A convoluted problem and heuristic solution for predicting ensembles and transitions for pseudoknots polymers

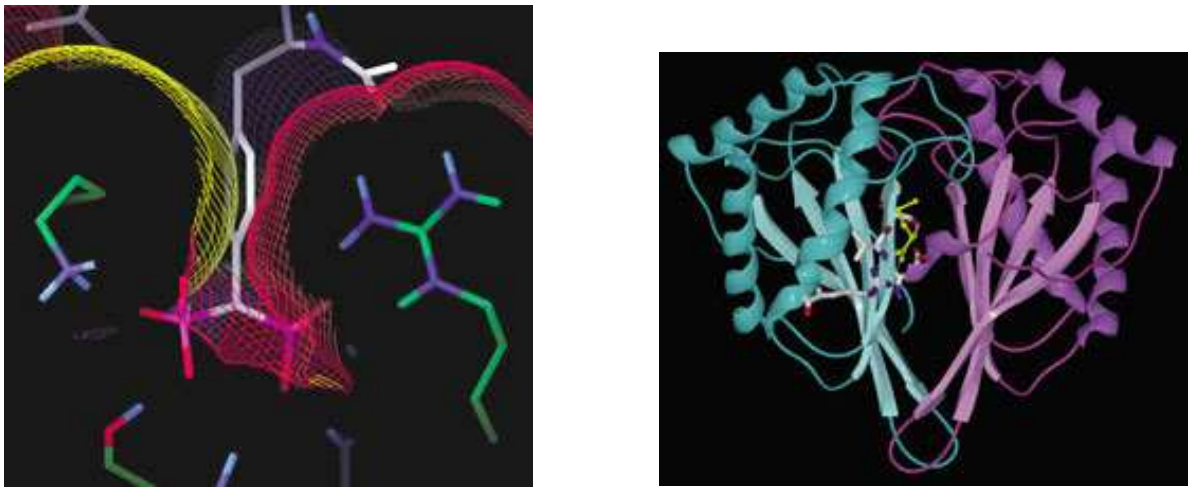
Ezekiel. F. Adebiyi, Ph.D
Department of Computer and Information Sciences
College of Science and Technology,
Covenant University,
PMB 1023, Ota, Ogun State, Nigeria.
Tel: +234 (0)1 7900724, 7901081 Ext 3021
E-mail: eadebiyi@sdsc.edu

May 28, 2007

The Structure-based method to drug design

- Focusing on malaria research
- Given a known anti-malaria drug (Drug/Compound, e.g chloroquine), in drug design/docking multi-tasking,
- the aim is to optimize its binding to the receptor (a drug target) using fragment based docking programs or other docking software.
- This way, one can design a new anti-malaria drug with improved activity and reduced liability to development of resistance by the plasmodium parasite.
- The procedure encapsulated above is called the structure-based method to drug design.

- Success story in this direction!



(a) Drug Target: Lux-S-a metalloenzyme in bacterial quorum sensing is shown bound to the small molecule methionine, (b) Docked: shown is the disphosphonomethyl group of one of Ariad pharceutical's compounds or drugs docked in the receptor binding site of Src SH2. (Bohacek, et al., J. Med. Chem., 44, 660, 2001)

- The starting point of a structural based method to drug design is the availability of drug compounds and drug targets.

- RNA has been identified to have great potential as drug targets and tRNA has been the most viable of them all (Herrman and Westhof, Combinatorial Chem. and High Throughput Screening, 3, 219-234, 2000).

- We have about five thousand, seven hundred and fourteen (5714) genes in P.f and

- currently, seventy-two (72) tRNA genes have been identified on chromosomes 2,3,4,5,6,7,11,12,13,14, sequences X95275 and X95276 of P.f.

- 114 structures produced experimentally using the X-ray crystallography and the NMR are available in the Protein Data Bank (henceforth PDB) at the University of California at San Diego (UCSD).

- As far as we know, non exist for any known tRNA genes.

- Note that experimental determination of these structures is very expensive and

- focus is on the automation of the process of structures prediction.

- This is otherwise known as the *RNA folding problem*.

- This is the focus of our work in this report.

tRNA structures prediction: Problem formulation and Previous results

- One of main drawback of predicting structures for tRNAs is because, tRNAs contain psendoknots.

- But solving the *RNA folding problem* is enhanced as we study the *RNA (that include tRNA) folding energy landscapes*.

- This is because the energy landscapes help us to:

