Computational Approaches for determination of Most Probable RNA Secondary Structure Using Different Thermodynamics Parameters

[1]Binod Kumar,

Assistant Professor, Computer Sc. Dept, ISTAR, Vallabh Vidyanagar, Anand, Gujarat, India.

[2]Chetan R Dudhagara,

Lecturer, Computer Sc. Dept., NV Patel College, Vallabh Vidyanagar , Anand, Gujarat, India.

[3]Dr. N. N. Jani,

Director, Faculty of Computer Sc. & Information Tech., Kadi Sarva Vishvavidyalaya Univ., Gandhinagar, India

ABSTRACT : Many bioinformatics studies require the analysis of RNA structures. More specifically, extensive work is done to elaborate efficient algorithms able to predict the 2-D folding structures of RNA. The core of RNA structure is a dynamic programming algorithm to predict RNA secondary structures from sequence based on the principle of minimizing free energy. In this paper the thermodynamic data have been used for RNA predictions.

In this paper the free energy minimization and the partition function code has been used to predict internal loops of any size in O (N^3) time. The free energy table for multibranch loops has been used by Dynalign. Base pair probabilities have been determined by the partition function calculation. Parameters controlling the prediction of suboptimal structures are Max % Energy Difference and Max Number of Structures. The fold module provides the basic implementation of RNA secondary structure prediction.

A Dynalign dot plot, a separate dot plot is generated for each of the two sequences involved. OligoScreen calculates the unimolecular and bimolecular folding free energies for a set of RNA oligonucleotides.

KEYWORDS: Minimum Free Energy, Partition Function, Dynalign, OligoWalk, RNA Folding.

1. INTRODUCTION

The computation of secondary structural folding of RNA or single-stranded DNA is a key element in many bioinformatics studies and has been extensively studied for many years. The firsts to propose an algorithm to predict the folding structure of RNA sequences is Waterman, Smith [1, 2]. This algorithm is based on **Dynamic Programming** with a complexity of O (n^3).

Following this pioneer work, several improvements have been done leading to different kinds of dynamic programming algorithms. We can cite: (1) the computation of the most stable structure through energy minimization running in O(n^3), introduced by Zuker and Stiegler [3] which outputs a single optimal structure and its corresponding energy ; (2) the computation of a partition function over all possible structures for deriving additional properties of the thermodynamic ensemble such as the base pairing probabilities of any base pair [4] ; (3) the computation of suboptimal structures which generates all structures within a given energy range of the optimal one.

2. RNA FOLDING ALGORITHIM

This section exposes the principles of the folding algorithm.

2.1 RNA STRUCTURE

RNA is transcribed (or synthesized) in cells as single strands of (ribose) nucleic acids. However, these sequences are not simply long strands of nucleotides. Rather, intra-strand base pairing will produce structures such as the one shown below.

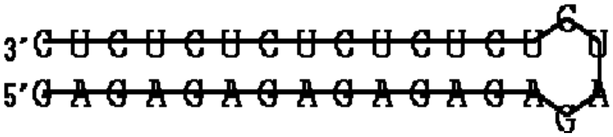

3'C U C U C U C U C U C U C U G U
5'G A G A G A G A G A G A G A G A

Figure 1: RNA Secondary Structure

In RNA, guanine and cytosine pair (GC) by forming a triple hydrogen bond, and adenine and uracil pair (AU) by a double hydrogen bond; additionally, guanine and uracil can form a single hydrogen bond base pair.

The stability of a particular secondary structure is a function of several constraints:

1 The number of GC versus AU and GU base pairs.(Higher energy bonds form more stable structures.)
2 The number of base pairs in a stem region. (Longer stems result in more bonds.)
3 The number of base pairs in a hairpin loop region.(Formation of loops with more than 10 or less than 5 bases requires more energy.)
4 The number of unpaired bases, whether interior loops or bulges. (Unpaired bases decrease the stability of the structure.)

2.2 ENERGY MODEL

The algorithm is designed to find the most stable structure of a RNA sequence. It is used for the search of micro RNAs where the stability of the secondary structure is an important feature.

A secondary structure is described by a list of base pairs *i · j* where each base forms at most one pair. The algorithm is based on a decomposition of the secondary structure into its constituent loops. Each loop is associated with an experimentally measured energy according to its sequence, length and type.

