

RNA Secondary Structures

Ivo Hofacker

Institute for Theoretical Chemistry, University of Vienna

<http://www.tbi.univie.ac.at/~ivo/>

AlgoSB 2019

Marseille, Janvier 2019

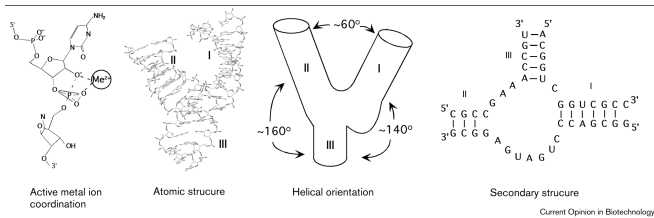
tbi

universität
wien

What is RNA Bioinformatics?

Algorithms on sequences don't care whether it's RNA or protein, so what's special?

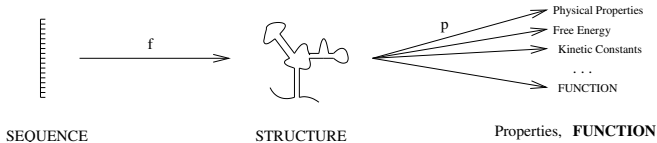
- In contrast to plain sequence analysis we're interested in *structure*
- When talking about *structural bioinformatics* most people think protein structure – RNA is different.
- Structure can be described at many levels
Most useful for RNA work is usually the *secondary structure*



Why look at Structure?

- Basic paradigm of structural biology:
Sequence \rightarrow Structure \rightarrow Function
- Understanding structure is a first step toward function
- Function is what we're ultimately interested in
Sequences is what we have in plenty
- Structure should be conserved more strongly than sequence

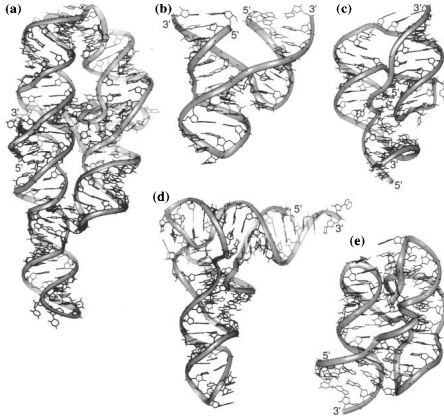
Structure based methods can succeed where sequence analysis fails



Structural Bioinformatics of RNA vs. Proteins

- Amino acid sequences are often better for homology search
Larger alphabet and redundancy of the genetic code makes protein alignment easier than nucleic acid alignment.
- Structural work on proteins often focuses on tertiary structures
- Many more protein than RNA tertiary structures are known
- Protein secondary structures are boring compared to RNA
they only tell which parts of a molecule are in helices/beta-sheets
RNA secondary structure contains interactions (base pairs)
- RNA secondary structures are computationally ease to handle
- Many biological functions can be understood purely on the 2D level

Many non-coding RNAs are Structural RNAs



- (a) Group I intron
- (b) Hammerhead ribozyme
- (c) HDV ribozyme
- (d) Yeast tRNA^{Phe}
- (e) L1 domain of 23S rRNA

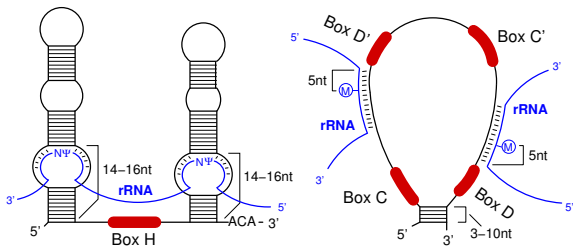
Hermann & Patel, JMB 294, 1999

All the “classical” ncRNAs depend on well-defined and evolutionary conserved **structure**

Example 1: snoRNAs

Small **nucleolar** RNAs are *trans*-acting ncRNAs

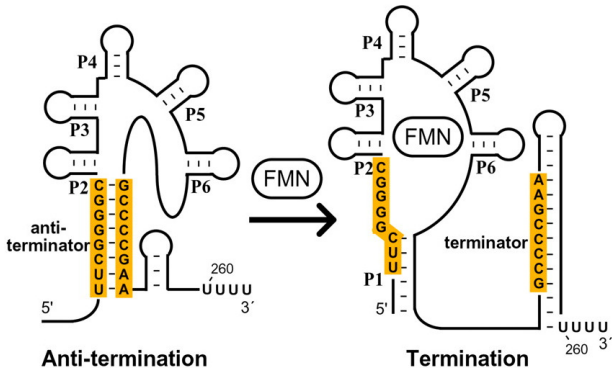
- SnoRNAs guide chemical modifications of rRNAs, snRNAs, and some mRNAs.
- Characteristic secondary structure and sequence motifs
- Two tasks: How to recognize a snoRNA from its sequence?
How to predict the targets of a given snoRNA?