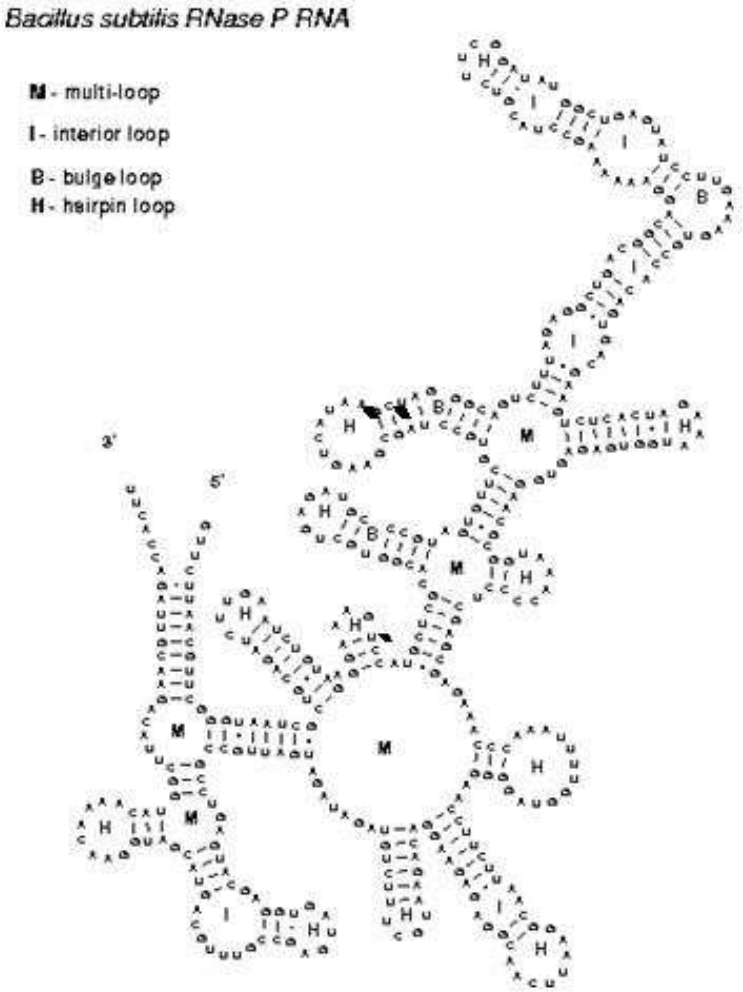
  1. know how the folding process takes place,
  2. predict the folding routes, and
  3. understands folding thermodynamics, cooperativity, intermediate states, transitions states and conformational transitions.

- The RNA folding is the process of folding the one-dimensional primary structure (sequence) into three-dimensional tertiary structure.

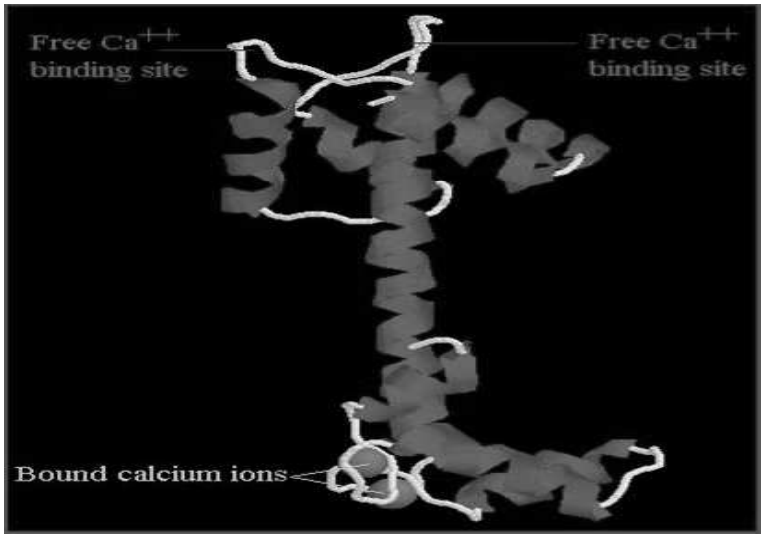
- Graphically, RNA folding process can be described as follows:

  Given is a one–dimensional primary structure (sequence)

- Next is the formation of the secondary structure, that is, of the pattern of Watson-Crick (and GU) base pairs.


Bacillus subtilis RNase P RNA
M - multi-loop
I - interior loop
B - bulge loop
H - heirpin loop
3'
5'

- And finally is the embedding of the contact structure in three dimensional space.



- Let represent a chain of an RNA sequence as $R = r_1, r_2, ..., r_n$, where r_i is called the (ribo)nucleotide.

- Each r_i belongs to the set $\{A, C, G, U\}$.

- We will refer to i as the i^th base of the sequence.

- A secondary structure, or folding, on R is a set S of ordered pairs, written as $i.j$, $1 \leq i < j \leq n$ satisfying:

  1. $j - i > 4$
  2. If $i.j$ and $i'.j'$ are 2 base pairs, (assuming without loss in generality that $i \leq i'$), then either:
     (a) $i = i'$ and $j = j'$ (they are the same base pair),
     (b) $i < j < i' < j'$($i.j$ precedes $i'.j'$ or unrelated links), or
     (c) $i < i' < j' < j$. (includes or nested links).

- *The last condition excludes pseudo-knots*. These occur when 2 base pairs, $i.j$ and $i'.j'$ satisfy $i < i' < j < j'$ (*cross links or linked links*).

- A brief overview of previous work is shown in the table below:

Table 1

Algorithm	ψ	*Abbr.*	*Remark*
deterministic			
Minimum Free Energy	-	*MFE*	*fast*
Kinetic Folding	+	*KIN*	*fast*
$5' - 3'$ *Folding*	+	*5-3*	*fast*
Partition Function	-	*PF*	*ensemble*
Maximum Matching	-	*MM*	*unrealistic*
stochastic			
Simulated Annealing	+	*SA*	*very slow*

Folding Algorithm for RNA Secondary Structures (ψ indicate Pseudo-knots can be included.)

