DNA, RNA structure: views through an algorithmic lens

Anne Condon
Computer Science Department
U. British Columbia
DNA, RNA structure: views through an algorithmic lens
“RNA, long upstaged by its more glamorous sibling, is turning out to have star qualities of its own” – J. Couzin, Science, 2002
DNA, RNA structure is vital to function

transfer RNA’s cloverleaf structure aids in the synthesis of proteins from messenger RNA

Stryer, *Biochemistry*
DNA, RNA structure is vital to function

some important RNA structures are not branched

Chen et al. JMB 1996
DNA, RNA structure is vital to function

some important RNA structures are not branched

Chang & Tinoco
JMB 1997
DNA, RNA structure is vital to function

some molecules can have two stable structures, with different functions

Schultes and Bartel

*Nature Reviews | Molecular Cell Biology*
DNA, RNA structure is vital to function

large RNA structures, such as this 16sRNA, contain few pseudoknots
DNA, RNA structure is vital to function

RNA shows promise as a therapeutic agent

Sullenger and Gilboa, Nature Insight, 2002
DNA, RNA structure is vital to function

RNA shows promise as a therapeutic agent

... this RNA aptamer blocks HIV replication in cells

Chaloin, L. et al. NAR 2002
DNA, RNA structure is vital to function

DNA is ideal for nanoscale construction

H. Yan et al.
Science, 2003
DNA, RNA structure is vital to function

DNA is ideal for nanoscale construction

H. Yan et al.
Science, 2003
DNA, RNA structure is vital to function

DNA is ideal for nanoscale construction:

“... rather than examining in detail what occurs in nature, we take the engineering approach of asking, what can we build?” E. Winfree

Rothemund et al.
JACS 2003
“There are strong reasons to conclude that DNA and protein based life was preceded by a simpler life form based primarily on RNA” – Gerald Joyce

DNA, RNA structure is vital to function

- 4.5 billion years ago: earth!
- 3.8 billion years ago: RNA world
- 3.6 billion years ago: DNA, proteins
- 0 billion years ago: life today
efficient, high-quality algorithms for nucleic acid structure prediction and design are useful!
“... the primary sequence determines the secondary structure which, in turn, determines its tertiary folding, whose formation alters only minimally the secondary structure” - Tinoco & Bustamente, JMB 1999
pseudoknots and hierarchical folding

• classification of pseudoknots differ:
  – “which bases [pair] to which other bases describes the secondary structure” - Aalberts & Hodas, NAR 2005
  – “pseudoknots are elements of RNA tertiary structure” - Hermann & Patel, JMB 1999

• the tertiary structure classification follows from the hierarchical folding hypothesis, which posits that pseudoknots form subsequent to formation of pseudoknot free structure
secondary structure representation

pseudoknotted

pseudoknot free
rationale:
• a structure is composed of loops
• the free energy of a structure is the sum of its loop energies*
• a strand folds into the structure with minimum free energy (mfe)

*under fixed conditions, relative to reference structure
prediction from the base sequence

rationale:

• a structure is composed of loops
• the free energy of a structure is the sum of its loop energies
• a strand folds into the structure with *minimum free energy* (mfe)

- hairpin loops
- internal loops
- multi-loops
prediction from the base sequence

rationale:

• a structure is composed of loops

• the free energy of a structure is the sum of its loop energies

• a strand folds into the structure with *minimum free energy (mfe)*

hairpin loops  internal loops  multi-loops  pseudo-loops
prediction from the base sequence

rationale:

- A structure is composed of loops.
- The free energy of a structure is the sum of its loop energies.
- A strand folds into the structure with minimum free energy (mfe).

hairpin loops  internal loops  multi-loops  pseudo-loops
prediction from the base sequence

rationale:

• a structure is composed of loops
• the free energy of a structure is the sum of its loop energies
• a strand folds into the structure with minimum free energy (mfe)

energy of loop depends on closing base pairs, number of unpaired bases, unpaired bases in loop adjacent to closing base pairs, whether or not closing base pairs cross other base pairs
prediction from the base sequence

- mfe prediction is NP-hard
  - Lyngso & Pederson; Akutsu 2000
- yet, many algorithms, ranging in complexity and generality

  Jacobson & Nussinov 1980
  Zuker & Steigler (Z&S) 1981
  Lyngso & Pederson 1999
  Rivas & Eddy (R&E) 1999
  Uemura et al. 1999
  Akutsu 2000
  Dirks and Pierce (D&P) 2003
complexity versus generality

- R&E: $O(n^6)$
- D&P: $O(n^5)$
- Z&S: $O(n^3)$
on algorithm generality: Z&S

\[ \begin{bmatrix} 0 \\ \vdots \end{bmatrix} = \text{min} \]

- pseudoknot free structures
on algorithm generality: R&E

\[ \begin{bmatrix} 0 \end{bmatrix} = \min \begin{bmatrix} \text{shapes} \end{bmatrix} \]

\[ \begin{bmatrix} \text{shapes} \end{bmatrix} = \min \begin{bmatrix} \text{shapes} \end{bmatrix} \]
on algorithm generality: R&E

\[
\begin{bmatrix}
0 & x & y & z \\
\end{bmatrix}
\]

= min

\[
\begin{bmatrix}
x & y & z \\
\end{bmatrix}
\]