Example 2: The Flavin Mononucleotide Riboswitch

Example for a *cis-acting* RNA element

The FMN-binding riboswitch in the free and bound state



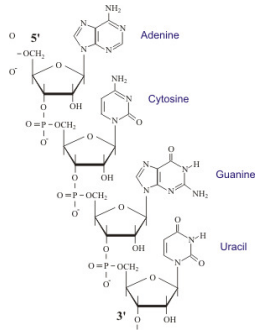
- How can we find novel riboswitches?
- Can we predict how a novel riboswitch works?
Up-, or down-regulation; transcriptional or translational control?

Basic Tasks

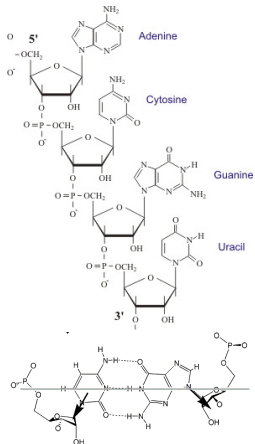
Mostly the same questions are the same as in sequence analysis
Here we want to base them on structure

- **Structure prediction** (sequence \rightarrow structure)
(single or multiple sequences, w/o pseudo-knots, tertiary structure)
- **Classification:**
 - Recognizing new members of a known RNA class (tRNAs, snoRNAs, ...)
 - Homology searches — Given a novel RNA find all the homologs where in Evolution did it first occur?
- ***De novo* prediction**
Annotate all functional RNAs in genomes or transcriptomes
- Many specialized problems for particular RNA classes
e.g. micro RNA target prediction

The RNA Molecule

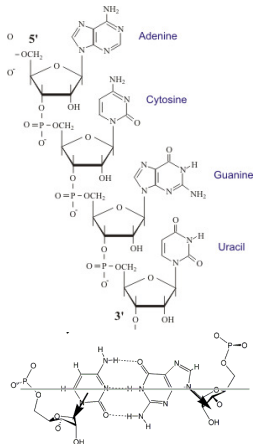


The RNA Molecule



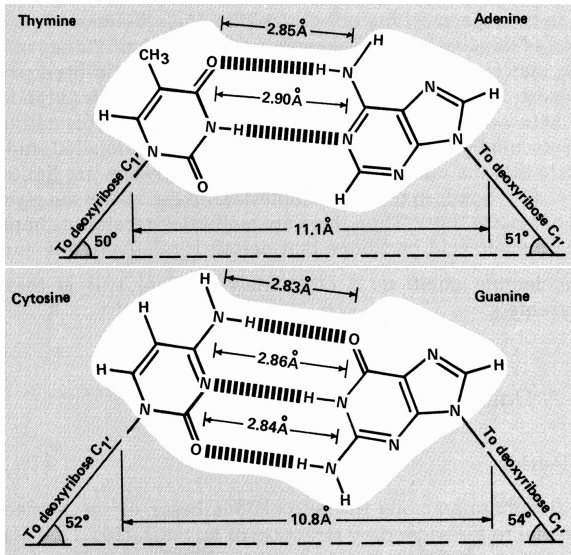
Double helices formed by Watson-Crick AU, UA, GC, CG, GU, UG pairs.

The RNA Molecule

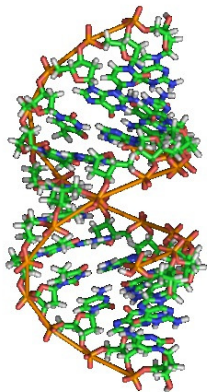


Double helices formed by Watson-Crick AU, UA, GC, CG, GU, UG pairs.

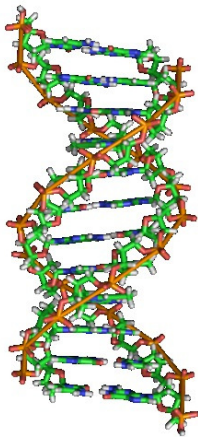
Isostericity of Watson-Crick Pairs



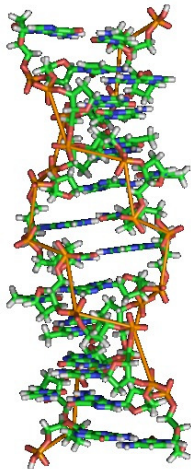
Double Helices



A-form



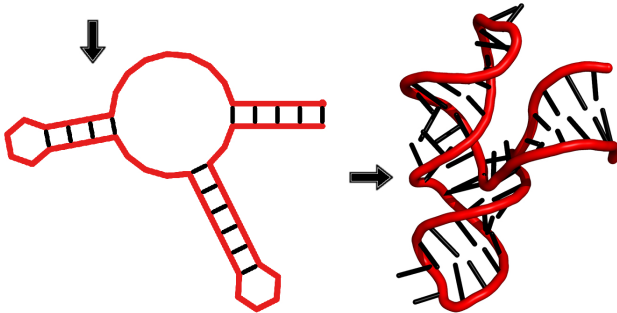
B-form



Z-form

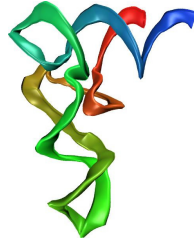
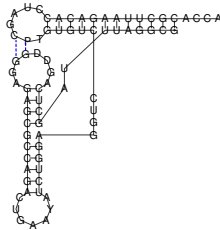
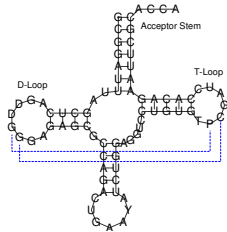
The RNA Folding Problem