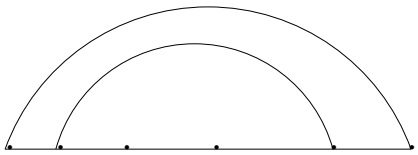
- Extended volume within loops and between the sub-structures are been neglected.

- Considerably more sophisticated and accurate is a model that treats the steric excluded volume of the chain explicitly. it.

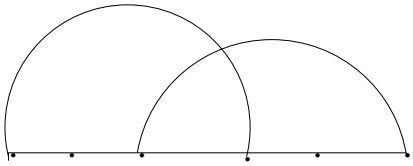
Polymer Graph

- Let's recall some details on the problem of the RNA folding problem:

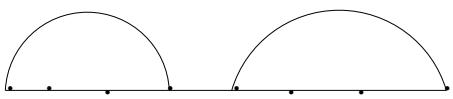
- If $i.j$ and $i'.j'$ are 2 base pairs, (assuming without loss in generality that $i \leq i'$), then either:
  1. $i = i'$ and $j = j'$ (they are the same base pair),
  2. $i < j < i' < j'$($i.j$ precedes $i'.j'$ or unrelated links), or
  3. $i < i' < j' < j$ (includes or nested links).
  4. $i < i' < j < j'$ (cross links or linked links).



(i) nested

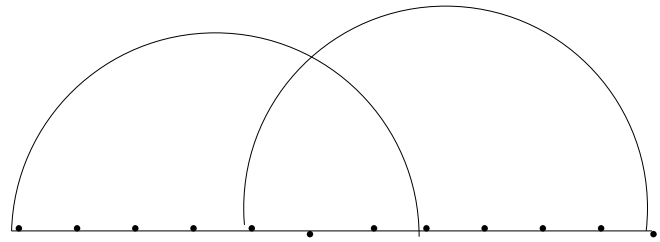


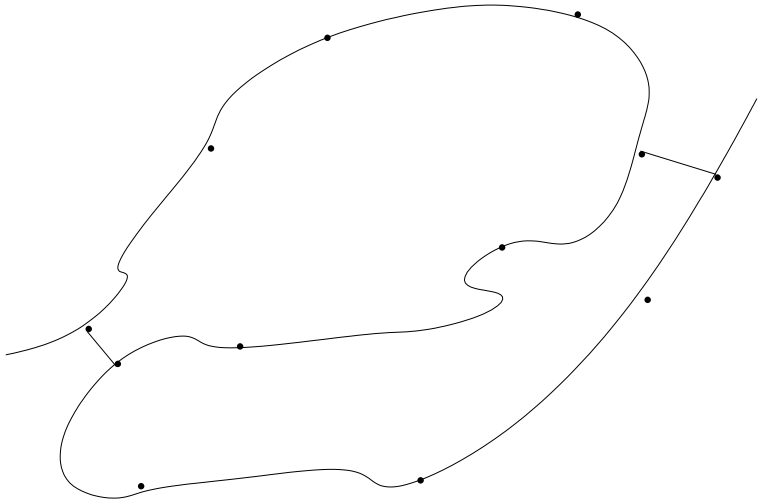
(ii) linked



(iii) unrelated

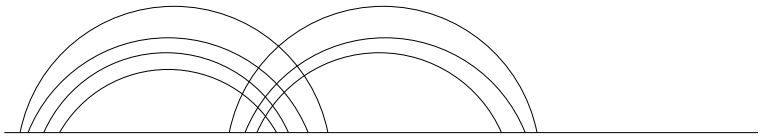
- A type of a polymer structure and the corresponding chain conformation:



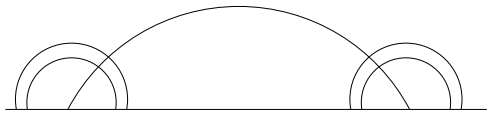


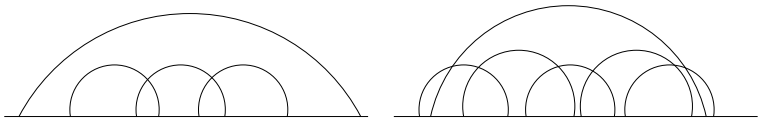
- Polymer graph encapsulates very well Pseudo-knots.

- Simple Pseudo-knots:



- Complex pseudo-knots:





A viable Computational pathway for the RNA folding problem

- This was derived by Ken Dill in the work with S-J. Chen.

- This can be enumerated as follows:

  1. Homopolymer RNA: Develop as general as possible, graph based mechanism to generate density of state, $g(n)$,

     $$g(n) = \sum graphs \cdot \#conformation\ per\ graph,$$

     for any number of contacts and conformation.

  2. Using the first step results, derive population, $P(n)$, of each state, T.

  3. Hetero-polymer RNA: Main work here is to develop a general code for Q, the partition function for any monomer sequence.

  4. Experiment:

     (a) Derived heat capacity, C_p, from the partition function.

