Sequence-Structure Relations of Single RNA Molecules and Cofolded RNA Complexes

Ulrike Mückstein

Institute for Theoretical Chemistry
University of Vienna

Vienna, 2005
Outline

Motivation

Folding RNA

RNA sequence - structure relations

Results

RNA sequence to structure mapping
Co-folding two RNA molecules

Discussion
RNA primary, secondary and tertiary structure
Outline

Motivation
- Folding RNA
  RNA sequence - structure relations

Results
- RNA sequence to structure mapping
- Co-folding two RNA molecules

Discussion
Relation between RNA sequences and their secondary structures

- RNA sequence to structure mapping
- Evolutionary dynamics of an RNA population in a flow reactor
- RNA-RNA interaction by co-folding two RNA molecules
RNA sequence to structure mapping
Evolutionary dynamics in a flow reactor
Evolutionary dynamics in a flow reactor

kind of tansition  continuous  discontinuous  distance

| Ω0 | 0   | 0.01 |
|Ω1 | 2   | 0.001 |
|Ω2 | 2   | 0.0001 |
|Ω3 | 2   |
|Ω4 | 2   |
|Ω5 | 2   |
|Ω6 | 2   |

shift

0  1  2  3  4  5  6  7

double flip

8  9  10  11  12  13  14  15

shift

16  17  18  19  20  21  22  23

flip

24  25  26  27  28  29  30  31

shift

32  33  34  35  36  37  38  39

start
Co-folding two RNA molecules
Co-folding two RNA molecules

[Diagram showing folding of RNA molecules]

J03132_CD54-1622
site 1 (1623 - 1641)

Positions in sequence

0

0.2

0.4

0.6

0.8

1

Probabilities

-27

-26

-25

-24

-23

-22

-21

-20

-19

-18

-17

-16

-15

-14

-13

-12

-11

-10

-9

-8

-7

-6

-5

-4

-3

-2

-1

0

Delta G

RNAup

Position in sequence

1610 1620 1630 1640 1650
Outline

Motivation
  Folding RNA
  RNA sequence - structure relations

Results
  RNA sequence to structure mapping
  Co-folding two RNA molecules

Discussion
Generic properties of RNA folding

- More sequences than structures
- Few common and many rare structures
- Common structures form extended neutral networks
- Shape space covering
Exhaustive folding and enumeration of the sequence spaces $I_{AUGC}^{(\ell=9)}$ and $I_{AUGC}^{(\ell=10)}$

PostgreSQL Database

- generate all sequences of length $= 9$
- and alphabet $\{A,U,G,C\}$
  - (262144)
- fold sequences
- decomposition into components
- insert sequences with stable minimal free energy structure sorted by structure
  - (3280 ~ 2%)

Hamming distance or base pair substitution

- structure
- G A C C U A G U C
- G A C C A A G U C

Stored Procedures

- Vigl Graph Library
- Vienna RNA package

Vienna RNA package

- decompose into components
- generate all sequences of length $= 9$
- and alphabet $\{A,U,G,C\}$
  - (262144)
- fold sequences
- insertion of sequences with stable minimal free energy structure sorted by structure
  - (3280 ~ 2%)
More sequences than structures

\[ 3^\ell \text{ structures } < 4^\ell \text{ sequences} \]

\(( ( ( \ldots ) ) )\) GCGAUGCGC

\(\mathcal{I}_{\text{AUGC}}^{(\ell=9)}\)
- stable mfe structure
  \(3239 \sim 2.3\%\)
- 4 different structures

\(\mathcal{I}_{\text{AUGC}}^{(\ell=10)}\)
- stable mfe structure
  \(40345 \sim 3.8\%\)
- 9 different structures
Few common and many rare structures

\[ I^{(\ell=9)}_{\text{AUGC}} \]

\[ I^{(\ell=10)}_{\text{AUGC}} \]

73.2% 15.5% 9.9% 1.4%
22.8% 20.6% 19.7% 16.7%
7.4% 4.7% 4.1% 2.0% 1.9%
Common structures form extended neutral networks

(((.....)))  73.2%

..(((.....)))  15.5%

.......(()(()))  9.9%

(......)  1.4%
Shape space covering

(((.....)).
.
(((((....))).)
((.....))

hamming distance
G A C C U A G U C
G A C C A A G U C

structure and hamming distance or base pair substitution

\[
\begin{align*}
C &\rightarrow G \\
A &\rightarrow U \\
A &\rightarrow G \\
G &\rightarrow C \\
\end{align*}
\]
Discussion

- topology of RNA sequence to structure mapping by exhaustive folding and enumeration of sequence spaces

- extended neutral networks allow a population of RNA molecules to explore sequence space by neutral drift

- shape space covering reduces the search space
Discussion

- topology of RNA sequence to structure mapping by exhaustive folding and enumeration of sequence spaces

- extended neutral networks allow a population of RNA molecules to explore sequence space by neutral drift

- shape space covering reduces the search space
Discussion

- topology of RNA sequence to structure mapping by exhaustive folding and enumeration of sequence spaces

- extended neutral networks allow a population of RNA molecules to explore sequence space by neutral drift

- shape space covering reduces the search space
Outlook

- functional RNAs work by interacting with other RNAs

- Camille Stephan-Otto Attolini et al. studied the co-folding map of interacting RNA sequences:
  - the co-folding map admits large neutral networks and long neutral paths similar to the folding map of single RNA molecules

- tool to compute RNA-RNA interaction by co-folding two RNA molecules
Outlook

- functional RNAs work by interacting with other RNAs

- Camille Stephan-Otto Attolini et al. studied the co-folding map of interacting RNA sequences:
  - the co-folding map admits large neutral networks and long neutral paths similar to the folding map of single RNA molecules

