RNA Secondary Structure Prediction via Energy Density Minimization

Can Alkan[1]*, Emre Karakoc[2]*, S. Cenk Sahinalp[2]*, Peter Unrau[3], H. Alexander Ebhardt[3], Kaizhong Zhang[4] and Jeremy Buhler[5]

[1] Department of Genome Sciences, University of Washington, USA
[2] School of Computing Science, Simon Fraser University, Canada
[3] Department of Molecular Biology and Biochemistry, Simon Fraser University, Canada
[4] Department of Computer Science, University of Western Ontario, Canada
[5] Department of Computer Science, Washington University in St Louis, USA

Abstract. There is a resurgence of interest in RNA secondary structure prediction problem (a.k.a. the RNA folding problem) due to the discovery of many new families of non-coding RNAs with a variety of functions. The vast majority of the computational tools for RNA secondary structure prediction are based on free energy minimization. Here the goal is to compute a non-conflicting collection of structural elements such as hairpins, bulges and loops, whose total free energy is as small as possible. Perhaps the most commonly used tool for structure prediction, mfold/RNAfold , is designed to fold a single RNA sequence. More recent methods, such as RNAscf and alifold are developed to improve the prediction quality of this tool by aiming to minimize the free energy of a number of functionally similar RNA sequences simultaneously. Typically, the (stack) prediction quality of the latter approach improves as the number of sequences to be folded and/or the similarity between the sequences increase. If the number of available RNA sequences to be folded is small then the predictive power of multiple sequence folding methods can deteriorate to that of the single sequence folding methods or worse.

In this paper we show that delocalizing the thermodynamic cost of forming an RNA substructure by considering the *energy density* of the substructure can significantly improve on secondary structure prediction via free energy minimization. We describe a new algorithm and a software tool that we call Densityfold , which aims to predict the secondary structure of an RNA sequence by minimizing the sum of energy densities of individual substructures. We show that when only one or a small number of input sequences are available, Densityfold can outperform all available alternatives. It is our hope that this approach will help to better understand the process of nucleation that leads to the formation of biologically relevant RNA substructures.

1 Introduction

Given an RNA sequence, *RNA secondary structure prediction problem* (sometimes referred to as the *RNA folding problem*) asks to compute all pairs of bases that form hydrogen bonds. Although the problem is one of the earliest in computational biology, it has attracted considerable fresh attention due to the recent discoveries of many non-coding RNA molecules such as siRNAs, riboswitches and catalytic RNAs with a variety of novel functions. (See for example, the September 2, 2005 issue of Science magazine,

* These authors contributed equally to the paper.

devoted to “Mapping RNA form and function”, which investigates the relationship between RNA structure and functionality [1].)

Much of the literature on RNA secondary structure prediction is devoted to the *free energy minimization* approach. This general methodology (which is sometimes called the *thermodynamic* approach) aims to compute the secondary structure by minimizing the total free energy of its substructures such as *stems*, *loops* and *bulges*.

Free energy minimization can be applied either to a single RNA sequence, or, simultaneously to a number of functionally similar RNA sequences. Free energy minimization of a single RNA sequence has been studied since early 70s [21] and a number of dynamic programming algorithms have been developed [17, 22, 13]. The popular Mfold and its more efficient version RNAfold (from the Vienna package) are implementations of these algorithms.

Despite a 25 year long effort to perfect secondary structure prediction of a single RNA sequence via energy minimization, the end result is still far from perfect. The limitations of this approach are usually attributed to the following factors. The total free energy is affected by tertiary interactions which are currently poorly understood and thus ignored in the energy tables [15] currently used by all structure prediction tools. There are also external, non-RNA related factors that play important roles during the folding process. Furthermore, the secondary structure of an RNA sequence is formed as the molecule is being transcribed. A highly stable substructure, formed only after a short prefix of the RNA sequence is transcribed, can often be preserved after the completion of the transcription, even though it may not conform to a secondary structure with the minimum free energy. [1]

In order to address these issues, much of the recent research on RNA secondary structure is focused on simultaneously predicting the secondary structure of many functionally similar RNA sequences. The intuition underlying this approach is that functional similarity is usually due to structural similarity, which, in many cases, correspond to sequence similarity. Because this approach can utilize the commonly observed covarying mutations among aligned base pairs in a stem, the accuracy of this approach can outperform single sequence structure prediction approach.

There are two main techniques for simultaneously predicting the secondary structure of multiple sequences via energy minimization.