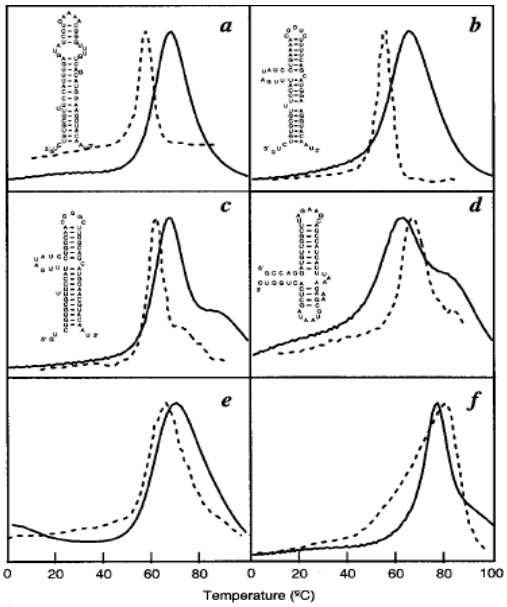
     (b) Adjust parameters to treat energies of non-canonical bases and any other tertiary interactions.

  5. Prediction of the energy landscapes for RNA molecules.

- Why energy landscapes:
  1. For designing faster and more robust computer methods for predicting native structures.
  2. To capture more figuratively, Ligand binding to proteins and RNA molecules and catalytic mechanisms.

- For structures of hairpin without pseudo-knots: S-J. Chen and Ken Dill implemented the items 1-5 above and demonstrate the use of the energy landscapes (PNAS, 2000).

- Key: Predicted (continuous lines) and experimental differential melting curves (dashed lines).



mutant α mRNA fragment (c) and E.coli 23s rRNA fragment (d)