GCGCUCUGAUGAGGCCGCAAGGCCGAAACUGCCGCAAGGCAGUCAGCGC



- Hierarchical folding: Secondary structure forms first then helices arrange to form tertiary structure
- Secondary structures cover most most of the folding energy
- Convenient and biologically useful description
- Computationally easy to handle
- Tertiary structure prediction needs knowledge of secondary structure

RNA Secondary Structures



A *secondary structure* is a list of base pairs (i, j) on a sequence x , with

- Any nucleotide (sequence position) can form at most one pair
- No pseudo-knots: No pairs (i, j) and (k, l) with $i < k < j < l$
- If (i, j) is a pair then $x_i x_j \in \{GC, CG, AU, UA, GU, UG\}$
- If (i, j) is a base pair, then $j - i > 3$

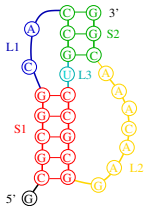
Pseudo knots

Excluding pseudo knots makes life easy, because

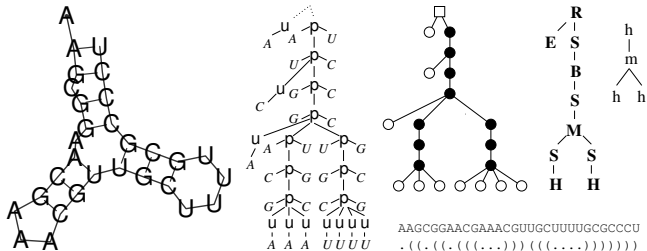
- It greatly simplifies the mathematical model
⇒ Simpler algorithms without pseudo knots
- Many pk-structures are sterically not feasible
- Energetics unknown, except for a few data on H-type pseudo-knots

On the other hand

- Pseudo knots can have important function
- First step toward tertiary structure
- H-type knots are tractable with extra computational effort

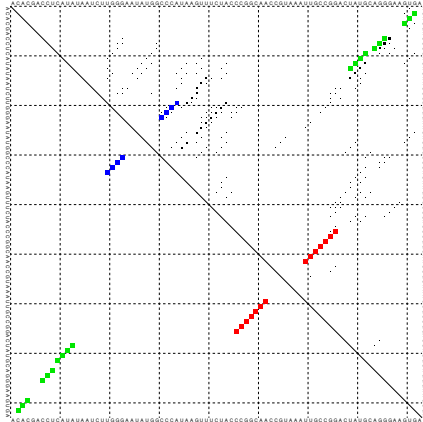


RNA Secondary Structure as Trees



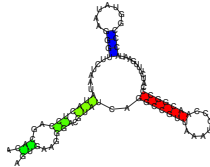
Many RNA algorithms are generalizations of well known *string* algorithms to *trees*

Representing Ensembles of Structures (thermodynamic equilibrium)



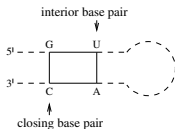
Ensembles of structures (thermodynamic equilibrium) are best represented by base pair probabilities.

A pair (i, j) with probability p is represented by a square in row i and column j with area p .

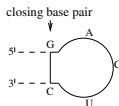


Loop Decomposition

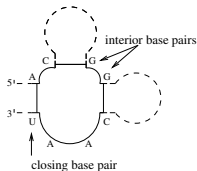
Secondary structures can be uniquely decomposed into loops.
Loops are the *faces* of the secondary structure graph.



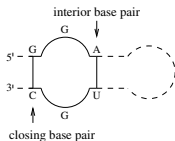
stacking pair



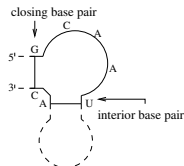
hairpin loop



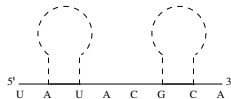
multi loop



interior loop



bulge

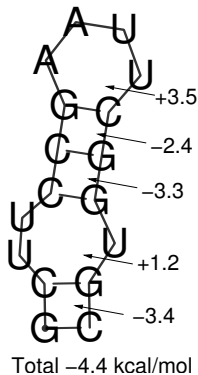


exterior loop

Nearest Neighbor Energies

Free energy of a structure approximated as the sum of loop energies

- Good approximation for most oligo-nucleotides
- Loop energies depend on loop type and size, with some sequence dependence
- Most relevant parameters measured experimentally, some still guesswork
- Free energies are dependent on temperature and ionic conditions
- Training parameters is becoming an alternative to experiment