= min

\[
\begin{bmatrix}
x & y & z \\
\end{bmatrix}
\]
on algorithm generality: R&E

• "we still lack a systematic \textit{a priori} characterization of the class of configurations that this algorithm can solve" (Rivas & Eddy)
R&E structures pass the following test:
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs

as long as possible

if structure is empty then accept
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
   remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
  remove nested, entwined,
  dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

```
repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
```
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

repeat
    remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept
R&E structures pass the following test:

\[
\text{repeat}
\]

\[
\text{remove } \text{nested, entwined, dangling, or empty arcs}
\]

\[
\text{as long as possible}
\]

\[
\text{if structure is empty then accept}
\]
R&E structures pass the following test:

repeat
  remove nested, entwined, dangling, or empty arcs
as long as possible
if structure is empty then accept

a non-R&E structure
### complexity versus generality

<table>
<thead>
<tr>
<th></th>
<th>Gut</th>
<th>RCSB</th>
<th>RNase</th>
<th>tmRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="#">#Str</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z&amp;S(^{(pkfree)})</td>
<td>O(n^3)</td>
<td>48</td>
<td>231</td>
<td>30</td>
</tr>
<tr>
<td>D&amp;P</td>
<td>O(n^5)</td>
<td>354</td>
<td>244</td>
<td>95</td>
</tr>
<tr>
<td>Akutsu</td>
<td>O(n^5)</td>
<td>354</td>
<td>246</td>
<td>95</td>
</tr>
<tr>
<td>R&amp;E</td>
<td>O(n^6)</td>
<td>369</td>
<td>274</td>
<td>468</td>
</tr>
</tbody>
</table>
complexity versus generality

- Complexity:
  - $O(n^6)$
  - $O(n^5)$
  - $O(n^3)$

- Generality:
  - Z&S: pseudoknot free structures
  - D&P: simple pseudoknots but not kissing hairpins
  - R&E: almost all structures
  - Uemura et al: kissing hairpins
what about hierarchical folding?

- note that none of the algorithms for pseudoknotted prediction are consistent with the hierarchical folding hypothesis
what about hierarchical folding?

• given a pkfree structure $S$, we can find, in $O(n^3)$ time, the pkfree structure $S'$ which, given $S$, has minimum energy
another dimension: accuracy

accuracy

complexity

73% (Mathews et al.)

generality
many possible approaches:
  • improve energy model
  • model co-transcriptional folding
  • revisit hierarchical folding hypothesis
  • ...

The diagram shows a three-dimensional space with axes for complexity, generality, and accuracy. A point at 73% accuracy is marked on the accuracy axis.
improving accuracy

RNA SSTRAND

[ Home | Search | Analysis | Submit structures | Help | About ]

The RNA Secondary Structure and Statistical Analysis Database

RNA SSTRAND contains known secondary structures of any type and organism. The ultimate goal of this database is to incorporate a comprehensive collection of known secondary structures, and to provide the scientific community with simple yet powerful ways of analysing, searching and updating the proposed database.

Search
Search for RNA SSTRAND entries, supports multiple search criteria

Analysis
Analyse one or a group of RNA secondary structures

Submit
Submit new RNA secondary structures to RNA SSTRAND

Help
Brief explanations of RNA SSTRAND input and output fields, also accessible via the ‘?’ links on any RNA SSTRAND page

About
Information about the database, its current holdings and conditions of use

Search RNA SSTRAND
RNA SSTRAND ID: [ ]
Search

Provenance of RNA SSTRAND structures:

<table>
<thead>
<tr>
<th>ID</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>588</td>
<td>Gutell Lab CRW</td>
</tr>
<tr>
<td>279</td>
<td>RCSB Protein Data Bank</td>
</tr>
<tr>
<td>1430</td>
<td>Rfam Database</td>
</tr>
<tr>
<td>468</td>
<td>RNase P Database</td>
</tr>
<tr>
<td>570</td>
<td>Sprinzl tRNA Database</td>
</tr>
<tr>
<td>6</td>
<td>tmRNA database</td>
</tr>
<tr>
<td>15</td>
<td>Other Sources</td>
</tr>
</tbody>
</table>

We hope that you find RNA SSTRAND useful. If you use the database or analysis provided on this website in your research, we ask you to acknowledge this by including the RNA SSTRAND URL: www.masoft.ca/sstrand, and by citing:

improving accuracy

Accuracy vs. length, molecules with pseudoknots, non-canonical bp broken
improving accuracy

- current energy model has 7,576 parameters, most (7,260) for small internal loops
  - some parameters experimentally determined; highly reliable
  - some extrapolated from energy rules and fit with known structures
- model has been refined over two decades, in a somewhat ad-hoc way
- can be simplified to a model with approx. 350 parameters
improving accuracy

True structure

Predicted structure
improving accuracy

True structure

Predicted structure
improving accuracy

True structure

Predicted structure
improving accuracy

• we’ve built a dataset that includes
  – 795 strands + mfe structures with experimentally measured free energies
  – 3,248 strand fragments with known mfe structures

• we’re using maximum likelihood, and constraint-based, approaches to determine best choice of parameters

• preliminary results on small dataset are encouraging (9% improvement in accuracy)
in closing

• accurate DNA, RNA secondary structure prediction poses many fun algorithmic and modeling challenges
  – many variants: predicting folding pathways, partition function, complexes composed of multiple molecules, ...

• RNA tertiary structure prediction from the base sequence is uncharted terrain!

• another major challenge is design of molecules that have desired structural or functional properties
thank you!

thanks also to:

– Mirela Andronescu, Beth Davy, Holger Hoos, David Mathews, Baharak Rastagari, Finbarr Tarrant, and Shelly Zhao who collaborated on this work
– NSERC, MITACS for their generous funding