- The first general technique, used in particular by the alifold program [10] of the Vienna package , assumes that the multiple alignment between the input RNA sequences (in the case of alifold , computed by the Clustal-W program [20]) corresponds to the alignment between their substructures. The structure is then derived by folding the multiple alignment of the sequences. Clearly this method crucially relies on the correctness of the multiple sequence alignment; thus its prediction quality is usually good for highly similar sequences (60% or more) but can be quite poor for more divergent sequences.
- The second general technique aims to compute the sequence alignment and the structure prediction simultaneously [19, 8, 16]. When formulated as a rigorous dynamic programming procedure, the computational complexity of this technique becomes very high; it requires $O(n^6)$ time even for two sequences and is NP-hard for

[1] Another crucial issue that limits the prediction accuracy of many energy minimization based tools is that they do not allow pseudoknots. This is due to the the fact that the energy minimization problem allowing arbitrary pseudoknots is NP-hard [3]. The only software tool we are aware of which allows certain types of pseudoknots (as described by [6]) is Pknots [18], which suffers from efficiency problems. Thus our current implementation does not allow any pseudoknots due to efficiency considerations; however it can easily be extended to allow the class of pseudoknots captured by Pknots .

multiple sequences [7]. In order to decrease the computational complexity, it may be possible to restrict the number of substructures from each RNA sequence to be aligned to the substructures from other sequences. In [5], this is done through a preprocessing step which detects all statistically significant potential stems of each RNA sequence by performing a local alignment between the sequence and its reverse complement. When computing the *consensus structure*, only those substructures from each RNA sequence which are enclosed by such stems are considered for being aligned to each other. This strategy is successfully implemented by the `RNAscf` program recently developed by Bafna et al. [5].

One final approach to multiple sequence structure prediction is the so called *consensus folding* technique. Rather than minimizing free energy, the consensus folding technique first extracts all potential stems of each input RNA sequence. The consensus structure is then computed through determining the largest set of compatible potential stems that are common to a significant majority of the RNA sequences. A good example that uses the consensus folding technique is the `comRNA` program [11] which, once all stems of length at least ℓ are extracted from individual sequences, computes the maximum number of compatible stems [2] that are common to at least k of the sequences via a graph theoretic approach. As one can expect, the consensus technique also relies on the availability of many sequences that are functionally (and hopefully structurally) similar.

1.1 Our approach

As described above, the most common objective in secondary structure prediction is total free energy minimization. In the context of multiple sequence structure prediction, this objective can be used in conjunction with additional criteria such as covariation in mutations on predicted stems etc., yet the effectiveness of such criteria very much depends on (1) the availability of sufficient number of RNA sequences with similar functions, and (2) reasonably high sequence similarity between the input sequences. When these two conditions are not met, single sequence energy minimization methods still provide the most accurate prediction. Furthermore, because multiple sequence folding methods generate consensus structures that involve those substructures found in the majority of the sequences, the stems they return get shorter and thus the number of correct base pairs they predict get worse with increasing number of input sequences.

The goal of this paper is to show that delocalizing the thermodynamic cost of forming an RNA substructure by considering the *energy density* of the substructure can improve on secondary structure prediction via free energy minimization. We describe a new algorithm and a software tool that we call `Densityfold` which aims to predict the secondary structure of an RNA sequence by minimizing the sum of energy densities of individual substructures. We believe that our approach may help understand the process of nucleation that is required to form biologically relevant RNA substructures.

Our starting observation is that potential stems that are most commonly realized in the actual secondary structure are those whose *free energy density* (i.e. length normalized free energy) is the lowest. Figure 1(a) depicts the known secondary structure of the *E.coli* 5S rRNA sequence. This sequence is one of the central examples used in [5] for illustrating the advantage of multiple sequence structure prediction approach (i.e. `RNAscf`) over single sequence structure prediction (i.e. `mfold/RNAfold`). Indeed,

[2] The notion of compatibility here allows the types of pseudoknots that are captured by the `Pknots` program.

the mfold/RNAfold prediction for this sequence is quite poor as can be seen in figure 1(d). However, although RNAscf prediction using 20 sequences from 5s rRNA family is quite good, as reported in [5], the accuracy of the prediction deteriorates considerably when only 3 sequences, *E.coli*, *asellus aquaticus* and *cyprinus carpio* are used; this is illustrated in figure 1(e).[3] The prediction accuracy of the alifold program is also poor as depicted in figure 1(f). Most importantly, all of the above programs miss the most significant stem (enclosed by the base pair involving nucleotides 79 and 97) depicted in figure 1(b); when normalized by length, the mfold/RNAfold free energy table entry of this base pair is the smallest among all entries. (Compare this to the prediction of our program Densityfold , given in figure 1(c).)

We believe that some of the accuracy loss in structure prediction via total energy minimization can be attributed to "chance stems" which are sometimes chosen over "actual stems" due to problems commonly encountered in *local sequence alignment*. A stem is basically a local alignment between the RNA sequence and its reverse complement. Some of the energy minimization approaches (e.g. RNAscf program [5]) explicitly perform a local alignment search between the input RNA sequence and its reverse complement, in order to extract all potential stems of interest. However not all significant potential stems are realized in the actual secondary structure.

