawa computed the Boltzmann pair probability $P(a_i, b_j)$ that a_i is aligned with b_j, for all $1 \le i \le n$ and $1 \le j \le m$, and then used $P(a_i, b_j)$ as the similarity score between a_i and b_j in the usual Needleman-Wunsch and Smith-Waterman algorithms. Do et al. lifted this method to the context of RNA secondary structure prediction.

```
1. void RNAborMEA(s, S_0, M)
2.   //M(i,j,k) is the score of MEA k-neighbor of S_0
3.   initialize M(i,j,k) = 0 for all 1 ≤ i,j ≤ n, 0 ≤ k ≤ n
4.   compute p_{i,j} for all 1 ≤ i ≤ j ≤ n (McCaskill's algorithm)
5.   for i = 1 to n
6.     q_i = 1 − Σ_{j≥i} p_{i,j} − Σ_{j<i} p_{j,i}
7.     //q_i is Boltzmann probability that i is unpaired
8.   for d = 0 to n − 1    // d is diagonal offset value
9.     for i = 1 to n − d
10.      j = i + d
11.      for k = 0 to n
12.        if j − i ≤ θ //θ unpaired bases in hairpin
13.          if k == 0
14.            M(i,j,k) = Σ_{r=i}^{j} β q_r
15.          else // k > 0
16.            break // for all k > 0 M(i,j,k) = 0
17.        else if j − i == θ + 1
18.          if (i,j) ∈ S_0 then
19.            M(i,j,0) = 2α p_{i,j} + Σ_{r=i+1}^{j−1} β q_r
20.            M(i,j,1) = Σ_{r=i}^{j} β q_r
21.            break //for k>1, M(i,j,k) = 0
22.          else // (i,j) ∉ S_0
23.            M(i,j,0) = Σ_{r=i}^{j} β q_r
24.            if basePair(i,j) then
25.              M(i,j,1) = 2α p_{i,j} + Σ_{r=i+1}^{j−1} β q_r
26.            break //for other cases M(i,j,k) = 0
27.        else // j − i > θ + 1
```

Figure 3. Initial portion of pseudocode for RNAborMEA algorithm, which continues in Figure 4. Given RNA sequence $\mathbf{s} = s_1, \ldots, s_n$ of length n, initial secondary structure S_0 of $\mathbf{s}$, RNAborMEA computes for all values of $0 \le k \le n$ that structure S with base pair distance k from S_0, which maximizes the value $M(i,j,k) = \sum_{(i,j)\in S} 2\alpha p_{i,j} + \sum_i$ unpaired in $_S\, \beta q_i$. The pseudocode actually computes only values $M(i,j,k)$ for all i,j,k; the MEA structures are obtained by backtracing. This algorithm clearly runs in $O(n^4)$ time with $O(n^3)$ space.

```
27.        else // j − i > θ + 1
28.          max = 0 // M(i,j,k) = max of following
29.          // Case 1: j unpaired in S[i,j]
30.          b_0 = d_BP(S_0[i,j−1], S_0[i,j])
31.          //b_0 = 1 if j paired in S_0[i,j], else 0
32.          val = M(i,j−1,k−b_0) + β q_j
33.          if val > max then
34.            max = val
35.            index = (0,0,0)
36.            //backtracking: j unpaired
37.          // Case 2: (i,j) ∈ S
38.          if basePair(i,j) //check if i,j can pair
39.            b_1 = d_BP(S_0[i+1,j−1] ∪ {(i,j)}, S_0[i,j])
40.            val = M(i+1,j−1,k−b_1) + 2α p_{i,j}
41.            if val > max then
42.              max = val
43.              index = (i, k−b_1, 0)
44.              //backtracking: (i,j) ∈ S
45.          // Case 3: (r,j) ∈ S for some i < r < j
46.          for r = i+1 to j − θ − 1
47.            if basePair(r,j)
48.              b_2 = d_BP(S_0[i,r−1] ∪ S_0[r+1,j−1] ∪ {(r,j)}, S_0[i,j])
49.              for k_0 = 0 to k − b_2
50.                k_1 = k − b_2 − k_0 //k_0 + k_1 + b_2 = k
51.                val = M(i,r−1,k_0) + M(r+1,j−1,k_1) + 2α p_{r,j}
52.                if val > max then
53.                  max = val
54.                  index = (r, k_0, k_1)
55.                  //backtracking: (r,j) ∈ S
56.          M(i,j,k) = max
57.          M(j,i,k) = index
```

