COMP 564: Pseudo-Knots and RNA-RNA interactions prediction

Jérôme Waldispühl
School of Computer Science
McGill University
Figure 2. Sequences and Structures of RNA Pseudoknots

http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.0030213
Figure 1. RNA Pseudoknot Architecture

http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.0030213
Pseudo-knot decomposition

(Rivas & Eddy, 1999)
Coverage

ABAB

ABCABC

ABACBC

ABCADBECEDE

ABACBDCD
Coverage

- ABAB: OK
- ABCABC: OK
- ABACBC: OK
- ABCADBECE: NO
- ABACBDCD: OK
Challenges

- Empirical values for thermodynamic parameters of pseudo-knots.
- Running time in $O(n^6)$, Space requirement in $O(n^4)$
- Extension partition function computation and rational sampling.
- Problem is NP-Hard if we consider stacking base pairs.
Results

Performance match MFOLD (The space of conformations increases drastically, and no wrong pseudo-knots were predicted).

tRNAs: 24 consistent predictions (cloverleaf shape) over 25 sequences, while MFOLD identifies only 19 cloverleafs.

HIV-1-RT ligand: 54 (over 63) predictions agreed with the pseudo-knoted structure found by comparative analysis (MFOLD could find only one the stems).

viral RNAs: 6 pseudo-knot structures identified (over 7 known).

(Rivas & Eddy, 1999)
Software for Pseudo-knot prediction

- Pknots (Rivas & Eddy, 1999)
- KineFold (Xayaphoummine et al., 2003)
- PknotsRG (Reeder et al., 2007)
- pKiss (Theis et al., 2010)
- IPknot (Sato et al., 2011)
Motivation

- Experimental and bioinformatical methods find novel ncRNAs *en masse*
- Give no hint as to the function of these novel ncRNAs
- Functional characterization of ncRNAs is difficult and slow
- Most ncRNAs function through interaction with other RNAs
- Identification of interaction partners is the easiest approach to learn about possible functions
- Most obvious in the case of miRNA target prediction
Well known Examples of RNA-RNA Interaction

- micro RNAs regulate mRNA translation
- snoRNAs guide methylation and pseudouridylation of rRNA
- some well studied bacterial examples
  - RyhB is transcribed under low Fe, binds several mRNA of Fe binding proteins (sdh, sodB) and leads to mRNA degradation
  - GadY interacts with the 3’ UTR of GadX and inhibits its degradation
  - DsrA is expressed at low temperatures and stimulates the translation of RpoS a translational regulator
  - OxyS is expressed under oxidative stress and inhibits translation of its targets RpoS and flhA
  - T-box motifs bind uncharged tRNAs to control transcription of aminoacyl synthetases
Interaction of OxyS and fhla

Binding of OxyS to fhla mRNA makes the ribosome binding site (start codon) inaccessible
Transcriptional control by T-box Motifs

Concentration of un-charged tRNAs controls transcription of its aminoacyl synthetase
Challenges

- Few well-studied examples
- Energetics of many interaction motifs are unknown
- Length of the interacting region is often quite small
- Binding is a concentration dependent process
- Folding kinetics rather than thermodynamics may play a role
- A single small RNA may have many targets
- RNA chaperones such as Hfq may be required for binding
- ncRNAs often act within RNPs, what’s the influence of the protein?
Overview of Prediction Strategies

- Co-folding by concatenation of two sequences, e.g. RNAcofold, pairfold, DINAMELT, Nupack
- Co-folding with pseudoknot-like structures, IRIS
- Using only inter-molecular interaction, i.e. assume that both molecules are unstructured by themselves. RNAhybrid, RNAduplex, codeRNAplex
- Combine interaction search with accessibility calculations. RNAup, RNAplfold + RNAplex, oligowalk
Simple Co-folding of two RNAs

- Poor man’s approach to cofolding:
  - Concatenate two RNAs using a short linker
  - Use conventional folding programs such as mfold

- Proper way:
  - Use modified folding algorithm that keeps track of the break between the strands
  - Any loop containing the break point is treated specially.
  - Implemented in the RNAcofold program of the Vienna RNA package

- Limited to structures that are pseudo-knot free for concatenated sequences.
Pair Probabilities from RNAcofold
Concentration Dependence of RNA-RNA interactions

Binding processes are always concentration dependent.
For two RNAs we have three reactions in equilibrium:

\[ A + B \rightleftharpoons AB \quad A + A \rightleftharpoons AA \quad B + B \rightleftharpoons BB \]

Compute concentrations of all five monomers and dimers.
UNAFold: prediction of RNA/DNA hybridization
(Dimitrov&Zuker,2004)

Motivation:
Let A and B be two polynucleotide sequences. In solution, UNAFold aims to predict the concentration of single stranded folded and unfolded A and B AND hybridization AA, BB and AB.

Principles:
• Simple modification of the McCaskill’s algorithm.
• Stacking energies computed from experimental measures.