The stability of a secondary structure is quantified as the amount of free energy released or used by forming base pairs. Positive free energy requires work to form a configuration; negative free energies release stored work. Free energies are additive, so one can determine the total free energy of a secondary structure by adding all the component free energies (units are kilocalories per mole). The more negative the free energy of a structure, the more likely is formation of that structure, because more stored energy is released. This fact is used to predict the secondary structure of a particular sequence.

To compute the minimum free energy of a sequence, empirical energy parameters are used. These parameters summarize free energy change (positive or negative) associated with all possible pairing configurations, including base pair stacks and internal base pairs, internal, bulge and hairpin loops, and various motifs which are know to occur with great frequency.

2.3 ALGORITHM

The dynamic programming [5] algorithm uses three tables: Q′$_{i,j}$ is the minimum energy of folding of a subsequence *i, j* given that bases *i* and *j* form a base pair; Q $_{i,j}$ and QM $_{i,j}$ are the minimum energy of folding of the subsequence *i, j* assuming that this subsequence is inside a multiloop and that it contains respectively at least one and two base pairs. A simplified model of the recursion relations can be written as:

$$Q'_{i,j} = \begin{cases} \min \left\{ \begin{array}{ll} Eh(i,j) & \text{hairpin loop} \\ Es(i,j) + Q'_{i+1,j-1} & \text{stacked pair} \\ \min\limits_{k,l \in [i,j]^2} Ei(i,j,k,l) + Q'_{k,l} & \text{interior loop} \\ QM_{i+1,j-1} & \text{multiloop} \end{array} \right\} & \textit{if pair } i \cdot j \textit{ is allowed} \\ \infty & \textit{if pair } i \cdot j \textit{ is not allowed} \end{cases} \qquad (1)$$

$$QM_{i,j} = \min_{i<k<j} \left(Q_{i,k} + Q_{k+1,j} \right) \qquad (2)$$

$$Q_{i,j} = \min \left\{ QM_{i,j}, \min(Q_{i+1,j}, Q_{i,j-1}), Q'_{i,j} \right\} \qquad (3)$$

Eh(*i,j*) E(*i,j,k,l*) E$_s$(*i,j*)are respectively energies of :

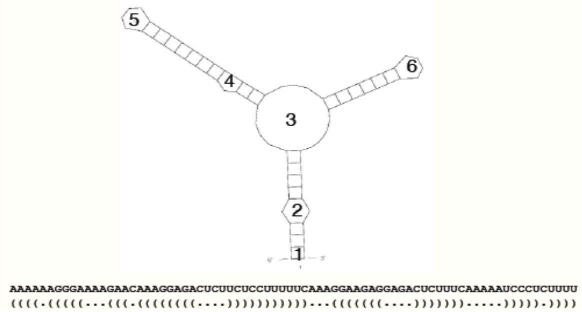


Figure 2: Secondary Structure .The secondary structure begins in 1 with stacked base pairs (two closing base pairs with both sides of the loop of length zero). 2 is an interior loop (two closing base pairs with both sides non null). 3 show a multiloop (several closing base pairs). 4 is a bulge loop (two closing base pairs with one loop side of length zero and the other greater than zero. 5 and 6 are hairpin loops (one closing base pair). The structure can also be written in a dot bracket representation where an unpaired base is a dot and a base pair is a matching pair of parenthesis. The free energy of the structure is the sum of the energies of its constituent loops.

– Eh(*i, j*): a hairpin loop closed by the pair *i · j*.
– Ei(*i, j, k, l*): an interior loop formed by the two base pairs *i · j, k · l*.
– Es(*i, j*): two stacked base pairs *i · j* and (*i* + 1)·(*j* − 1).

These functions compute energies through the use of lookup tables containing energy parameters according to the size and sequence of the loop.

E_j being the minimum free energy of subsequence 1 . . . *j*, the minimum free energy E_n of the whole sequence is then obtained through the recursion:

$$E_j = \min\left\{E_{j-1}, \min_{1<k<j}(E_{k-1} + Q'_{k,j})\right\} \qquad (4)$$

Dynamic programming using this recursion computes the minimum free energy of a sequence of length n in $O(n^2 \cdot L^2 + n^3)$ by restricting the loop size of interior loops to L. The corresponding secondary structure is then obtained by a trace-back procedure.