Figure 4. Pseudocode for RNAborMEA algorithm. Given RNA sequence $\mathbf{s} = s_1, \ldots, s_n$ of length n, initial secondary structure S_0 of $\mathbf{s}$, RNAborMEA computes for all values of $0 \le k \le n$ that structure S with base pair distance k from S_0, which maximizes the value $M(i,j,k) = \sum_{(i,j)\in S} 2\alpha p_{i,j} + \sum_i$ unpaired in $_S\, \beta q_i$. The pseudocode actually computes only values $M(i,j,k)$ for all i,j,k; the MEA structures are obtained by backtracing. This algorithm clearly runs in $O(n^4)$ time with $O(n^3)$ space.

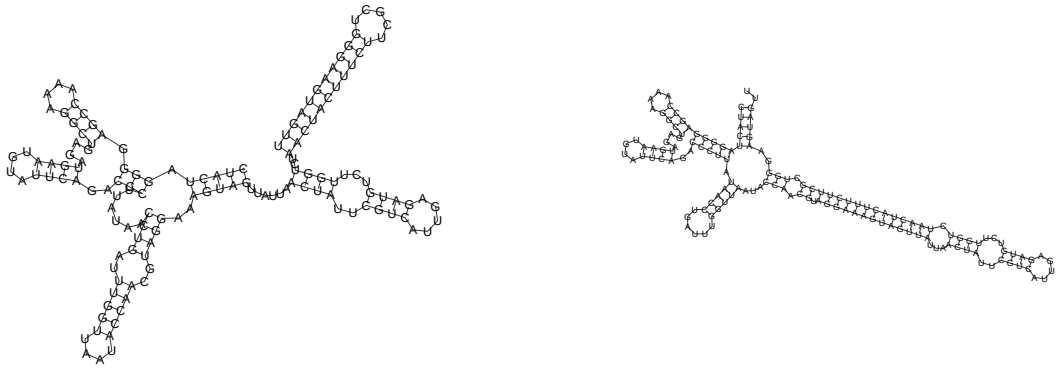


Figure 1. Sample outputs from RNAborMEA on a TPP-riboswitch, AF269819/1811-1669. We took the TPP riboswitch aptamer from the Rfam database [12], then extracted right-flanking nucleotides from the corresponding EMBL file. Displayed from left to right are the structures $MEA(0)$ and $MEA(61)$ (the structure $MEA(52)$ is similar to that of $MEA(61)$ and corresponds to a free energy local minimum in the left figure.) The structure $MEA(61)$ had the highest MEA score over all structural neighbors, including the original structure $\mathcal{S}_0 = MEA(0)$, and had free energy, -46.0 kcal/mol, that was equal to that of the initial structure $\mathcal{S}_0 = MEA(0)$, which is the minimum free energy structure for the given sequence.

the minimum number of unpaired bases in a hairpin loop; i.e. there must be at least θ unpaired bases in a hairpin loop.

3) *Nonexistence of pseudoknots:* If (i, j) and (k, ℓ) belong to S, then it is not the case that $i < k < j < \ell$.
4) *No base triples:* If (i, j) and (i, k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$.