Stacked Pairs

- Major source of stabilizing energy
- all 21 combinations measured, accuracy at least 0.1 kcal/mol
- include the hydrogen bonding energy of pair formation
- energies of tandem G-U pairs depend on context, i.e. violate the nearest-neighbor model



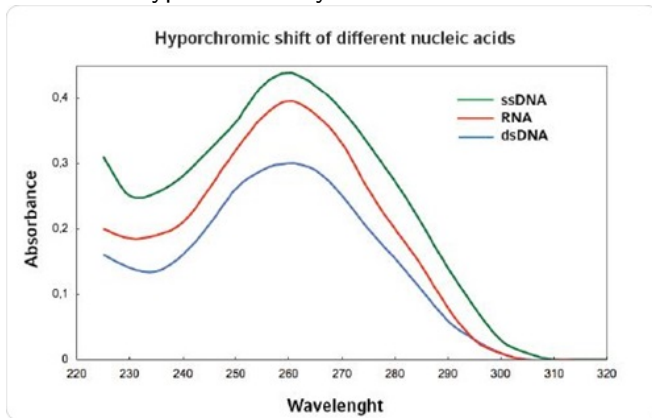
-2.4
kcal/mol

		$(i + 1, j - 1)$					
		CG	GC	GU	UG	AU	UA
(i, j)	CG	-2.4	-3.3	-2.1	-1.4	-2.1	-2.1
	GC	-3.3	-3.4	-2.5	-1.5	-2.2	-2.4
	GU	-2.1	-2.5	1.3	-0.5	-1.4	-1.3
	UG	-1.4	-1.5	-0.5	0.3	-0.6	-1.0
	AU	-2.1	-2.2	-1.4	-0.6	-1.1	-0.9
	UA	-2.1	-2.4	-1.3	-1.0	-0.9	-1.3

For comparison: Thermal energy $RT \approx 0.6\text{kcal/mol}$ at 37C.

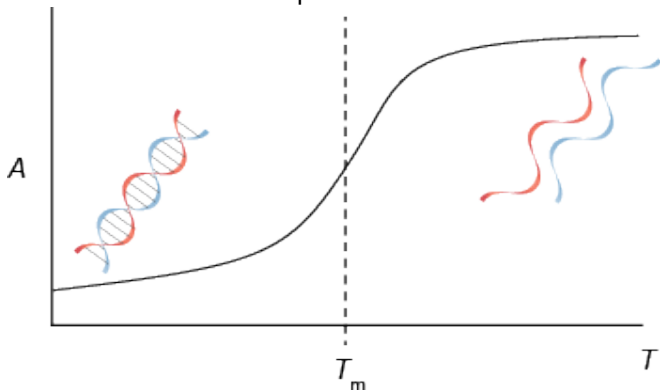
Determining Nearest Neighbor Parameters

UV absorption can distinguish between double and single stranded nucleic Acids — Hyperchromicity



Determining Nearest Neighbor Parameters

UV absorption (at 260nm) can be used to follow the unfolding transition as a function of temperature.



Analyzing UV Melting Curves

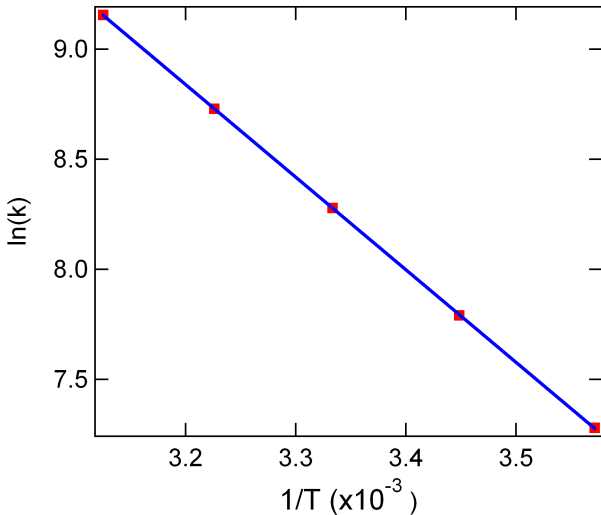
Assume a 2-state model

$$f_{unfolded}(T) = \frac{A(T) - A(T_{min})}{A(T_{max}) - A(T_{min})}$$

$$K_{eq} = \frac{f_{unfolded}}{f_{folded}} = e^{-\Delta G/RT}$$

Van't Hoff Analysis

$$\ln K_{eq} = -\frac{\Delta G}{RT} = -\frac{\Delta H}{RT} + \frac{\Delta S}{R}$$



Algorithms for Secondary Structure Prediction

RNA structures can be predicted by *Dynamic programming* algorithms in many variants.