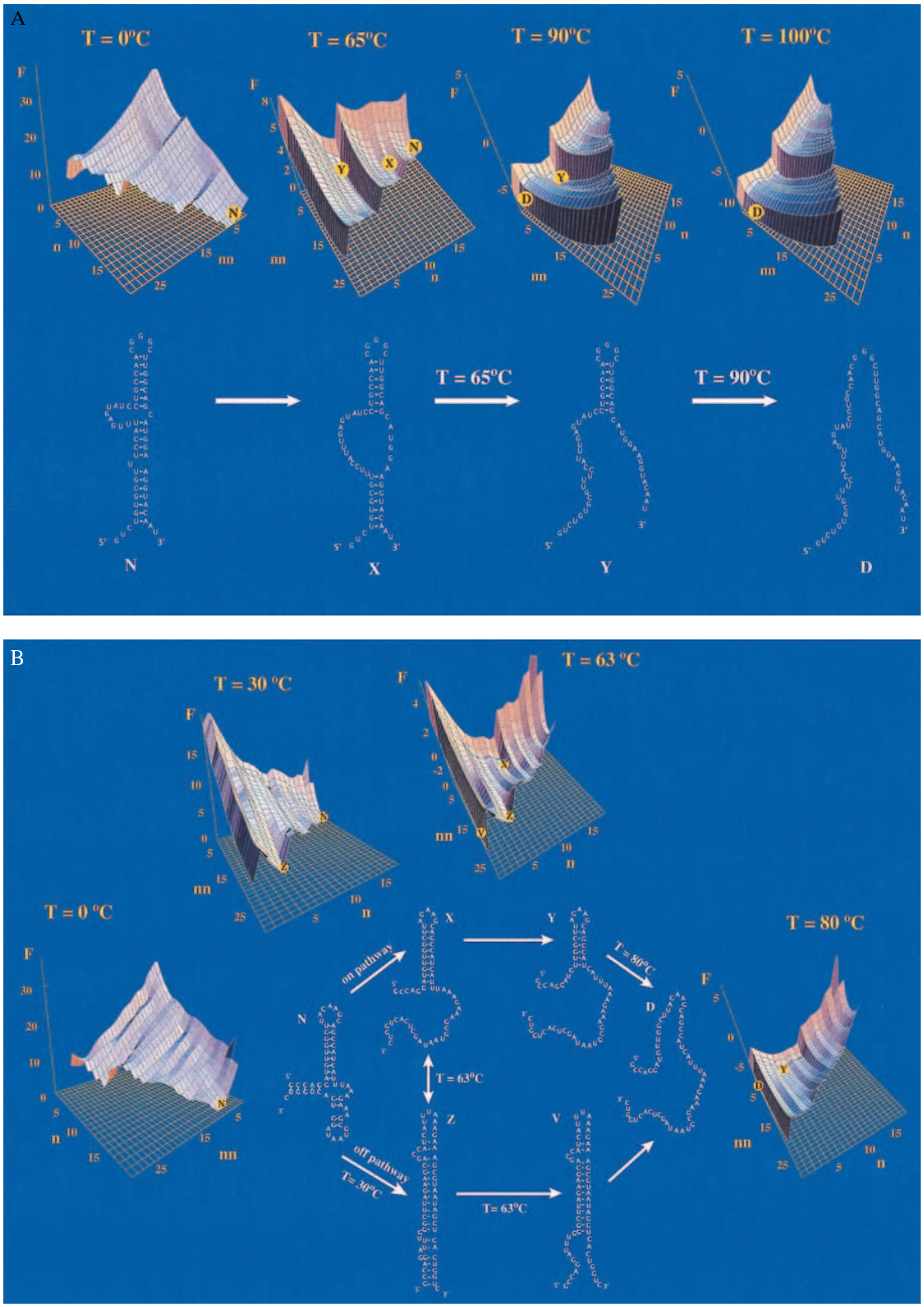


Fig. 3. The energy landscapes $F(n,nn)$ vs. Temperature T (in °C) for (*A*) mutant α mRNA fragment (Fig. 2*c*) and (*B*) *E. coli* 23S rRNA fragment (Fig. 2*d*). is the free energy for a state with n native contacts and nn non-native contacts, where the native and non-native contacts are defined according to the structure N. The free energies (in kcal/mol) are relative to the native states. Under native conditions, the number of non-native contacts need not equ because loops can bump into the chain in ways that involve no stable contact. Stable states are valleys, highlighted by the symbols on the energy land

- The computational pathway above becomes convoluted as we attempt to introduce pseudo-knots.

- An independent work that try this was also perform by Kopeikin and Chen.