The preceding definition provides for an inductive construction of the set of all secondary structures for a given RNA sequence $a_1, \ldots, a_n$. For all values of $d = 0, \ldots, n$ and all values of $i = 1, \ldots, n - d$, the collection $\mathbb{S}_{i,i+d}$ of all secondary structures for $a_i, \ldots, a_{i+d}$ is defined as follows. If $0 \leq d \leq \theta$, then $\mathbb{S}_{i,i+d} = \{\emptyset\}$; i.e. the only secondary structure for $a_i, \ldots, a_{i+d}$ is the empty structure containing no base pairs (due to the requirement that all hairpins contain at least θ unpaired bases). If $d > \theta$ and $\mathbb{S}_{i,j}$ has been defined by recursion for all $i \leq j < i + d$, then

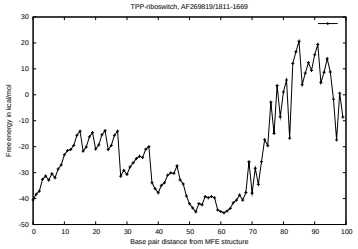
- Any secondary structure of $a_i, \ldots, a_{i+d-1}$ is a secondary structure for $a_i, \ldots, a_{i+d}$, in which a_{i+d} is unpaired.
- If a_i, a_j can form a Watson-Crick or wobble base pair, then for any secondary structure $\mathcal{S}$ for $a_{i+1}, \ldots, a_{i+d-1}$, the structure $\mathcal{S} \cup \{(i, j)\}$ is a secondary structure for $a_i, \ldots, a_{i+d}$.
- For any intermediate value $i + 1 \leq r \leq j - \theta - 1$, if a_r, a_j can form a Watson-Crick or wobble base pair, then for any secondary structure $\mathcal{S}$ for $a_i, \ldots, a_{r-1}$ and any secondary structure $\mathcal{T}$ for $a_{r+1}, \ldots, a_{j-1}$, the structure $\mathcal{S} \cup \mathcal{T} \cup \{(r, j)\}$ is a secondary structure for $a_i, \ldots, a_{i+d}$.

Given two secondary structures $\mathcal{S}, \mathcal{T}$, we define the *base pair distance* between $\mathcal{S}, \mathcal{T}$, denoted by $d_{BP}(\mathcal{S}, \mathcal{T})$, to be the cardinality of the symmetric difference of $\mathcal{S}, \mathcal{T}$; i.e. $d_{BP}(\mathcal{S}, \mathcal{T}) = |(\mathcal{S} - \mathcal{T}) \cup (\mathcal{T} - \mathcal{S})|$.

III. Algorithm description

Given an RNA sequence $a = a_1, \ldots, a_n$, a secondary structure $\mathcal{S}_0$ of a, and a maximum desired value $Kmax \leq n$, the RNAborMEA algorithm computes, for each $1 \leq i < j \leq n$ and each $0 \leq k \leq Kmax \leq n$, the maximum *score* $M(i, j, k)$

$$\sum_{(i,j)\in\mathcal{S}} 2\alpha p_{i,j} + \sum_{i \text{ unpaired}} \beta q_i$$



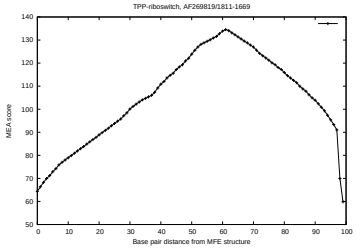


Figure 2. *(Left)* Free energy for all $MEA(k)$ structural neighbors, $0 \leq k \leq 99$, of the TPP-riboswitch, AF269819/1811-1669, described in the previous figure. Clearly, $MEA(0)$ and $MEA(61)$ have the least energy, -46.0 kcal/mol, and $MEA(61)$ has the largest MEA score, 134.555, of all secondary structures for the given RNA sequence. *(Right)* MEA score for all $MEA(k)$ structural neighbors, $0 \leq k \leq 99$, of the TPP-riboswitch, AF269819/1811-1669, described in the previous figure. Clearly, $MEA(61)$ has the largest MEA score, 134.555, of all secondary structures for the given RNA sequence.

where the first sum is taken over all base pairs (i, j) belonging to $\mathcal{S}$, the second sum is taken over all unpaired positions in $\mathcal{S}$, and where $p_{i,j}$ [resp. q_i] is the probability that i, j are paired [resp. i is unpaired] in the ensemble of low energy structures, and $\alpha, \beta > 0$ are weights. Our computational experiments, as in Figure 2, were carried out with default values of 1 for α, β. (See Equation 1 for the formal definition of Boltzmann base pairing probability $p_{i,j}$.)