In the context of searching for significant alignments, the problems attributed to Smith-Waterman approach is usually considered to be a result of:
(1) the *shadow effect*, which refers to long alignments with relatively low conservation levels often having a higher score (and thus higher priority) than short alignments with higher conservation levels, and
(2) the *mosaic effect*, which refers to two highly conserved alignments with close proximity being identified as a single alignment, hiding the poorly aligned interval in between.
It is possible that the stem discovery process, which is performed either explicitly (e.g. in RNAscf) or implicitly (e.g. in mfold), may encounter with similar problems. For example, two potential stems, which, by chance, occur in close proximity, can easily be chosen over a conflicting longer stem due to the mosaic effect: the free energy penalty of an internal loop (which will be left in between the two chance stems) is often insignificant compared to the benefit of "merging" two stems.

In the context of local sequence alignment, the impact of these effects could be reduced by the use of *normalized sequence alignment* introduced by Arslan, Egecioglu and Pevzner [4]. The normalized local alignment problem asks to find a pair of substrings with maximum possible alignment score, normalized by their length ($+L$, a user defined parameter to avoid "trivial" alignments of length 1).

Inspired by this approach we propose to apply a *normalized free energy* or *energy density* criteria to compute the secondary structure of one or more RNA sequences. The algorithms we present aim to minimize the sum of *energy densities* of the substructures of an RNA secondary structure.[4] The *energy density of a base pair* is defined as the free energy of the substructure that starts with the base pair, normalized by the length of the underlying sequence. The energy density of an unpaired base is then defined to be the energy density of the closest base pair that encloses it. The overall objective of secondary structure prediction is thus to minimize the total energy density of all bases, paired and unpaired, in the RNA sequence.

[3] This example is particularly interesting as the independent mfold/RNAfold prediction for some of these sequences are very accurate.

[4] Note that, unlike the Arslan, Egecioglu, Pevzner approach we do not need to introduce an additive factor, L, artificially: a base pair in an RNA structure has at least three nucleotides in between.

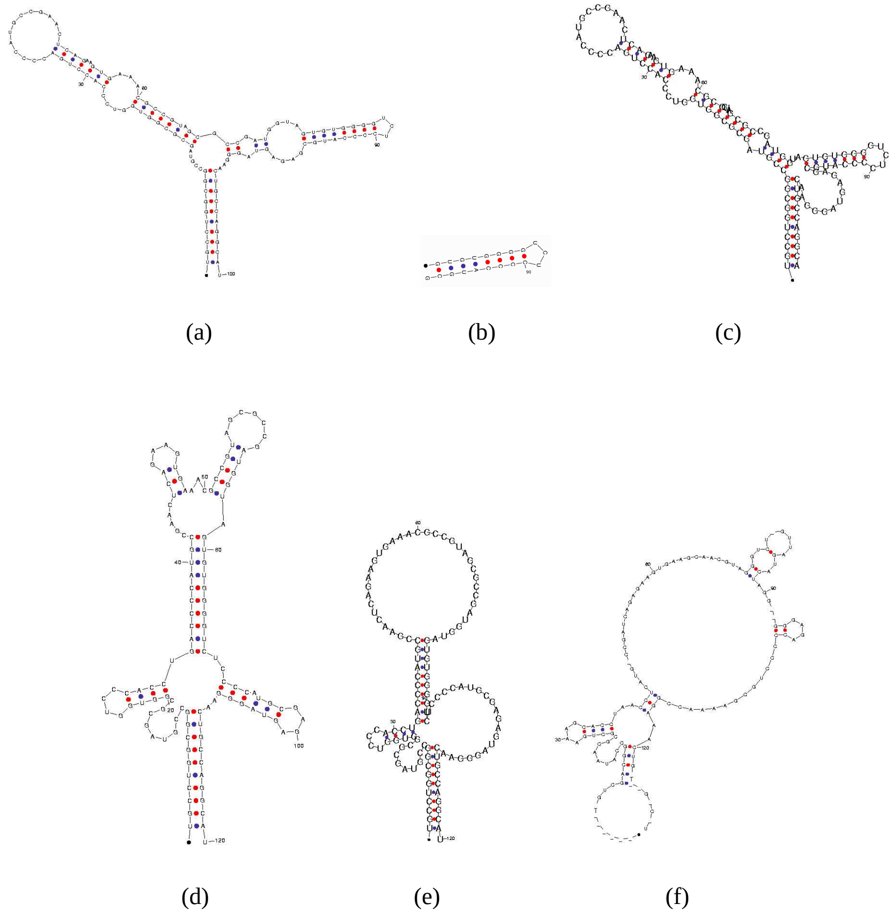

Fig. 1. (a) Known secondary structure of the *E.coli* 5S rRNA sequence. (b) The substructure with minimum energy density (missed by mfold/RNAfold, RNAscf and alifold programs). (c) Structure prediction by our Densityfold program. We capture the substructure with minimum energy density and correctly predict 28 of the 37 base pairs in the known structure. (d) Structure prediction by mfold/RNAfold program - only 10 of the 37 base pairs correctly predicted (e) Structure prediction by RNAscf program (consensus with the the *asellus aquaticus* and *cyprinus carpio* 5S rRNA sequences) - only 10 of the 37 base pairs correctly predicted (f) Structure prediction by alifold program (consensus with the *asellus aquaticus* and *cyprinus carpio* 5S rRNA sequences) - only 3 of the 37 base pairs correctly predicted.