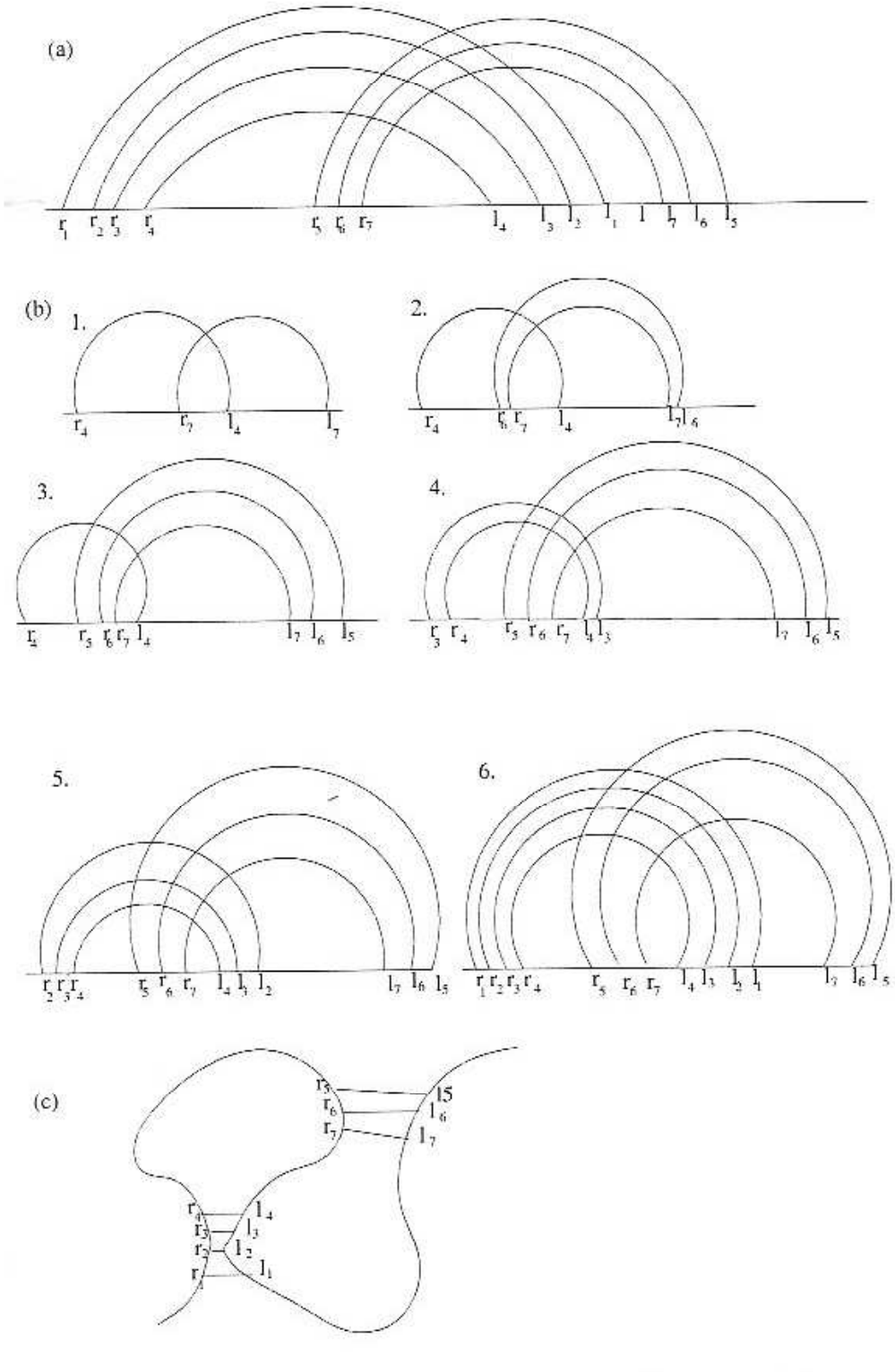
Our Findings

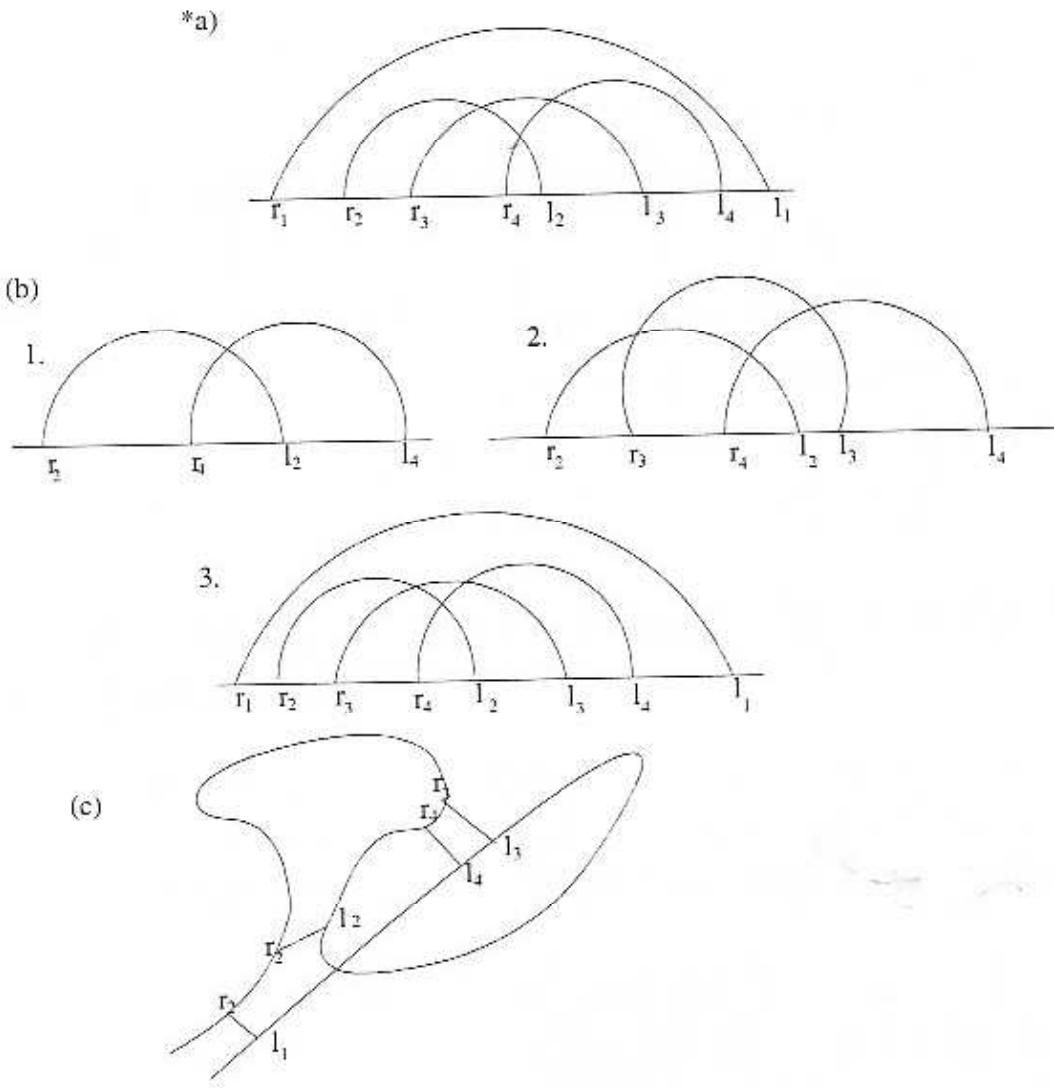
- Central challenge in the prediction of the RNA folding thermodynamics: **computation of the partition function**.