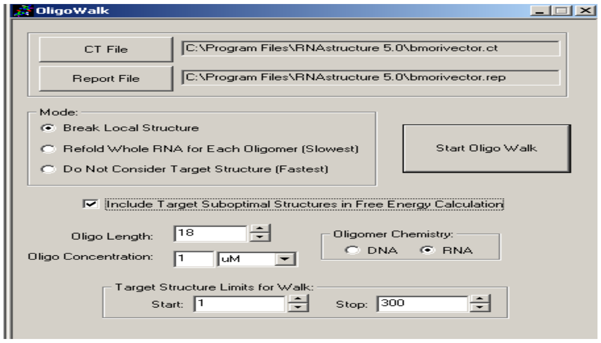
2.4 STRING NOTATION

Several representations of secondary structure have been utilized, each with different advantages. The planar graph representation shown above gives an intuition for the shape of an RNA sequence, but the same structure could also be represented in string notation. In string notation, balanced parenthesis is used to indicate paired bases, and periods are used to indicate unpaired bases. The secondary structure in the above figure is given as ((((((((((((((....)))))))))))))) in string notation.

The number of possible secondary structures (*S*) of *n* bases with *k* base pairs is given as

$$S(n,k) = \frac{1}{k}\binom{n-k}{k-1}\binom{n-k+1}{k-1} \qquad (5)$$

2.5 OLIGOWALK



CT File: C:\Program Files\RNAstructure 5.0\bmorivector.ct

Report File: C:\Program Files\RNAstructure 5.0\bmorivector.rep

Mode:
- ● Break Local Structure
- ○ Refold Whole RNA for Each Oligomer (Slowest)
- ○ Do Not Consider Target Structure (Fastest)

☑ Include Target Suboptimal Structures in Free Energy Calculation

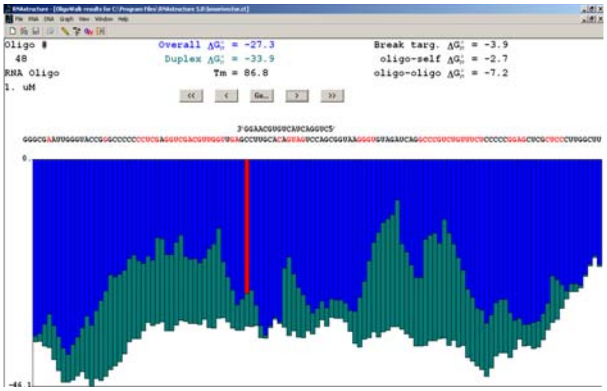
Oligo Length: 18

Oligo Concentration: 1 uM

Oligomer Chemistry: ○ DNA ● RNA

Target Structure Limits for Walk: Start: 1 Stop: 300

Figure 3: OligoWalk


Oligo # 48
RNA Oligo
1. uM

Overall ΔG_{37}° = -27.3
Duplex ΔG_{37}° = -33.9
Tm = 86.8

Break targ. ΔG_{37}° = -3.9
oligo-self ΔG_{37}° = -2.7
oligo-oligo ΔG_{37}° = -7.2

Figure 3: Graph of OligoWalk

This provides [6] a rapid search interface for picking out an oligonucleotide that binds strongly to its target, according to thermodynamic data. It not only calculates energy required to break target structure, but also structure due to oligonucleotide folding, if any.

The given ΔG's are as follows:

1. Overall ΔG: This is the net ΔG in kcal/mol of oligo-target binding, when all contributions are considered, including breaking target structure and oligo self-structure, if any. A more negative value indicates tighter binding.
2. Duplex ΔG (Binding ΔG): This is the ΔG of the oligo-target binding from unstructured states.
3. Break targ. ΔG: This number provides the energy penalty (hence it's usually positive) due to breaking of intramolecular target base pairs when oligo is bound.
4. Oligo-self ΔG: This provides the ΔG of intramolecular oligo structure. If it is zero, there is no stable intramolecular structure. A negative number indicates this self structure is stable, making for unfavorable oligo-target binding.
5. Oligo-oligo ΔG: This is the ΔG of intermolecular oligo structure. It is non-zero if one oligomer molecule can bind to another. A negative number indicates a stable oligo-oligo duplex, making for unfavorable oligo-target binding.

T_m: This is a melt temperature in degrees C for the duplex formation, i.e. the temperature at which half the target strands are bound with oligomer.