The algorithms we describe in this paper also enables one to minimize a linear combination of the total energy density and total free energy of an RNA sequence. Based on these algorithms, we developed the Densityfold program for folding a single sequence and the MDensityfold program for folding multiple sequences. We tested the predictive power of our programs on the RNA sequence families used by

Bafna et al. [5] to measure the performance of the RNAscf program. We compare Densityfold and MDensityfold against all major competitors based on energy density minimization criteria - more specifically mfold/RNAfold , the best example of single sequence energy minimization, RNAscf , the best example of multiple sequence energy minimization without an alignment and alifold , the best example of multiple sequence energy minimization with an alignment. We show that when only one or a small number of functionally similar sequences are available, Densityfold can outperform the competitors, establishing the validity of energy density criteria as an alternative to the total energy criteria for RNA secondary structure prediction.

In the remainder of the paper we first describe a dynamic programming approach for predicting the secondary structure of an RNA sequence by minimizing the total free energy density. Then we show how to generalize this approach to minimize a linear combination of the free energy density and total free energy, a criteria that seems to capture the secondary structure of longer sequences. Because the running time of the most general approach is exponential with the maximum number of branches allowed in a multibranch loop we show how to approximate the energy density of such loops through a divide and conquer approach which must be performed iteratively until a satisfactory approximation is achieved. We finally provide some experimental results.

2 Energy Density Minimization for a Single RNA Sequence

We start with description of our dynamic programming formulation for minimizing the total free energy density of the secondary structure of an RNA sequence. We denote the input sequence by $S = S[1 : n]$; the i^{th} base of S is denoted by $S[i]$ and $S[i].S[j]$ denotes a base pair. Given input sequence S, its secondary structure $ST(S)$ is a collection of base pairs $S[i].S[j]$. A substructure $ST(S[i,j])$ is always defined for a base pair $S[i].S[j]$ and corresponds to the structure of the substring $S[i,j]$ within $ST(S)$. The base pair $S[i].S[j]$ is said to *enclose* the substructure $ST(S[i,j])$. The free energy of the substructure $ST(S[i,j])$ is denoted by $E_S(i,j)$. Thus the free energy density of $ST(S[i,j])$, denoted by $ED_S(i,j)$, is defined to be $E_S(i,j)/(j-i+1)$.

The notion of the free energy density of a substructure enables us to attribute an energy density value to each base $S[i]$. The individual energy density of $S[i]$, denoted $ED(i)$ is defined as the energy density of the smallest substructure that encloses $S[i]$. More specifically, let k be the largest index in S such that $S[k].S[\ell]$ form a base pair in $ST(S)$ for some ℓ with the property that $k < i < \ell$. Then the energy density attributed to $S[i]$ is $ED_S(k,\ell)$.

Our goal is to compute a secondary structure where the total energy density attributed to the bases is minimum possible. In this section we show how to minimize the total free energy density for S.

We first give some notation. The values of the following thermodynamic energy functions are provided in [15].

1. $eH(i,j)$: free energy of a hairpin loop enclosed by the base pair $S[i].S[j]$.
2. $eS(i,j)$: free energy of the base pair $S[i].S[j]$ provided that it forms a stacking pair with $S[i+1].S[j-1]$.
3. $eBI(i,j,i',j')$: free energy of the internal loop or a bulge that starts with base pair $S[i].S[j]$ and ends with base pair $S[i'].S[j']$ (an internal loop becomes a bulge if $i' = i+1$ or $j' = j-1$).
4. $eM(i,j,i_1,j_1,i_2,j_2,\ldots,i_k,j_k)$: free energy of a multibranch loop that starts with base pair $S[i].S[j]$ and branches out with basepairs $S[i_1,j_1], S[i_2,j_2],\ldots,S[i_k,j_k]$.

5. $eDA(j, j-1)$: free energy of an unpaired dangling base $S[j]$ when $S[j-1]$ forms a base pair with any other base (used for approximating eM).

By using the above functions we need to compute the following tables that correspond to total energies and energy densities of potential substructures.