The dynamic programming computation of $M(i, j, k)$ is performed by recursion on increasing values of $j - i$ for all values $1 \leq i \leq j \leq n$ and $0 \leq k \leq Kmax$. The value of $M(i, j, k)$, stored in the upper triangular portion of matrix M, will involve taking the maximum over three cases, which correspond to the inductive construction of all secondary structures on $a_i, \ldots, a_j$, as described in the previous section. At the same time, the value $M(j, i, k)$, stored in the lower triangular portion of matrix M, will consist of a triple r, k_0, k_1 of numbers, such that the following *approximately*[2] holds. *(i)* If $r = 0$ then $M(i, j, k)$ is maximized by a k-neighbor $\mathcal{S}$ of $\mathcal{S}_0[i, j]$ for the subsequence $a_i, \ldots, a_j$ in which a_j is unpaired. In this case, $k_0 = k$ and $k_1 = 0$. *(ii)* If $r = i$, then $M(i, j, k)$ is maximized by a k-neighbor $\mathcal{S}$ of $\mathcal{S}_0[i, j]$ for the subsequence $a_i, \ldots, a_j$ in which base pair $(i, j) \in \mathcal{S}$. In this case, $k_0 = 0$ and $k_1 = k - 1$. *(i)* If $i < r \leq j - \theta - 1$ then $M(i, j, k)$ is maximized by a k-neighbor $\mathcal{S}$ of $\mathcal{S}_0[i, j]$ for the subsequence $a_i, \ldots, a_j$ in which base pair $(r, j) \in \mathcal{S}$. The left portion of $\mathcal{S}$, which is $\mathcal{S}[i, r - 1]$ will be a k_0 neighbor of $\mathcal{S}[i, r - 1]$, while the right portion of $\mathcal{S}$, which is $\mathcal{S}[r, j]$ must contain the base pair (r, j) and itself be a k_1 neighbor of $\mathcal{S}[r, j]$. In summary, the values r, k_0, k_1 will be used in computing the traceback, where the maximum expected accurate structure that is a k-neighbor of $\mathcal{S}[i, j]$ will be constructed by one of the following: *(i)* MEA k-neighbor of $\mathcal{S}[i, j - 1]$, in the event that a_j is unpaired in $[i, j]$; *(ii)* MEA $k-1$-neighbor of $\mathcal{S}[i+1, j-1]$, in the event that a_i, a_j form a base pair; *(iii)* MEA k_0-neighbor of $\mathcal{S}[i, r - 1]$ and the MEA k_1-neighbor of $\mathcal{S}[r, j]$, where $k_0 + k_1 = k$, in the event that a_r, a_j form a base pair.

Pseudocode for the algorithm `RNAborMEA` is given in Figures 3 and 4. An array M of size $n \times n \times Kmax$ is requires to store the MEA scores in $M(i, j, k)$ for all subsequences $[i, j]$ and all base pair distances $0 \leq k \leq Kmax$ be-

[2]In this section, we provide the motivating idea. The actual algorithm description, which deviates slightly from the description here, is given in the next section and in Figures 3 and 4.

tween structures $\mathcal{S}[i,j]$ and initially given structure $\mathcal{S}_0[i,j]$. For $1 \leq i \leq j \leq n$ and all $0 \leq k \leq Kmax$, the pseudocode in Figure 4 stores a value of the form (x,y,z) in the lower triangular portion, $M(j,i,k)$, of the array. Here, $x=0$ indicates that the optimal structure on $[i,j]$, i.e. having maximum MEA score over all k-neighbors of $\mathcal{S}_0[i,j]$, is obtained by not pairing j with any nucleotide in $[i,j]$; for values $x>0$, hence $x \in [i, j-\theta-1]$, the optimal k-neighbor of $\mathcal{S}_0[i,j]$ is obtained by pairing x with j. The values y,z correspond to the values k_0,k_1, such that: *(i)* if $x=0$, then the optimal k-neighbor of $\mathcal{S}_0[i,j]$ is obtained by first computing the optimal k_0-neighbor of $\mathcal{S}_0[i,j-1]$, where $k_0=k-b_0$, then leaving j unpaired; *(ii)* if $x=i$, then the optimal k-neighbor of $\mathcal{S}_0[i,j]$ is obtained by first computing the optimal k_1-neighbor of $\mathcal{S}_0[i+1,j-1]$, where $k_1=k-b_1$, then adding the enclosing base pair (i,j); *(iii)* if $x=r \in [i+1, j-\theta-1]$, then the optimal k-neighbor of $\mathcal{S}_0[i,j]$ is obtained by first computing the optimal k_0-neighbor of $\mathcal{S}_0[i,r-1]$ as well as the optimal k_1-neighbor of $\mathcal{S}_0[r+1,j-1]$, then adding the base pair (r,j). This last calculation must be done over all values k_0,k_1 such that $k_0+k_1=k$. Using the values $M(j,i,k)=(x,y,z)$, the traceback can be easily computed by recursion; see Figure 5 for pseudocode of traceback.