- tool to compute RNA-RNA interaction by co-folding two RNA molecules
Outlook

- functional RNAs work by interacting with other RNAs

- Camille Stephan-Otto Attolini et al. studied the co-folding map of interacting RNA sequences:
  - the co-folding map admits large neutral networks and long neutral paths similar to the folding map of single RNA molecules

- tool to compute RNA-RNA interaction by co-folding two RNA molecules
Outlook

- functional RNAs work by interacting with other RNAs

- Camille Stephan-Otto Attolini et al. studied the co-folding map of interacting RNA sequences:
  - the co-folding map admits large neutral networks and long neutral paths similar to the folding map of single RNA molecules

- tool to compute RNA-RNA interaction by co-folding two RNA molecules
Outline

Motivation
Folding RNA
RNA sequence - structure relations

Results
RNA sequence to structure mapping
Co-folding two RNA molecules

Discussion
Loop decomposition of RNA Secondary Structure

\[ F(S) = \sum_{L \in S} F_L. \]
Basic Algorithms

- McCaskill: Dynamic programming algorithm for calculation of the full equilibrium partition function $Z$ for RNA secondary structure in $O(N^3)$

$$Z = \sum_S e^{-[F(S)/kT]}$$

- Vienna RNA Package
Basic Algorithms

- McCaskill: Dynamic programming algorithm for calculation of the full equilibrium partition function $Z$ for RNA secondary structure in $O(N^3)$

$$Z = \sum_S e^{-[F(S)/kT]}$$

- Vienna RNA Package
Complete ensemble of secondary structures

Probability of an unstructured region
Complete ensemble of secondary structures

Probability of an unstructured region

Probability of interaction in unstructured regions
Information provided by our algorithm

- structural context of a binding site
  - $P_u[i, j]$ region $[i, j]$ contains no secondary structure

- location of possible binding sites
  - $P^*[i, j]$ probability of a regional interaction

- energetics of RNA-RNA interaction
  - $\Delta G = \Delta G_u + \Delta G_h$
Information provided by our algorithm

- structural context of a binding site
  - \( P_u[i, j] \) region \([i, j]\) contains no secondary structure

- location of possible binding sites
  - \( P^*[i, j] \) probability of a regional interaction

- energetics of RNA-RNA interaction
  - \( \Delta G = \Delta G_u + \Delta G_h \)
Information provided by our algorithm

- structural context of a binding site
  - $P_u[i, j]$ region $[i, j]$ contains no secondary structure

- location of possible binding sites
  - $P^*[i, j]$ probability of a regional interaction

- energetics of RNA-RNA interaction
  - $\Delta G = \Delta G_u + \Delta G_h$
$P_u[i, j]$ region $[i, j]$ contains no secondary structure

\[
P_u[i, j] = \underbrace{Z[1, i-1] \times 1 \times Z[j+1, N]}_{\text{exterior}} + \sum_{p<i<j<q} P_{pq} \frac{Z_{pq}[i, j]}{Z^b[p, q]} .
\]
Interaction of region \([i, j]\) with region \([i^*, j^*]\)

\[
Z^l[i, j, i^*, j^*] = \sum_{i<k<j\atop i^*>k^*>j^*} Z^l[i, k, i^*, k^*] e^{-\frac{1}{kT} l(k, k^*; j, j^*)}.
\]
conditional probability given that two molecules bind at all

\[ P^*[i, j] = \frac{Z^*[i, j]}{\sum_{k<l} Z^*[k, l]} . \]

optimal free energy of a regional interaction

\[ \Delta G_i = \min_{k \leq i \leq l} \{ \Delta G[k, l] \} . \]

\[ \Delta G[i, j] = (-kT) \ln Z^*[i, j] . \]
probability and optimal free energy of an interaction

- Conditional probability given that two molecules bind at all

\[ P^*[i, j] = \frac{Z^*[i, j]}{\sum_{k < l} Z^*[k, l]} \]

\[ Z^*[i, j] = P_u[i, j] \sum_{i^* > j^*} Z'[i, j, i^*, j^*]. \]

- Optimal free energy of a regional interaction

\[ \Delta G_i = \min_{k \leq i \leq l} \{ \Delta G[k, l] \}. \]

\[ \Delta G[i, j] = (-kT) \ln Z^*[i, j]. \]
Interaction of CD54 mRNA with siRNA si1622

Interaction of CD54 mRNA with siRNA si1622

Interaction of CD54 mRNA with siRNA si1622

Local RNA target structure influences siRNA efficacy

Discussion

- Realistic description of interaction between small RNAs and their targets
- Structural context of a possible binding site
- Location of possible binding sites
- Energetics of RNA-RNA interaction
- Time complexity is $O(N^3)$, memory requirement $O(N^2)$
Discussion

- Realistic description of interaction between small RNAs and their targets
- Structural context of a possible binding site
- Location of possible binding sites
- Energetics of RNA-RNA interaction
- Time complexity is $O(N^3)$, memory requirement $O(N^2)$
Discussion

- Realistic description of interaction between small RNAs and their targets
- Structural context of a possible binding site
- Location of possible binding sites
- Energetics of RNA-RNA interaction
- Time complexity is $O(N^3)$, memory requirement $O(N^2)$
Discussion

- Realistic description of interaction between small RNAs and their targets
- Structural context of a possible binding site
- Location of possible binding sites
- Energetics of RNA-RNA interaction
- Time complexity is $O(N^3)$, memory requirement $O(N^2)$
Discussion

- Realistic description of interaction between small RNAs and their targets
- Structural context of a possible binding site
- Location of possible binding sites
- Energetics of RNA-RNA interaction
- Time complexity is $O(N^3)$, memory requirement $O(N^2)$