Results:
Reproduce experimental observations
Sfold: Accessibility prediction through Boltzmann sampling (Ding & Lawrence, 2001)

Sample secondary structures using a stochastic backtracking procedure:

**Principle:**
- Estimate accessibility (not base paired) of each nucleotide in the sample set.
- Identify the hybridization regions.
Structures (not) Predicted by RNAcofold

knot-free  pseudo-knotted
Predicting more complex Structures

Without restricting allowed structure motif RNA-RNA interaction is NP-complete

- The most general algorithms (Alkan 2006, Pervouchine 2004) allow structures where
  - Intra-molecular pairs form pseudo-knot free structures
  - Inter-molecular pairs are not allowed to cross
- Run time is too slow for most purposes ($O(n^3 \cdot m^3)$)
Fast Interaction Search

Methods for fast interaction search

- Search for sequence complementarity by BLAST
- Better: Interaction search using thermodynamics
- Simplified folding algorithm without intra-molecular pairs.
- Runs in $O(n \cdot m)$ time.
- Used in RNAhybrid (miRNA target prediction), RNAduplex, RNAplex

What's the effect of neglecting intra-molecular structure?
RNA-mRNA interaction interaction energies (from RNAduplex)
red: ncRNA candidates from RNAz, grey: shuffled sequences.
Enrichments relative to randomly chosen conserved regions:
I: 2.3, II: 1.9, III: 1.4, IV: 1.1
Combining Interaction and Accessibility

Two ingredients for efficient hybridization
- Complementarity
- Accessibility

How to quantify these?
Complementarity $\rightarrow$ interaction energy
Accessibility $\rightarrow$ probability to be unpaired
RNA Hybridization as a two Step Process
Example: ompN and RybB

GCCAC-------TGCTTTTTCTTGTGCCCCATTTT-GTGGA--------GC-CCATCAACCCGCCATTTCCGTT---CAAG-GTTGTTGGGTGTTTTTT

MFE -38.2 kcal/mol
Cost of opening 23.6 kcal/mol

AGGTCAAACAACGC-AGAAACAATATT---TAAGTGCCGCACACGACGCGCGTGTCGTT-GGCTTCGCCCCTACTGTTACCGTTATGAAAAGAAACC-3′

-24 kcal/mol
Example: ompN and RybB

\[ \Delta G_{\text{open}} = 1.6 + 3.9 \text{ kcal/mol}, \quad \Delta \Delta G = -16 \text{ kcal/mol} \]
The RNAup Approach

- Compute probability that a site at \([i..j]\) is unpaired (equivalent to the energy \(\Delta G_{\text{open}}\) needed to force it open).
- Consider all possible ways of binding to the region \([i..j]\) to compute the interaction energy \(\Delta G_{\text{interact}}\).
- Total binding energy is the sum of these contributions: \(\Delta \Delta G = \Delta G_{\text{open}} + \Delta G_{\text{interact}}\).
- Currently, restrict interactions to a single region.
Computing Accessibility

\( \Delta G_{\text{open}} \) is equivalent to the probability that the region \([i..j]\) is unpaired in equilibrium

\[
\Delta G_{\text{open}} = -RT \ln P^u[i,j]
\]

- Constrained folding \( \Delta G_{\text{open}} = \Delta G^{\text{constr}} - \Delta G^{\text{free}} \)
- Boltzmann sampling, works for short regions only
- Direct computation by modified folding algorithm
Computing Accessibility

\( \Delta G_{\text{open}} \) is equivalent to the probability that the region \([i..j]\) is unpaired in equilibrium \( \Delta G_{\text{open}} = -RT \ln P^u[i,j] \)

- Constrained folding \( \Delta G_{\text{open}} = \Delta G^{\text{constr}} - \Delta G^{\text{free}} \)
- Boltzmann sampling, works for short regions only
- Direct computation by modified folding algorithm

\[
P^u[i,j] = \frac{Z_{1,i-1} Z_{j+1,n}}{Z_n} + \sum_{h<i,j<l} p_{h,l} \cdot \text{Prob}([i,j]|(k,l))
\]
RNAup

Structural Information
RNAup

Structural Information

[Graph showing structural information with probabilities and position in sequence]
RNAup

Structural Information
Example: siRNA Binding

Data taken from Schubert et al 2006
A scanning Version of RNAup

Can we adapt this method for fast searching in large databases?

- *Local* folding algorithms can scan very large sequences by restricting the size of local structures to some maximum $L$.
- RNAplfold computes the probability that regions of length $u$ are unpaired by averaging over all windows of length $L$.
- Runtime is linear in the length of the database $O(n \cdot L^2)$.
A scanning Version of RNAup

Can we adapt this method for fast searching in large databases?

- *Local* folding algorithms can scan very large sequences by restricting the size of local structures to some maximum $L$.
- RNAplfold computes the probability that regions of length $u$ are unpaired by averaging over all windows of length $L$.
- Runtime is linear in the length of the database $\mathcal{O}(n \cdot L^2)$.

Computes average over all windows containing the region

$$
\pi^L[i, j] = \frac{1}{L - (j - i) + 1} \sum_{u = j - L}^{i} P^{u, L}[i, j]
$$
RNAup
RNAplfold -u
sfold

UCUAGAAAGUUUUCAC
AAAGCUAACAGGUAC
CUCGAGAAGUUUUCACAAAGCUAACACCGGAAGUUUUCACAAAGCUAACAAUCGCGGGCCCUAGAGCGGCCGCUUCGAGCAGACAUGAUAAGAUAC
AUUGAUGAGUUUGGACAAACCACAACUAGAAUGCAGUGAAAAAAAUGC
UUUAUUUGUGAAAUUUGUGAUGCUAUUGCUUUAUUUGUAACCAUUAUAAGCUGCAAU
AAACA
Accessibility of miRNA targets

NON WORKING -36.5 kcal/mol

WORKING -28.3 kcal/mol
Accessibility and miRNA targets

miRNA 26

Interaction only

16%

incl. Accessibility

0.6%
Accessibility predicts siRNA efficiency

Data provided by Dharmacon