1. $ED(j)$: minimum total free energy density of a secondary structure for substring $S[1, j]$.
2. $E(j)$: free energy of the energy density minimized secondary structure for substring $S[1, j]$.
3. $ED_S(i, j)$: minimum total free energy density of a secondary structure for $S[i, j]$, provided that $S[i].S[j]$ is a base pair.
4. $E_S(i, j)$: free energy of the energy density minimized secondary structure for the substring $S[i, j]$, provided that $S[i].S[j]$ is a base pair.
5. $ED_{BI}(i, j)$: minimum total free energy density of a secondary structure for $S[i, j]$, provided that there is a bulge or an internal loop starting with base pair $S[i].S[j]$.
6. $E_{BI}(i, j)$: free energy of an energy density minimized structure for $S[i, j]$, provided that a bulge or an internal loop starting with base pair $S[i].S[j]$.
7. $ED_M(i, j)$: minimum total free energy density of a secondary structure for $S[i, j]$, such that there is a multibranch loop starting with base pair $S[i].S[j]$.
8. $E_M(i, j)$: free energy of an energy density minimized structure for $S[i, j]$, provided there is a multibranch loop starting with base pair $S[i].S[j]$.

The above tables are computed via the following dynamic programming formulation. Note that as per mfold/RNAfold method we do not have any penalty for the unpaired bases at the very ends of the secondary structure.

$$ED(j) = \min \left\{ \begin{array}{l} ED(j-1) \\ \min_{1 \leq i \leq j-1} \{ED(i-1) + ED_S(i, j)\} \end{array} \right\}$$

$$ED_S(i, j) = \min \left\{ \begin{array}{lr} +\infty, & (i) \\ eH(i, j), & (ii) \\ 2\frac{eS(i,j)+E_S(i+1,j-1)}{j-i+1} + ED_S(i+1, j-1), & (iii) \\ ED_{BI}(i, j), & (iv) \\ ED_M(i, j) & (v) \end{array} \right\}$$

Note that in the item (iii) of the formula above, we need to use constant number 2. This is because the energy values are assigned to single bases; and the item (iii) computes the energy density of a substructure where the energy values of two different bases contribute to the total energy density of the corresponding substructure.

$$ED_{BI}(i, j) = \min_{i', j' | i < i' < j' < j} \left\{ \frac{eBI(i, j, i', j') + E_S(i', j')}{j - i + 1} \cdot [(i' - i) + (j - j')] + ED_S(i', j') \right\}$$

$$ED_M(i, j) = \min_{\substack{i_1, j_1, .., i_k, j_k | \\ i < i_1 < j_1 < ..i_k < j_k < j}} \left\{ \begin{array}{l} \frac{e_M(i,j,i_1,j_1,..i_k,j_k)+E_S(i_1,j_1)+...E_S(i_k,j_k)}{j-i+1} \\ \cdot [i_1 - i.. + j - j_k] + [ED_S(i_1, j_1) + ..ED_S(i_k, j_k)] \end{array} \right\}$$

For each (i, j), once the total energy density under the three possible structures (stack, bulge/internal loop and multibranch loop) are computed, the corresponding free energies can be computed as follows.

$$E_S(i,j) = \left\{ \begin{array}{ll} (i): & +\infty, \\ (ii): & eH(i,j), \\ (iii): & E_S(i+1,j-1) + eS(i,j), \\ (iv): & E_{BI}(i,j), \\ (v): & E_M(i,j) \end{array} \right\}$$

$$E_{BI}(i,j) = eBI(i,j,i',j') + E_S(i',j') \quad \mathit{for}\ i',j'\ \mathit{computed\ above}$$

$$\begin{array}{l} E_M(i,j) = eBI(i,j,i_1,j_1,\ldots i_k,j_k) + E_S(i_1,j_1)\ldots + E_S(i_k,j_k) \\ \mathit{for}\ i_1,j_1\ldots i_k,j_k\ \mathit{computed\ above} \end{array}$$

The algorithm above assumes that the maximum number of branches in a multibranch loop is k. Under this assumption the running time of the algorithm is $O(n^{k+2})$ and the space complexity is $O(n^2)$. Clearly this is not very practical for large values of k. Thus for $k > 2$ we make a number of simplifying assumptions on the free energy of a multibranch loop akin to the assumptions made by the mfold/RNAfold method. In particular we assume that the multibranch loop energy $eM(i,j,i_1,j_1,\ldots i_k,j_k)$ is a linear function of the number of unpaired bases and the dangling energies of the bases that follow the base pairs in the multibranch loop, namely $eDA(i+1,i), eDA(j-1,j),\ldots$. This assumption helps mfold/RNAfold to partition a multibranch loop into two iteratively, so that its minimum possible free energy can be computed in time linear with the size of the loop.

However, because we want to minimize the normalized free energy of the multibranch loop, which is non-linear, we can not apply the same divide-and-conquer approach directly. Thus we provide an alternative formulation which (at least in practice) converges to the correct value of the multibranch loop energy density in a small number of iterations. We describe this formulation in the next section.