- Minimum free energy structure (Zuker & Stiegler '81)
- Optimal and *certain* suboptimal structures (Zuker '89)
- All structures within an energy range (Wuchty et al. '99)
- Partition function and base pair probabilities (McCaskill '90)
- Stochastic suboptimals (Ding & Lawrence '01)
- Maximum expected accuracy structures (Do et al '06)
- Consensus structure prediction from alignment
(Knudsen & Hein '99, Hofacker et al. '02)
- Minimum free energy with pseudo-knots (Rivas & Eddy '99)
- *Extended* secondary structures with non-canonical pairs
(Parisien & Major '08, Höner et al '11)

The RNA Conformation Space

The number of secondary structures for a sequence $x = x_1 \dots x_n$ can be enumerated recursively:



$$S_{ij} = S_{i+1,j} + \sum_{k=i+1}^j S_{i+1,k-1} S_{k+1,j} \text{pair}(x_k, x_j)$$

$\text{pair}(x_k, x_j) = 1$ if $x_k x_j$ is a canonical pair

(GC, CG, AU, UA, GU, UG)

otherwise $\text{pair}(x_k, x_j) = 0$.

For typical sequences the number of conformations grows as


$$\bar{S}_{1n} \sim n^{-\frac{3}{2}} 1.85^n$$

Solving the Folding Problem

Toy model for RNA folding: assign energies to base pairs $\varepsilon(x, y)$.

Easily solved by **Dynamic Programming**, i.e.:

Recursive computation with tabulation intermediate results.


$$E_{ij} = \min_{i < k \leq j} \left\{ E_{i+1, j}; \left(E_{i+1, k-1} + E_{k+1, j} + \varepsilon(x_i, x_k) \right) \right\}$$

- E_{1n} is the best possible energy for our sequence x .
- Backtracing through the E table yields the corresponding structure.
- The Algorithm requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ CPU time.

In practice this toy model is not good enough!

For serious predictions, we need to use loop dependent energies.

Folding using Nussinov's Algorithm

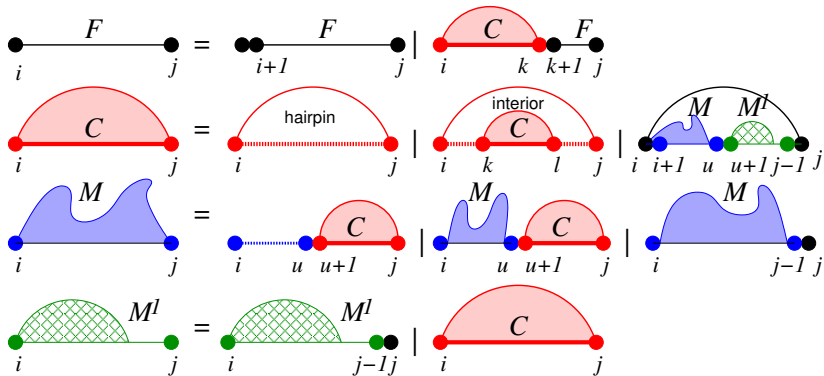
$$\varepsilon(C, G) = \varepsilon(G, C) = -3 \quad \varepsilon(A, U) = \varepsilon(U, A) = -2$$

$$\varepsilon(G, U) = \varepsilon(U, G) = -1$$

$$E_{ij} = E_{i+1,j} \mid E_{i+1,k-1} + E_{k+1,j} + \varepsilon(x_i, x_k)$$

.	())
A	G	C	A	C	A	C	A	G	G	C	
0	0	0	0	0	0	-3	-3	-3	-6	-9	A
	0	0	0	0	0	-3	-3	-3	-6	-9	G
		0	0	0	0	0	0	-3	-6	-6	C
			0	0	0	0	0	-3	-3	-3	A
				0	0	0	0	-3	-3	-3	C
					0	0	0	0	0	0	A
						0	0	0	0	0	C
							0	0	0	0	A
								0	0	0	G
									0	0	G
										0	C

Folding with Loop based Energies



F_{ij} free energy of the optimal substructure on the subsequence $x[i..j]$.

C_{ij} optimal free energy on $x[i..j]$, where (i, j) pair.

M_{ij} $x[i..j]$ is part of a multiloop and contains at least one pair.

M_{ij}^1 same as M_{ij} but contains exactly one component closed by (i, h) .

Minimum Free Energy Folding and Accuracy

MFE folding predicts, the optimal, most probable, structure.

- + easy to program, calculate, and interpret
- one structure to represent the equilibrium ensemble
- no indication of reliability

Accuracy measured by comparison with (phylogenetic) model structures.

- about 40-70% accuracy with current parameters
- accuracy decreases with sequence length
- accurate structures can usually be found within small energy increment of the MFE

Both inaccurate parameters and approximations of the energy model to blame for poor predictions.

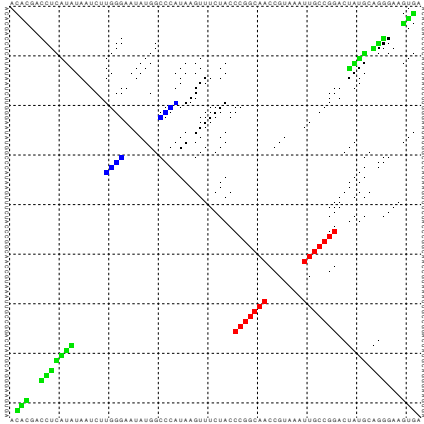
Whenever possible don't rely on a single structure!