The middle part of the screen shows the current oligo (3'->5') bound to the target (5'->3'). Target bases appear red if they are paired in the folded target structure and are black otherwise. Target base numbers are given.

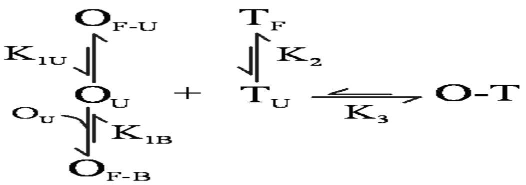
At the bottom, some ΔG values are displayed graphically. The default is to show binding ΔG in green and overall ΔG in blue (red for the currently shown oligo). Due to breaking of target and oligo structure, the blue bars are generally smaller than the green bars. Downward bars indicate negative ΔG and upward bars indicate positive ΔG. All bars start at zero energy.

2.5.1 EQUILIBRIA AND CALCULATIONS

2.5.1.1 OLIGOWALK EQUILIBRIUM

When designing an antisense oligonucleotide (oligomers) that binds with high affinity, it is desirable to consider the structure of the target RNA strand and the antisense oligomer. Specifically, for an oligomer to bind tightly, it should be complementary to a stretch of target RNA that has little self-structure. Also, the oligomer should have little self-structure, either intramolecular or bimolecular. Breaking up any self-structure amounts to a binding penalty.

OligoWalk considers the following equilibrium:



In this reaction, O is the oligomer, T is the target, O-T is the oligomer-target complex. OF is self-structured oligomer either unimolecular (U) or bimolecular (B) and TF is self-structured target (unimolecular). Bimolecular target-target interactions are neglected because the concentration of target is low. OU is unfolded oligomer and TU is unfolded target in the region of oligomer complementarity. These structures are in equilibrium with each other, with equilibrium constants K_{1U}, K_{1B}, K_2, and K_3.

2.5.1.2THE T_M CALCULATION IN OLIGOWALK

OligoWalk calculates [7] a melt temperature for the duplex formation of antisense-target binding. This calculation neglects target structure and antisense oligonucleotide structure.

Consider the equilibrium of:

$$O_{RC} + T_{RC} \xleftrightarrow[K]{} O - T \qquad (6)$$

where the random coil oligomer binds to random coil target with an equilibrium constant K. If we assume that:

[OR.C.] >> [TR.C.]

then the T_m will be the temperature at which half the target is bound or:

$$[TR.C.] = [O\text{-}T] = [Target]\ Total/2 \qquad (7)$$

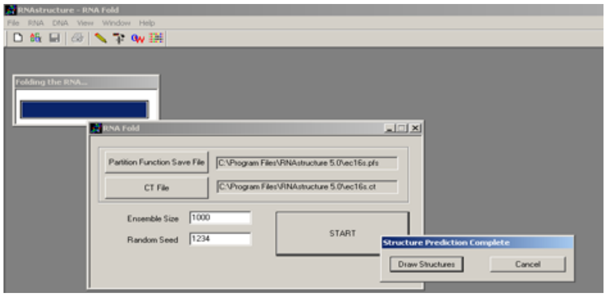
Knowing that:
$\Delta G = \Delta H - T\Delta S$ and $\Delta G = -RT\ \mathbf{ln}\ (K)$ (8)

$$\text{then } T_m = \frac{\Delta H}{\Delta S + R\ln[oligomer]_{Total}} \qquad (9)$$

where R is the gas constant and T_m is in K. OligoWalk converts T_m to degrees C.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 PARTITION FUNCTION



Partition Function Save File: C:\Program Files\RNAstructure 5.0\ec16s.pfs
CT File: C:\Program Files\RNAstructure 5.0\ec16s.ct
Ensemble Size: 1000
Random Seed: 1234