3 Minimizing a linear combination of the energy density and energy

The initial tests we performed on the above dynamic programming formulation provided good outcomes for short RNA sequences; however as the sequence length increased, the predictive performance of this formulation deteriorated considerably. We noticed that although the energy density itself can help identify short structural motifs well, it may not provide the right criteria for “stitching them together”. Thus, in this section we describe a modified version of the dynamic programming formulation we gave above for energy density minimization. The goal of this modified version is to minimize a linear combination of the energy density and the total free energy. More specifically, for any $x \in \{S, BI, M\}$ let $ELC_x(i,j) = ED_x(i,j) + \sigma \cdot E_x(i,j)$. The function we would like to optimize is thus $ELC(n) = ED(n) + E(n)$.

$$ELC(j) = \min \left\{ \begin{array}{l} ELC(j-1) \\ \min_{1\leq i\leq j-1} \{ELC(i-1) + ELC_S(i,j)\} \end{array} \right\}$$

$$ELC_S(i,j) = \min \left\{ \begin{array}{lr} +\infty, & (i) \\ eH(i,j)\cdot(1+\sigma), & (ii) \\ 2\frac{eS(i,j)+E_S(i+1,j-1)}{j-i+1} + ELC_S(i+1,j-1) + \sigma\cdot eS(i,j), & (iii) \\ ELC_{BI}(i,j), & (iv) \\ ELC_M(i,j) & (v) \end{array} \right\}$$

$$ELC_{BI}(i,j) = \min_{i',j'|i<i'<j'<j} \left\{ \begin{array}{l} \frac{eBI(i,j,i',j')+E_S(i',j')}{j-i+1} \cdot [(i'-i)+(j-j')] \\ +ELC_S(i',j') + \sigma \cdot eBI(i,j,i',j') \end{array} \right\}$$

For computing the value of our optimization function for multibranch loops efficiently we have to perform an approximation to the multibranch loop energy density through a divide and conquer approach For this we have to define a new energy table $\overline{ELC}_M^{[i,j]}(k,\ell) = \overline{ED}_M^{[i,j]}(k,\ell) + \sigma \cdot \overline{E}_M^{[i,j]}(k,\ell)$ where $\overline{E}_M^{[i,j]}(k,\ell)$ and $\overline{ED}_M^{[i,j]}(k,\ell)$ are the free energy and the energy density of the optimal substructures for $S[k,\ell]$ provided that both $S[k]$ and $S[\ell]$ are on a multibranch loop starting with the base pair $S[i].S[j]$.

$$ELC_M(i,j) = \sigma \cdot a + \min_{i<k<j} \left\{ \overline{ELC}_M^{[i,j]}(i,k) + \overline{ELC}_M^{[i,j]}(k+1,j) \right\}$$

Here a is the multibranch loop opening score. Define:

$$\overline{b} = \frac{\widehat{E}_M(i,j)}{(j-i+1)}$$

where $\widehat{E}_M(i,j)$ is an estimation (a lower bound) for $E_M(i,j)$ of the optimal structure. The initial value of $\widehat{E}_M(i,j)$ is obtained through the following dynamic programming routine.

$$\widehat{E}_M(i,j) = a + \min_{i<k<j} \left\{ \overline{E}_M(i,k) + \overline{E}_M(k+1,j) \right\}$$

$$\overline{E}_M(k,k) = b$$

$$\overline{E}_M(k,\ell) = \min \left\{ \begin{array}{l} E_S(k,\ell) + c + eDA(k-1,k) + eDA(\ell,\ell+1) \\ \min_{k\leq h<\ell}\{\overline{E}_M(k,h) + \overline{E}_M(h+1,\ell)\} \end{array} \right\}$$

Here c is the contribution for each base pair on the multibranch loop and b is the unpaired base penalty. Based on this initial estimation $\widehat{E}_M(i,j)$ we have:

$$\overline{ELC}_M^{[i,j]}(k,k) = \overline{b} + \sigma \cdot b$$

$$\overline{ELC}_M^{[i,j]}(k,\ell) = \min \left\{ \begin{array}{l} ELC_S(k,\ell) + \sigma \cdot [c + eDA(k-1,k) + eDA(\ell,\ell+1)] \\ min_{k\leq h<\ell}\{\overline{ELC}_M^{[i,j]}(k,h) + \overline{ELC}_M^{[i,j]}(h+1,\ell)\} \end{array} \right\}$$