In a manner similar[3] to the pseudocode of Figures 3 and 4, we have developed a program to compute the pseudo-partition function values

$$Z_{i,j}^{(k)} = \sum_{\mathcal{S} \text{ on } [i,j], d_{BP}(\mathcal{S}_0,\mathcal{S})=k} \exp(\text{MEA}(\mathcal{S}/RT))$$

We then graphed the Boltzmann probabilities $\frac{Z_{1,n}^{(k)}}{Z_{1,n}}$ as well as the uniform probabilities $\frac{N_{1,n}^{(k)}}{N_{1,n}}$, where $N_{1,n}^{(k)}$ is the number of k-neighbors, and $N_{1,n}$ is the total number of secondary structures. When $RT=n$, which normalizes the MEA score to a maximum of 1, it appears that the Boltzmann distribution is the *same* as the uniform distribution, as illustrated in figures and data that cannot be shown, due to space restrictions.

IV. Results

We extended the `RNAborMEA` program to support *structural* constraints; i.e. where structures are required to contain certain designated base pairs or for certain designated positions to be unpaired. Taking the *B. subtilis* XPT riboswitch, whose GENE ON and GENE OFF structures were experimentally determined by in-line probing [21], we applied `RNAborMEA` to all purine riboswitch aptamers from the Rfam database [12], where additional flanking nucleotides were extracted from the EMBL database. Using the structural alignment program `Gardenia` [6], we determined values k_0,k_1 for the most structurally similar structures $MEA(k_0)$ to the XPT GENE OFF structure, resp. $MEA(k_1)$ to the XPT GENE ON structure. Due to space constraints, we can only show one sample result in Figure 6

Quite to our surprise, there appears to be little to no correlation between the structures $MFE(k)$ output by `RNAbor` [10] and the structures $MEA(k)$ output by our current program `RNAborMEA`. Thus our current program provides a different manner of probing increasingly distant structural neighbors of a given RNA structure.

[3]Essentially, one replaces the operation of taking the *maximum* by the a summation, and one replaces the MEA score by the pseudo-Boltzmann factor $\exp(MEA(\mathcal{S})/RT)$.

Acknowledgements

Research of P. Clote and F. Lou was funded by the Digiteo Foundation, in the form of a Digiteo Chair of Excellence to P. Clote. Additionally, P. Clote was partially by the National Science Foundation under grants DMS-0817971 and DBI-0543506. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. Abreu-Goodger and E. Merino. RibEx: A web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res*, 33:W690–W692, 2005.

[2] T.-H. Chang and H.-D. Huang and L.-C. Wu and C.-T. Yeh and B.-J. Liu and J.-T. Horng. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA*, 15(7), 2009.

[3] R.C. Olsthoorn and S. Mertens and F.T. Brederode and J.F. Bol. A conformational switch at the $3'$ end of a plant virus RNA regulates viral replication. *EMBO J.*, 18:4856–4864, 1999.

[4] D. Repsilber and S. Wiese and M. Rachen and A.W. Schroder and D. Riesner and G. Steger. Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (−)-stranded RNA by temperature-gradient gel. *RNA*, 5:574–584, 1999.