Figure 4: Calculation of RNA Partition Function

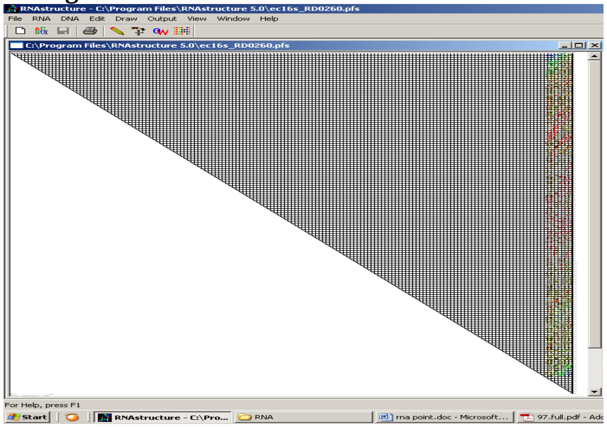
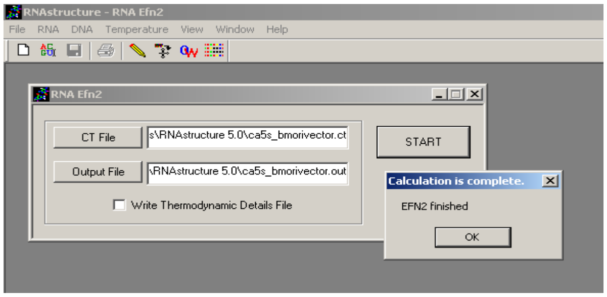

RNAstructure - C:\Program Files\RNAstructure 5.0\ec16s_RD0260.pfs
File
RNA
DNA
Edit
Draw
Output
View
Window
Help
C:\Program Files\RNAstructure 5.0\ec16s_RD0260.pfs
For Help, press F1
Start
RNAstructure - C:\Pro...
RNA
rna point.doc - Microsoft...
97.full.pdf - Adc

Figure 5: Graph of Partition Function of RNA

The partition function calculation has been used to predict the base pairing probabilities for all possible canonical base pairs in a sequence. The predicted probabilities are displayed in a probability dot plot, or predicted secondary structures can be color annotated with these probabilities. More probable pairs are more likely to be correctly predicted than less probable pairs.

3.2 RNA ENERGY FUNCTION 2



RNAstructure - RNA Efn2

RNA Efn2

CT File: s\RNAstructure 5.0\ca5s_bmorivector.ct

Output File: \RNAstructure 5.0\ca5s_bmorivector.out

□ Write Thermodynamic Details File

Calculation is complete.

EFN2 finished

Figure 6: RNA Energy Function Calculation

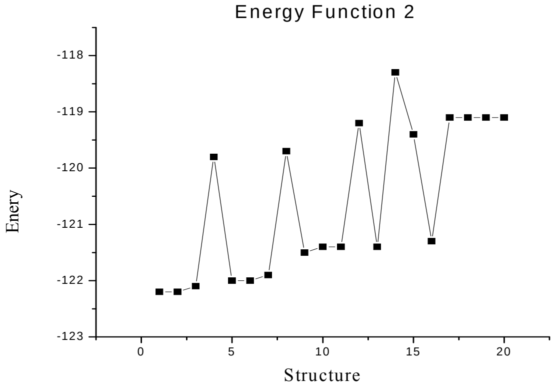
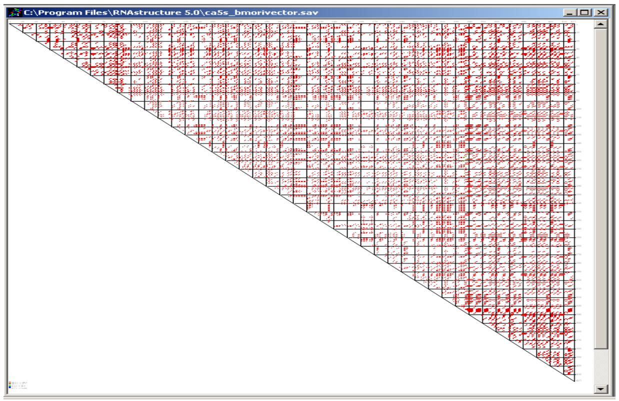


Figure 7: Energy vs. Structures (20) Graph

RNA Partition Function [8] determines the free energy (-ve) of a secondary structure saved.

3.3 PARTITION FUNCTION DOT PLOT

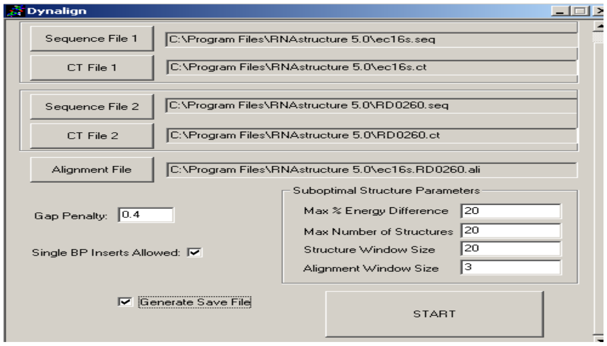


C:\Program Files\RNAstructure 5.0\ca5s_bmorivector.sav

Figure 8: Partition Function Dot Plot

Partition function dot plots display not the energy of a pair, but the probability of that pair existing, as predicted by the partition function. Dots are displayed as -log$_{10}$ (probability). The default view shows all possible pairs of any probability.