The corresponding energies of the substructures are as in the previous section:

$$E_S(i,j) = \left\{ \begin{array}{ll} (i): & +\infty, \\ (ii): & eH(i,j), \\ (iii): & E_S(i+1,j-1) + eS(i,j), \\ (iv): & E_{BI}(i,j), \\ (v): & E_M(i,j) \end{array} \right\}$$

$$E_{BI}(i,j) = eBI(i,j,i',j') + E_S(i',j') \quad for\ i',j'\ computed\ above$$

$$E_M(i,j) = eM(i,j,i_1,j_1,\ldots i_k,j_k) + E_S(i_1,j_1)\ldots + E_S(i_k,j_k)$$
$$for\ i_1,j_1\ldots i_k,j_k\ computed\ above$$

Note that if $E_M(i,j) \geq \widehat{E}_M(i,j) + \epsilon$ for some user defined (small) value of ϵ we set $\widehat{E}_M(i,j) = \widehat{E}_M(i,j) + \epsilon$ and re-iterate the above procedure for computing $ELC_M(i,j)$. The reader can easily verify that the running time of this dynamic programming algorithm is $O(n^4)$.

3.1 Multiple Sequence Energy Density Minimization

The dynamic programming algorithm for minimizing $ELC(n)$ for a single sequence is generalizable to multiple sequences without difficulty. Here we follow the general approach taken by the alifold program: we start with the multiple sequence alignment of the input sequences (obtained by the Clustal-W program) and fold the aligned sequences simultaneously, with the objective of minimizing the sum of energy densities of all bases from each sequence. This is somewhat different from the alifold and RNAscf methods as both of them assigns the *maximum energy* among aligned substructures to the energy of the consensus structure. We assign the *total energy and total energy density* of the aligned substructures to the energy and, respectively, energy density of the consensus structure. The gaps are also included in the calculations as a base.

The reader can verify that for m sequences the running time of this dynamic programming algorithm is $O(m \cdot n^4)$.

4 Experimental Results and Discussion

We implemented and tested the performance of our algorithms for minimizing the linear combination of the energy density and the total free energy of a single sequence as well as of multiple sequences, respectively called Densityfold and MDensityfold . Our test set is comprised of the same 12 RNA families from the Rfam database [9] used by Bafna et al. [5] for testing the performance of RNAscf program. Using this test set, we compared the performance of Densityfold and MDensityfold with varying values of σ (which determines the contribution of the total energy to the optimization function) against mfold/RNAfold , the best single sequence energy minimization program, alifold the best multiple sequence energy minimization program that uses the alignment between the input sequences, and RNAscf the best multiple sequence energy minimization program that computes the alignment and the folding simultaneously. In the context of multiple sequence folding, our goal is to demonstrate the predictive power of MDensityfold when only a limited number of sequences are available; thus we only report on the jointly predicted structures of a pair of sequences, randomly selected from each family.

The most common measure for demonstrating the predictive power of a single sequence secondary structure determination method is the number of correct base pairs (see for example [11]). Unfortunately the Rfam database only provides the consensus structure of a family and not individual sequences; thus it is not possible to reliably count the number of predicted base pairs which appear in the actual structure of an individual sequence and vice versa. To overcome this problem Bafna et al. used an alternative, *stack counting* measure [5] which is defined as the number of actual stacks and predicted stacks that overlap. As mentioned in [5] this measure is intended for comparing methods that explicitly extract stacks - which is not performed by most of the methods we compare.

We thus measure the predictive power of the programs we tested under the *structural edit distance* measure [12, 14]. which considers the differences between two RNA molecules in terms of both sequence/stack composition and structural elements. Given the *tree representation* of two RNA secondary structures, where each branch is labeled with a stack and every node represents a loop, their structural edit distance is defined to be the minimum possible sum of edit distances between the stack compositions of branch pairs and sequences of node pairs that are aligned to each other.

We computed the structural edit distances between the actual (consensus) structure of each of the 12 test families and the structure predictions by each test program via the RNA_align tool, publicly available on the web [2]. A distance of 0 corresponds to an identical sequence and structure, i.e. a perfect prediction. A higher distance value implies a poorer prediction.

	Single sequence methods				Multiple sequence methods		
Name	mfold/	Densityfold			MDensityfold	RNAscf	alifold
(Rfam_id)	RNAfold	$\sigma = 1.5$	$\sigma = 3$	$\sigma = 5$			
5s_rRNA (RF00001)	149	84	89	89	92	134	122
Rhino_CRE (RF00220)	94	93	93	93	77	88	30
ctRNA_pGA1 (RF00236)	45	83	83	83	48	91	44
glmS (RF00234)	194	288	230	230	189	249	198
Hammerhead_3 (RF00008)	2	2	2	2	74	2	88
Intron_gpII (RF00029)	100	93	103	103	85	113	78
Lysine (RF00168)	182	256	194	186	178	131	173
Purine (RF00167)	64	103	103	103	133	56	141
Sam_riboswitch (RF00162)	124	129	129	99	110	133	121
Thiamine (RF00059)	156	170	179	149	187	179	149
tRNA (RF00005)	31	67	67	67	50	31	32
ykok (RF00380)	158	200	189	189	168	203	157

Table 1. Structural edit distances between the actual (consensus) structure of a family and the predicted structures by each one of the programs tested.