Limits to prediction accuracy

A number of different factors contribute to the inaccuracy of our predictions:

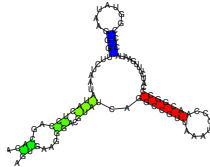
- Beyond secondary structure:
Pseudo-knots, non-canonical pairs, 3D structure
- Deviation from “standard” conditions
Ion concentrations (esp. Mg^{2+}), temperature
- Interaction with other molecules:
RNA interact with proteins other RNAs and metabolites
- Folding kinetics:
native structure \neq ground state structure

Representing Ensembles of Structures (thermodynamic equilibrium)



Ensembles of structures (thermodynamic equilibrium) are best represented by base pair probabilities.

A pair (i, j) with probability p is represented by a square in row i and column j with area p .



Partition Function, Boltzmann probabilities and Pair Probabilities

The partition function Z is the fundamental quantity of statistical mechanics. All thermodynamic properties can be derived from it.

Partition Function, Boltzmann probabilities and Pair Probabilities

The partition function Z is the fundamental quantity of statistical mechanics. All thermodynamic properties can be derived from it.

- Partition function $Z = \sum_s \exp\left(\frac{-E(s)}{RT}\right)$
- Boltzmann probability of a structure s : $p(s) = \frac{1}{Z} \exp\left(\frac{-E(s)}{RT}\right)$
- Expected value of any quantity A : $\langle A \rangle = \sum_s A(s)p(s)$
- Free energy $F = -RT \ln Z$
- Entropy $S = -\frac{\partial F}{\partial T}$

Partition Function, Boltzmann probabilities and Pair Probabilities

The partition function Z is the fundamental quantity of statistical mechanics. All thermodynamic properties can be derived from it.

- Partition function $Z = \sum_s \exp\left(\frac{-E(s)}{RT}\right)$
- Boltzmann probability of a structure s : $p(s) = \frac{1}{Z} \exp\left(\frac{-E(s)}{RT}\right)$
- Expected value of any quantity A : $\langle A \rangle = \sum_s A(s)p(s)$
- Free energy $F = -RT \ln Z$
- Entropy $S = -\frac{\partial F}{\partial T}$

Constrained Partition Functions

Probability of some feature A . Compute the constraint partition function

$$Z^A = \sum_{s \text{ has feature } A} \exp\left(\frac{-E(s)}{RT}\right), \quad p(A) = \frac{Z^A}{Z}$$

- Probability of forming the pair (i, j)
- probability of forming an aptamer structure
- probability of presenting a binding site
- ...

Computing Partition Function and Pair Probabilities

For simplicity back to the Nussinov model:

Computing the partition function $Z = \sum_{\Psi} \exp(-E(\Psi)/RT)$ is simple:

$$Z_{ij} = Z_{i+1,j} + \sum_{k, (i,k) \text{ pairs}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\varepsilon_{ik}/RT).$$

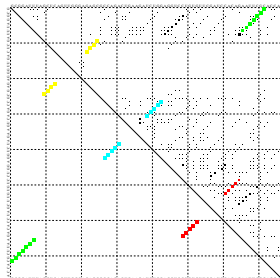
In conjunction with the partition function \hat{Z}_{ij} for structures *outside* the subsequence $x[i..j]$ we can compute pair probabilities:

$$p_{ij} = \hat{Z}_{ij} Z_{i+1,j-1} \exp(-\varepsilon_{ij}/RT) / Z.$$

Pair Probabilities

Equilibrium probabilities for all possible pairs can be calculated via McCaskill's partition function algorithm:

- provides a rigorous description of structures in thermodynamic equilibrium
- probability dot plots contain entropic terms not included in energy dot plots
- starting point for measures of reliability and well-definedness
- not as easy to interpret as a small set of structures



Well-defined Regions

Pair probabilities can help judge the *reliability* of a prediction.

Well-definedness: Are there many structural alternatives?

e.g. “ensemble diversity” (returned by RNAfold) measures dissimilarity of structural alternatives

Computed directly from pair probabilities $\langle d \rangle = \sum_{i,j} p_{ij} \cdot (1 - p_{ij})$

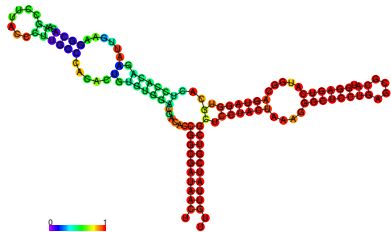
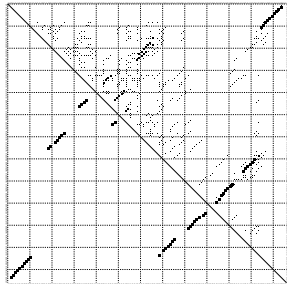
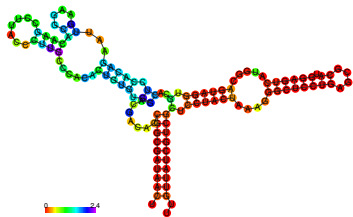
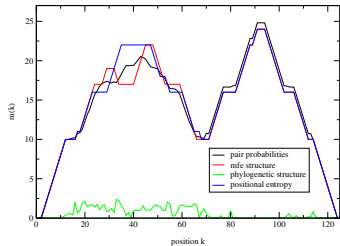
Local reliability: Which parts the prediction we can trust?