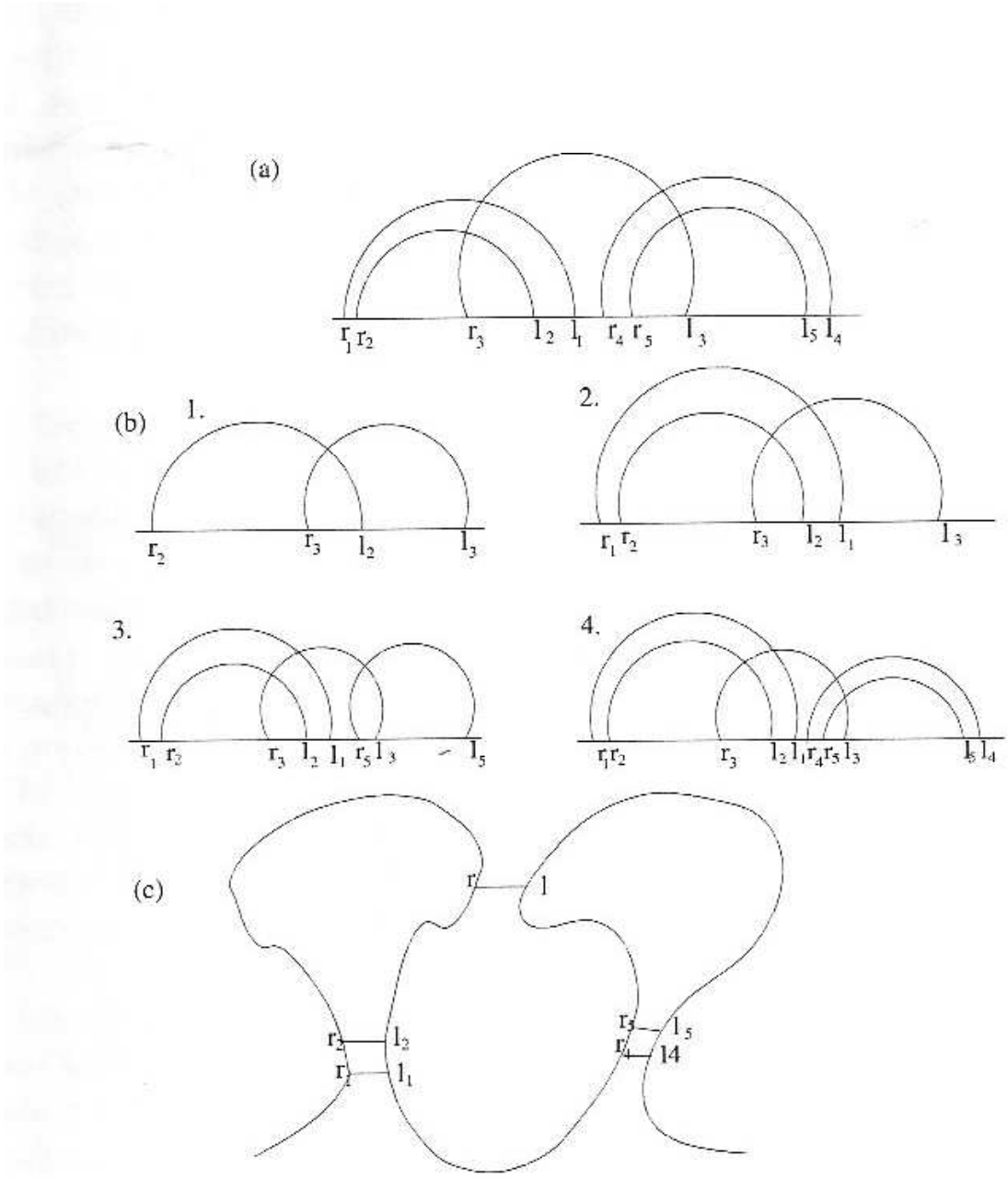
- Partition function can be computed either by summing over graphs, or by summing over energy levels.

- Both approaches include the sub-problem of graphs counting.

- We developed in this work, a breadth first searching algorithm for graph counting in the presence of crossing links.

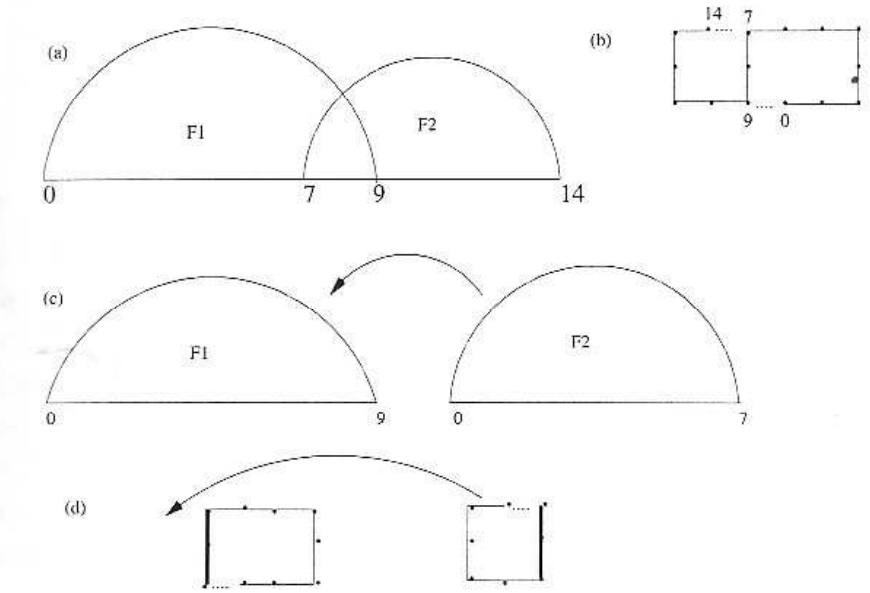

(a)
(b)
1.
2.
3.
4.
5.
6.
(c)


*a)
r1 r2 r3 r4 l2 l3 l4 l1
(b)
1.
r2 r1 l2 l1
2.
r2 r3 r4 l2 l3 l1
3.
r1 r2 r3 r4 l2 l3 l4 l1
(c)
r3
r4
l3
l4
l2
r2
r1
l1