3.4 DYNALIGN



Dynalign

Sequence File 1: C:\Program Files\RNAstructure 5.0\ec16s.seq

CT File 1: C:\Program Files\RNAstructure 5.0\ec16s.ct

Sequence File 2: C:\Program Files\RNAstructure 5.0\RD0260.seq

CT File 2: C:\Program Files\RNAstructure 5.0\RD0260.ct

Alignment File: C:\Program Files\RNAstructure 5.0\ec16s.RD0260.ali

Gap Penalty: 0.4

Single BP Inserts Allowed: ☑

Suboptimal Structure Parameters

Max % Energy Difference: 20

Max Number of Structures: 20

Structure Window Size: 20

Alignment Window Size: 3

☑ Generate Save File

Figure 9: Dynalign Calculation

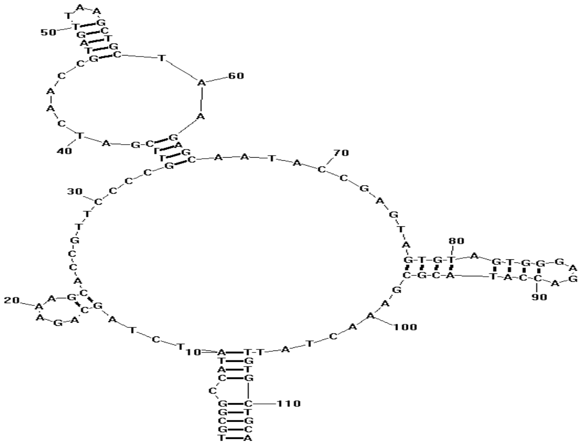
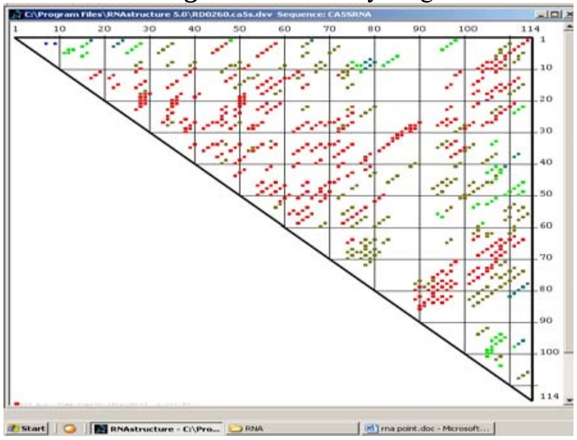


Figure 10: RNA Dynalign



C:\Program Files\RNAstructure 5.0\RD0260.ca5s.dsv Sequence: CA5SRNA

Figure 11: Dynalign Dot plot

Dynalign [9] uses the mutual information of the two sequences to constrain secondary structure prediction. This can result in a large improvement in the accuracy of secondary structure prediction. The algorithm generates an alignment of the two sequences, but does not depend on sequence similarity.

For Dynalign dot plots, a separate dot plot is generated for each of the two sequences involved.