High probability pairs are almost always correct.

Secondary structure plots can be colored by pair probability or “positional entropy” to highlight reliable and unreliable regions.

positional entropy is computed from pair probabilities as $S(k) = - \sum_i p_{ik} \ln p_{ik}$

Structures with reliability annotation



Alternatives to the MFE as “best” Structure

Maximum Expected Accuracy (MEA)

- Select the structure expected to have the most base pairs correct
- \rightarrow Maximize the sum of pair probabilities $\langle A \rangle = \sum_{(i,j) \in S} p_{ij}$
- Computed by “Nussinov”-like dynamic programming

Centroid Structure

- Choose the structure that is “nearest” to all other structure in the Boltzman ensemble
- Minimize $\langle d(S) \rangle = \sum_{(i,j) \in S} p_{ij} + \sum_{(i,j) \notin S} (1 - p_{ij})$
- Trivial solution: just pick all pairs with probability > 0.5

No free lunch: Centroid and MEA structure may correspond to a very unlikely structure!

Suboptimal Structures

Zuker's p -suboptimal structures (mfold)

For each pair (i, j) generate the best structure containing that pair.

- + easy to compute by multiple backtracking
- + present a user user with few, representative alternatives
- important alternatives can be missed
- somewhat arbitrary selection of representatives

Complete suboptimal folding

generates **all** structures within given energy range from the mfe.

- + Calculate any thermodynamic average
- + Investigate the energy landscapes, e.g. find low-lying local minima
- huge amount of data, only for moderately short sequences

Stochastic suboptimal structure

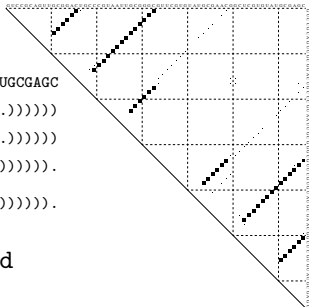
generate Boltzmann weighted sample through stochastic backtracking

Dot Plots and Suboptimal Structures

Zuker suboptimal structures may miss important alternatives

```
GUCCGCAGUUGC GGACUGCCCGUAAUUGCGGGCGUUCGUGUAUGCGAACGGCUCGUGUAUGCGAGC
((((((...))))).((((((...)))))(((((((...))))).((((((...))))))
((((((((((((((...))))))))))))))((((((...))))).((((((...))))))
((((((((((((((((((...)))))))))))))))).((((((((((((((((((...)))))))))))))))).
((((((...))))).((((((...))))).((((((((((((((((((...)))))))))))))).
```

Only the first 3 structures are found by `mfold`



Variations on Backtracing

What we need:

- Let π be a partial structure consisting of
 - Ω_π a set of already known base pairs
 - Υ_π a list of sequence intervals with unknown structure
- $E(\pi)$ best energy that can be obtained by completing π
- \mathfrak{S} a stack of partial structures to be processed

Variations on Backtracing

MFE backtracing:

$\emptyset \rightarrow \mathcal{G}$.

while $\mathcal{G} \neq \emptyset$ **do**

$\pi \leftarrow \mathcal{G}$;

if π is complete **then** output π

$[i, j] = I \in \Upsilon_\pi$.

$\pi' = \pi \blacktriangleleft(i)$

if $E(\pi') = E_{\text{opt}}$

then $\pi' \rightarrow \mathcal{G}$; **next**;

for all $k \in [i, j]$ **do**

$\pi' = \pi \blacktriangleleft(i, k)$

if $E(\pi') = E_{\text{opt}}$

then $\pi' \rightarrow \mathcal{G}$; **next**;

Variations on Backtracing

Wuchty suboptimals

$\emptyset \rightarrow \mathcal{G}$.

while $\mathcal{G} \neq \emptyset$ **do**

$\pi \leftarrow \mathcal{G}$;

if π is complete **then** output π

$[i, j] = I \in \Upsilon_\pi$

$\pi' = \pi \blacktriangleleft(i)$

if $E(\pi') \leq E_{\text{opt}} + \Delta E$

then $\pi' \rightarrow \mathcal{G}$;

for all $k \in [i, j]$ **do**

$\pi' = \pi \blacktriangleleft(i, k)$

if $E(\pi') \leq E_{\text{opt}} + \Delta E$

then $\pi' \rightarrow \mathcal{G}$;