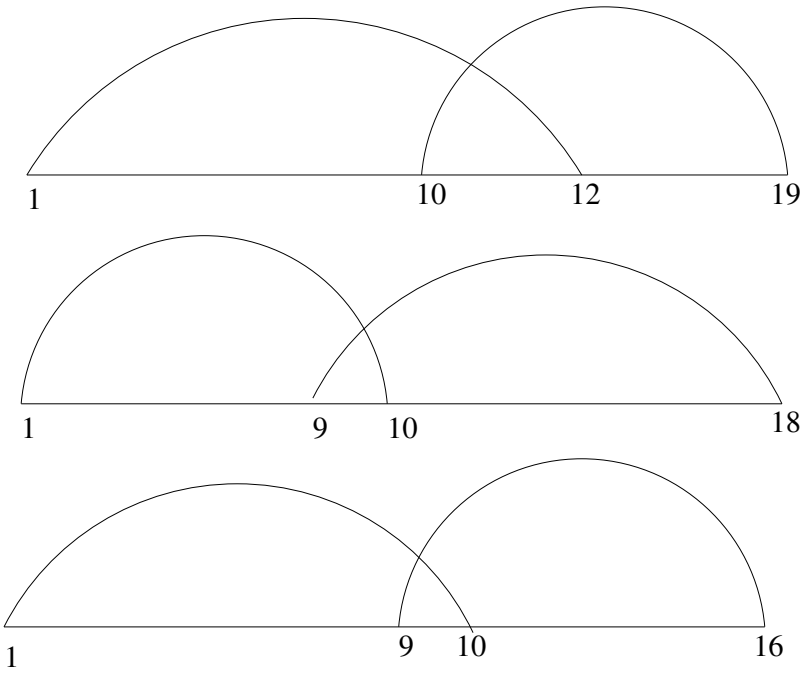
- We extend next the fire hose model of S-J. Chen and Ken Dill to a pair canal fire hose model.

- We then adapt Chen and Dill nearest neighbor model for combining the connectors in the fire hose.



- Application to conformations with two crossing links.

(a)

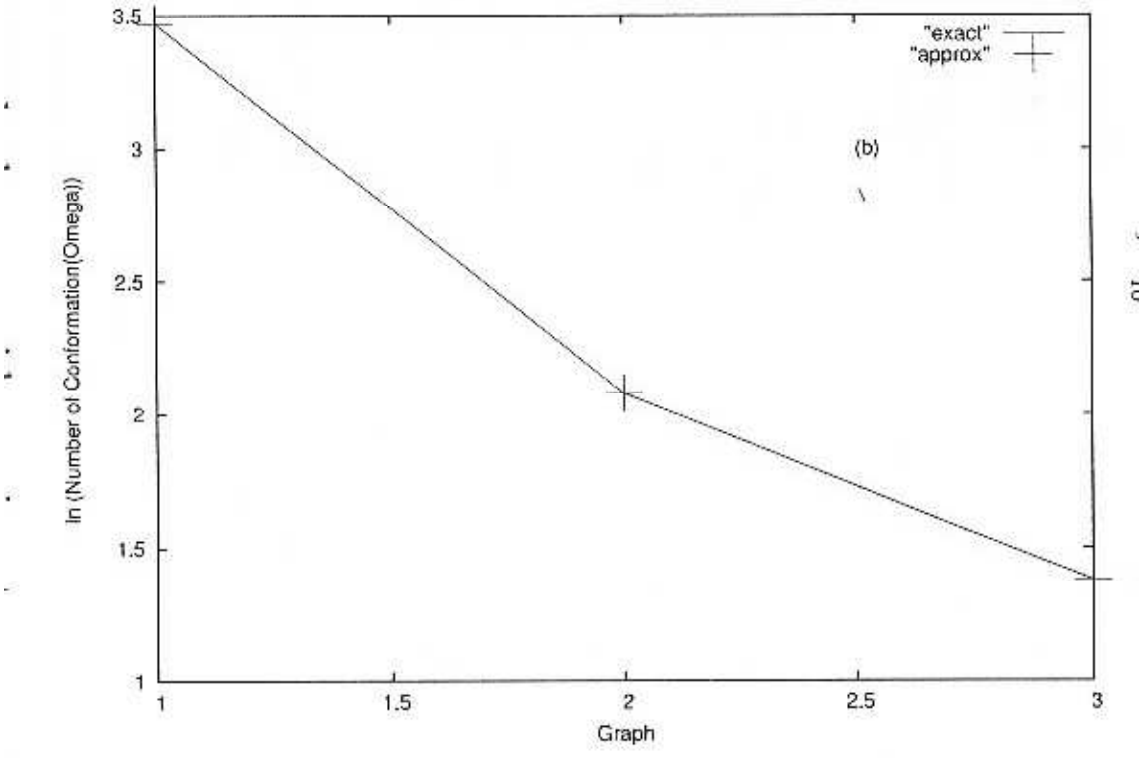


Graph

1.

2.

3.



Key: Our method prediction (indicated with +s) and the exact enumerating with a self avoiding walk (SAW) (indicated with solid line).

Conclusion

- We are still at the first step of the Ken Dill computational pathway, we enumerated above and
- target is to use this for structure-based drug design for malaria treatments.
- New work in this direction has recently been done by Cao and Chen.

Acknowledgment

- Covenant University,
- Adam Lucas,
- Ken Dill,
- J-S. Chen,
- Wen Zhang, and
- Timothy Ibiyemi.