The results of our comparative tests are summarized in the table above. (In addition, figure 1 demonstrates the outcome of Densityfold on the *E.coli* 5s_rRNA sequence (from RF00001 family) with that of mfold/RNAfold, alifold and RNAscf.) We used the default parameters in all programs we tested. We list the outcome of Densityfold for $\sigma = 1.5,\ 3.0$ and 5.0, and list the outcome of MDensityfold for the best possible σ value. As can be seen, Densityfold is at the top or near the top for most of the families. Densityfold with $\sigma = 5.0$ is always better than Densityfold with $\sigma = 3.0$. However Densityfold with $\sigma = 1.5$ outperforms both in a number of examples. Note that as σ approaches to ∞ the outcome of Densityfold gets more and more similar to the outcome of mfold/RNAfold.[5]

[5] In fact, we observed that for the families tested $\sigma = 100$ gives almost indistinguishable results to that by mfold/RNAfold.

However Densityfold with $\sigma = 5.0$ (the highest value we report) significantly outperforms mfold/RNAfold in a number of examples. Furthermore there is no clear winner between Densityfold and MDensityfold , each one outperforming the other in almost equal number of examples. However, in general, the longer the sequence gets, the better MDensityfold seemed to perform.

In conclusion, Densityfold demonstrates that an energy density minimization objective is a valid alternative to the total energy minimization objective. It can be used both on a single sequence or on multiple sequences. Our goal for the future is to test non-linear combinations of energy density and total energy as well as non-linear normalizations of the free energy as objective functions; we hope that such variations can explain the better performance of MDensityfold over Densityfold on longer sequences.

References

1. Mapping RNA Form & Function. *Science* **309 (5740)**, 2 September 2005.
2. RNA_align tool. http://www.csd.uwo.ca/faculty/kzhang/rna/
3. Akutsu, T., Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discr. Appl. Math.* **104 (1-3)**, 45-62, 2000.
4. Arslan, A.N., Egecioglu, O., & Pevzner, P.A., A New Approach to Sequence Comparison: Normalized Sequence Alignment. *Proc. RECOMB*, **ACM**, 2-11, 2001.
5. Bafna, V., Tang, H. & Zhang, S., Consensus Folding of Unaligned RNA Sequences Revisited. *Proc. RECOMB*, **LNBI 3500**, 172-187, 2005.
6. Condon, A., Davy, B., Rastegari, B., Zhao, S., Tarrant F., Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.* **320 (1)**, 35-50, 2004.
7. Davydov, E.& Batzoglou, S., A Computational Model for RNA Multiple Structural Alignment. *Proc. Symp. on Combinatorial Pattern Matching*, **LNCS 3103**, 254-269, 2004.
8. Gorodkin, J., Heyer, L. & Stormo, G., Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucl. Acids Res.* **25 (18)**, 3724-3732, 1997.
9. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S., Rfam: an RNA family database. *Nucl. Acids Res.* **31 (1)**, 439-441. 2003.
10. Hofacker, I., Fekete, M., Stadler, P., Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319 (5)**, 1059-1066, 2002.
11. Ji, Y., Xu, X., Stormo, G.D., A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* **20 (10)**, 1591-1602, 2004.
12. Lin, G., Ma, B., Zhang, K., Edit distance between two RNA structures. *Proc. RECOMB* **ACM**, 211-220, 2001.
13. Lyngso, R.B., Zuker, M. & Pedersen, C.N.S., Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics* **15 (6)**, 440-445, 1999.
14. Ma, B., Wang, L. & Zhang, K., Computing similarity between RNA structures *Theoretical Computer Science* **276 (1-2)**, 111-132, 2002.
15. Mathews, D., Sabina, J., Zuker, M. & Turner, D., Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288 (5)**, 911-940, 1999.
16. Mathews, D.& Turner, D., Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* **317 (2)**, 191-203, 2002.
17. Nussinov, R. & Jacobson, A., Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Nat. Acad. Sci. USA* **77 (11)**, 6309-6313, 1980.
18. Rivas, E. & Eddy, S.R., A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285 (5)**, 2053-2068, 1999.
19. Sankoff, D., Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM J. Appl. Math.* **45**, 810-825, 1985

20. Thompson, J., Higgins, D. & Gibson, T., Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680, 1994.
21. Tinoco, I., Uhlenbeck, O.& Levine, M, Estimation of secondary structure in ribonucleic acids. *Nature* **230 (5293)**, 362-367, 1971.
22. Zuker, M. & Stiegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9 (1)**, 133-148, 1981
23. Zuker, M., On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48-52, 1989.