3.5 RNA FOLD

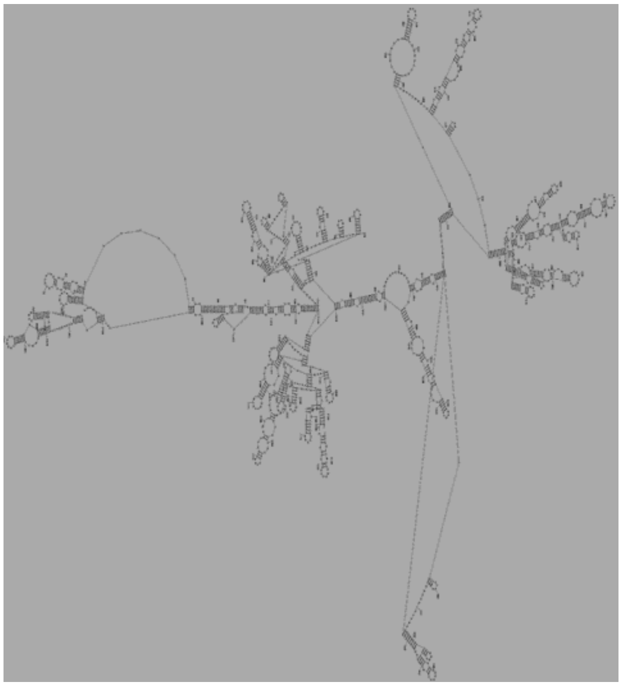


Figure 12: Fold RNA single strand

The fold module provides the basic implementation of RNA secondary structure prediction.

The folding of two strands functions in much the same way as folding of a single strand.

3.6 OLIGOSCREEN

Table 1: Calculation of the Unimolecular and Bimolecular Folding Free Energies

Sequence	DGbimolecular	DGunimolecular	DGduplex	DG2BPat5'	DG2BPat3'
GGGCCAAUGCGA	-9.3	-0.3	-23.1	6.6	4.3
UUUAAACCGGCC	-7.1	0.0	-18.2	1.8	6.7
GGGAUGCA	-3.1	0.0	-13.1	6.6	5.0
CGGAUUCGA	-9.6	0.0	-12.7	5.7	4.3
GGCAUUCGGG	-2.7	0.0	-18.1	6.7	6.6

OligoScreen [10] calculates the unimolecular and bimolecular folding free energies for a set of RNA oligonucleotides.

The OligoScreen parameters are a subset of those calculated by OligoWalk.

4. CONCLUSION

By incorporating experimental information as a free energy change term in RNAstructure, we determine the structures of RNA using RNA Fold and RNA Dynalign.

Dynalign predicts a set of low energy structures and alignments, called suboptimal structures. A set of parameters are used to define how many suboptimal structures to generate and how different from each other the suboptimal structures should be. RNA, "Fold Bimolecular" allows for folding of two distinct strands. Energy minimization methods have been so well refined that a series of energetically feasible models and the most thermodynamically probable structural models may be computed.

Partition Function is used to predict the base pairing probabilities for all possible canonical base pairs in a sequence. RNA Energy Function with free energy (-ve) determines stability in a secondary structure. OligoScreen calculates a hybridization free energy for those strands annealed to a complementary RNA target. Maximum Expected Accuracy predicts a specific subset of structures composed of probable base pairs and single-stranded nucleotides.

RNA structural analysis may be used to search reliably through genomic sequences for genes that encode these RNA molecules. The successful analysis of these types of RNA molecules could be readily extensible to other classes of RNA molecules.

5. REFERENCES

[1] Waterman, M., Smith, T.: RNA secondary structure: a complete mathematical analysis. Math. Biosci 42 (1978) 257–266

[2] Nussinov, R., Pieczenik, G., Griggs, J., Kleitman, D.: Algorithms for loop matchings. SIAM J. Appl. Math 35(1) (1978) 68–82

[3] Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9(1) (1981) 133–148

[4] McCaskill, J.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29(6-7) (1990) 1105–19.

[5] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. "Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure." *Proceedings of the National Academy of Sciences USA*, 101:7287-7292. (2004).*Nucleic Acids Res*. **19:** 2707–2714.

[6] D.H. Mathews, M.E. Burkard, S.M. Freier, J.R. Wyatt, and D.H. Turner. "Predicting Oligonucleotide Affinity to Nucleic Acid Targets." *RNA*, 5:1458-1469. (1999).

[7] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure." *Journal of Molecular Biology*, 288:911-940. (1999).

[8] D.H. Mathews. "Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization." *RNA*, 10:1178-1190. (2004).

[9] A. Harmanci, G. Sharna, and D.H. Mathews. "Efficient Pairwise RNA Structure Prediction Using Probabilistic Alignment Constraints in Dynalign." *BMC Bioinformatics*, 8:130-150. (2007).

[10] D.V. Mateeva, D.H. Mathews, A.D. Tsodikov, S.A. Shabalina, R.F. Gesteland, J.F. Atkins, and S.M. Freier. "Thermodynamic Criteria for High Hit Rate Antisense Oligonucleotide Design." *Nucleic Acids Research*, 31:4989-4994. (2003).