Variations on Backtracing

Stochastic backtracing

$\emptyset \rightarrow \mathfrak{S}$.

while $\mathfrak{S} \neq \emptyset$ **do**

$\pi \leftarrow \mathfrak{S}$;

if π is complete **then** output π

$[i, j] = I \in \Upsilon_\pi$

$r = Z(\pi) \cdot \text{rand}()$;

$\pi' = \pi \blacktriangleleft(i)$; $r = r - Z(\pi')$

if $r \leq 0$ **then** $\pi' \rightarrow \mathfrak{S}$; **next**;

for all $k \in [i, j]$ **do**

$\pi' = \pi \blacktriangleleft(i, k)$; $r = r - Z(\pi')$

if $r \leq 0$ **then** $\pi' \rightarrow \mathfrak{S}$; **next**;

SCFG based Methods

Stochastic Context Free Grammars

- An extension to HMMs that can handle long range correlations
- Formally, a CFG is a tuple $G = (V, \alpha, S, R)$ of Alphabet α , nonterminals V , Start symbol S , and rewrite rules R
- *Stochastic* CFG adds probabilities to each rule
- Generates sequences using a series of rewriting rules
- Well known repertoire of algorithms that work on any SCFG

Nussinov Algorithm as SCFG


$$S \rightarrow aS|cS|gS|uS$$
$$S \rightarrow PS$$
$$S \rightarrow \emptyset$$
$$P \rightarrow aSu|uSa|cSg|gSc|gSu|uSg$$

- Applying grammar rules produces an RNA sequence
- The parse tree of productions used corresponds to the structure
- A *stochastic* CFG assigns probabilities to each production rule
- CYK algorithm finds the parse tree (structure) most likely to produce the given sequence

SCFGs vs Energy directed folding

Many RNA related algorithms come in two flavors, based on either *energy directed folding* or *stochastic context free grammars*

physics based model

parameters from experiment

probabilistic inference

parameters learned from data

Algorithms are mostly analogous, but terminology is different!

MFE folding

pair probabilities

CYK algorithm

Inside/Outside algorithm

Beyond classical secondary structure

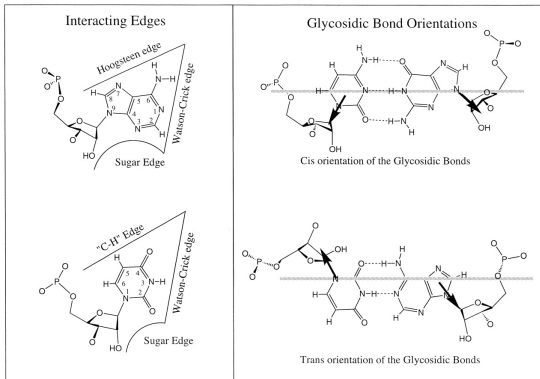
State of the art in RNA secondary structure prediction:

- Classical secondary structure considers only 6 types of pairs
CG,GC,AU,UA,GU,UG
- All base pairs are assumed to be Watson-Crick type
- Loops are drawn as unstructured bubbles
- Energy model assigns free energies to each loop

Can we improve the level of detail?

Leontis-Westhof classification

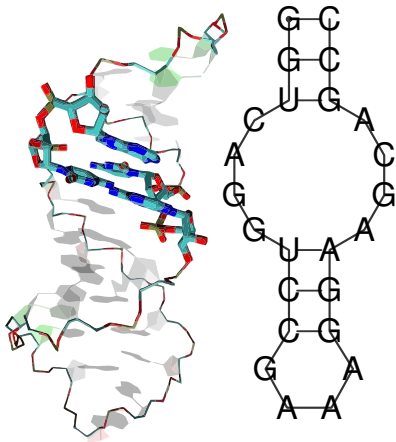
Watson-Crick base pairing is only part of the story!



Overall 12 different pairing types for any two bases

Starting point for defining recurring tertiary structure *motifs*

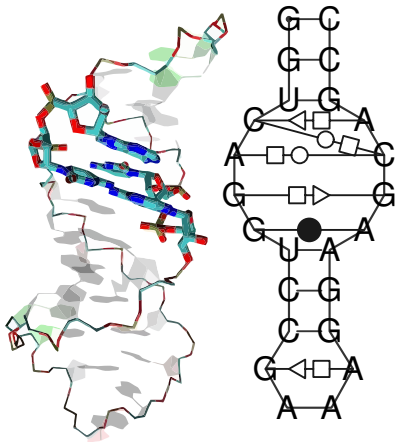
2D and 3D Structures



Looking at known tertiary structures we find that:

- “Loops” are not unstructured
- full of non-canonical base pairs
- all 16 $[ACGU] \times [ACGU]$ combinations form pairs
- the same two nucleotides can pair in many different ways

2D and 3D Structures



Looking at known tertiary structures we find that:

- “Loops” are not unstructured
- full of non-canonical base pairs
- all 16 $[ACGU] \times [ACGU]$ combinations form pairs
- the same two nucleotides can pair in many different ways