Progress toward Predicting Viral RNA Structure from Sequence:
How Parallel Computing can Help Solve the RNA Folding Problem

Susan J. Schroeder
University of Oklahoma
October 7, 2008
We finished the genome map, now we can’t figure out how to fold it!

Science (1989) 243, p.786
RNA
Sequence → Structure → Function

Guanine Riboswitch

5’GGACAUAAUAU
CGCGUGGAUAUG
GCACGGAUGUUC
CUACCGGGGCACC
GUAAUGUCCGA
CUAUGUCCA

5’GCGGAUUUAG2M
CUCAGUDHUDP2GG
AGAGCGM2CCAGA
C0MUG0MAGYAU0PS
C5MUGGAG5MUC
C5MUGUGUSMUPSC
GA1MUCACAGAA
UUCGACCA

5’GGACAUAUAAU
CGCGUGGAUAUG
GCACGGAUGUUC
CUACCGGGGCACC
GUAAUGUCCGA
CUAUGUCCA

tRNA

Guanine Riboswitch
RNA Folding Problem

• Folding a polymer with negative charge
• Watson-Crick base pairing
• Hierarchical folding $1^\circ \rightarrow 2^\circ \rightarrow 3^\circ$

Figure from Dill & Chan (1997) Nat. Struct. Biol. Vol. 4, pp. 10-19
RNA Secondary Structure has Helices and Loops

Nussinov & Jacobson 1980 PNAS v 11 p 6310 Fig.1
An example of phylogenetic alignment and structure prediction for RNase P

Frank et al. 2000 RNA v 6 p1895
How Dynamic Programming Algorithms Calculate RNA Secondary Structure

• Stochastic context-free grammar defines possible base pairs as 1 of 4 possible cases

• Recursion statement finds maximum value for each small subset of RNA sequence

• Fill an array with scores for each substructure

• Traceback through the array to find the lowest free energy structure

• $O(N^2)$ memory storage

• $O(N^3)$ runtime
Websites for Folding Algorithms to Predict RNA 2°

Mfold
http://www.bioinfo.rpi.edu/applications/mfold

Vienna package
http://www.tbi/univie.ac.at/~ivo/RNA

RNAstructure
http://rna.urmc.rochester.edu

Sfold
http://sfold.wadsworth.org

PKNOTS
http://selab.wustl.edu

STAR4.4
http://biology.leidenuniv.nl/~batenburg/STRAbout.html
Representations of RNA Secondary Structure

dots & parentheses

tree

graph

merged landscape

Wuchty 2003, Nucl. Acids Res. v 31, p 1115 Fig. 7
Gruber et al. 2008, Nucl. Acids Res. v 36, p. W73 Fig. 1
RNAStructure Predicts Secondary Structure Well

<table>
<thead>
<tr>
<th>RNA</th>
<th>Lowest $\Delta G^o$ Structure</th>
<th>Best Suboptimal Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>73%</td>
<td>87%</td>
</tr>
<tr>
<td>Group II introns</td>
<td>88%</td>
<td>94%</td>
</tr>
<tr>
<td>tRNA</td>
<td>87%</td>
<td>97%</td>
</tr>
<tr>
<td>5 S rRNA</td>
<td>74%</td>
<td>96%</td>
</tr>
<tr>
<td>Group I introns</td>
<td>69%</td>
<td>84%</td>
</tr>
<tr>
<td>SRP RNA</td>
<td>66%</td>
<td>88%</td>
</tr>
<tr>
<td>Rnase P</td>
<td>63%</td>
<td>76%</td>
</tr>
<tr>
<td>23 S rRNA (as domains)</td>
<td>(74%)</td>
<td>(88%)</td>
</tr>
<tr>
<td>16 S rRNA (as domains)</td>
<td>(61%)</td>
<td>(76%)</td>
</tr>
</tbody>
</table>

Mfold and RNAStructure
Sample Suboptimal Structures

Lowest Free Energy Structure  Suboptimal Structure 1  Suboptimal Structure 2  Structure that Cannot be Predicted

Mathews, 2006, J. Mol. Biol. V359 p. 528 Fig. 2
How Wuchty’s Algorithm is like a Tree
STMV RNA folding problem

- Crystal structure to 1.8 Å resolution (Larson et al., 1998)
- 59% of the 1,058 RNA nucleotides are in helices
- RNA is icosahedrally averaged
- Identity of nucleotides in helices remains obscure
- Structure of 41% of the RNA remains unknown

Figure reproduced from VIPER website
Reddy et al. 2001
Current model for STMV RNA

Nussinov algorithms for maximizing matches and blocks

Nussinov et al. 1978 SIAM v35, p. 71,78 Fig. 1, 5
Combinatorial Search of STMV RNA

- Locate potential helical structures between pairing bases $i$ and $j$
- Assemble non-overlapping potential helices $(i,j)$ with $(p>j,q)$ or $(k>i+l, k+l<q<j)$
- Nested searching identifies “helices within helices”

- Over 144,000 perfect 6-pair helices, but no possible simultaneous combination of 30 helices in STMV RNA
Many more possible structures contain 30 imperfect helices in the STMV sequence
Chemical modification data restrains possible base pairing

5’AAGAUGU^AAA^CCAGGA3’
3’CUGCA AA GGUCUCU5’

5’AAGAUGU^AAA^CCAGGA3’
3’CUGCA AA GGUCUCU5’

5’AAGAUGU^AAA^CCAGGA3’
3’CUGCA A AA GGUCUCU5’

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU

5’AAGAUGU^AAA^CCAGGA
3’CUGCA AA GGUCUCU
Including Results from Chemical Probing Improves Secondary Structure Prediction

_E. coli_ 5S rRNA

Folded with constraints from _in vivo_ chemical modification

LOWEST 26.3 %
BEST 86.8 %

LOWEST 86.8 %
BEST 97.4 %

Mathews et al. (2004)  _Proc. Natl. Acad. Sci._ 101, 7290 Figure 1
3 Restraints Can Change the Lowest Energy Fold

Native -341.1 kcal/mol

A527, A532, A537 restrained to be single stranded -341.0 kcal/mol
Free Energy Landscape of STMV RNA

- MFE
- Wuchty
- Zuker
- Chemical data
- Future data
- # Helices>1mm
- # Helices,1 mm
- Goal

$\Delta G$
How can Parallel Computing Help Solve the RNA Folding Problem?

- Utilize tree structure of RNA secondary structure prediction
- Expand range of free energies that can be computed for an RNA free energy landscape
- Explore more possible RNA structures
Acknowledgements

- Deb Mathews, University California Riverside
- David Mathews, University of Rochester
- Jeanmarie Verchot-Lubicz, Oklahoma State University
- Cal Lemke, Oklahoma University greenhouses

- Lab Members:
  Xiaobo Gu, Steven Harris, Koree Clanton-Arrowood, Brina Gendhar, Shelly Sedberry, Brian Doherty, Ted Gibbons, Sean Lavelle, John McGurk, Becky Myers, Mai-Thao Nguyen, Samantha Seaton, Jon Stone

- Funding