[5] P. Bengert and T. Dandekar. Riboswitch finder – A tool for identification of riboswitch RNAs. *Nucleic Acids Res*, 32:W154–W159, 2004.

[6] Guillaume Blin, Alain Denise, Serge Dulucq, Claire Herbach, and Hlne Touz. Alignments of RNA structures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010.

[7] M. T. Cheah, A. Wachter, N. Sudarsan, and R. R. Breaker. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(7143):497–500, May 2007.

[8] C. B. Do, M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15(2):330–340, February 2005.

```
 1. void traceback(i, j, k, M, paren)
 2.   //perform traceback on [i,j] for k-neighbors of S_0[i,j]
 3.   if j − i > θ and M(i,j,k) > 0
 4.     (r, k_0, k_1) = M(j,i,k)
 5.     if r > 0 //j pairs with r in [i,j]
 6.       paren[r] = '('
          //note that paren has dummy char '$' at position 0
 7.       paren[j] = ')'
 8.       traceback(r + 1, j − 1, k1, M, paren)
 9.       traceback(i, r − 1, k_0, M, paren)
10.     else //r = 0, so j not paired in [i,j]
11.       traceback(i, j − 1, k_0, M, paren)
12.   return
```

Figure 5. Pseudocode for the traceback computed by our RNAborMEA algorithm.

```
>  X83878/168-267
UUACAAUAUAAUAGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUG
((((((...........(((((((((.....(((((.......)))))).........(((((.......))))))..)))))))).....(
11     .................(((((((...(.(((((.......))))).).........(((((.......))))))..)))))))).
80     (((...)).).(.((((((...)((((((..(.((((((...)..))))))).)((....))((((((.......))))))).).(((((((
```

Figure 6. Given riboswitch sequence X83878/168-267 with initial structure the minimum free energy structure, the structure $MEA(11)$ is most similar to the XPT GENE ON structure, with Gardenia similarity 155.5, while its similarity to XPT GENE OFF structure is a much lower 66.0. The $MEA(80)$ structure is most similar to the XPT GENE OFF structures, with Gardenia similarity 101.5, while the its similarity to the XPT GENE ON structure is a low 5.0. Maximum expected accurate structural neighbors $MEA(k)$, for $0 \leq k \leq 150$ were computed by RNAborMEA.

[9] Thomas Franch, Alexander P. Gultyaev, and Kenn Gerdes. Programmed cell death by hok/sok of plasmid r1: Processing at the hok mRNA 3H-end triggers structural rearrangements that allow translation and antisense RNA binding. *J. Mol. Biol.*, 273:38–51, 1997.

[10] E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062, August 2007.

[11] E. Freyhult, V. Moulton, and P. Clote. RNAbor: a web server for RNA structural neighbors. *Nucleic. Acids. Res.*, 35(Web):W305–W309, July 2007.

[12] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. Rfam: updates to the RNA families database. *Nucleic. Acids. Res.*, 37(Database):D136–D140, January 2009.

[13] H. Kiryu, T. Kin, and K. Asai. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, 23(4):434–441, February 2007.

[14] Z. J. Lu, J. W. Gloor, and D. H. Mathews. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA.*, 15(10):1805–1813, October 2009.

[15] M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, and R.R. Breaker. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell*, 113(5):577–586, 2003.

[16] M. Mandal and R. R. Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, 11(1):29–35, January 2004.

[17] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[18] S. Miyazawa. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, 8:999–1009, 1994.

[19] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.

[20] P. S. Ray, J. Jia, P. Yao, M. Majumder, M. Hatzoglou, and P. L. Fox. A stress-responsive RNA switch regulates VEGFA expression. *Nature*, 457(7231):915–919, February 2009.

[21] A. Serganov, Y.R. Yuan, O. Pikovskaya, A. Polonskaia, L. Malinina, A.T. Phan, C. Hobartner, R. Micura, R.R. Breaker, and D.J. Patel. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.*, 11(12):1729–1741, 2004.

[22] T. Xia, Jr. J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35, 1999.

[23] M. Zuker. On finding all suboptimal foldings of an rna molecule. *Science*, 244(7):48–52, 1989.