The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2584-4

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

RNA-Secondary Structure Prediction

Martin Weigt
Université Pierre et Marie Curie
Paris
France

RNA - Secondary structure prediction

RNA function:

- messenger RNA
- transfer RNA
- catalytique RNA
- regulatory RNA
- Ribosome

} structured !

secondary & tertiary structure

⇓

base and row !

2D Structure: Internal base pairing in an RNA

=> forms local helices

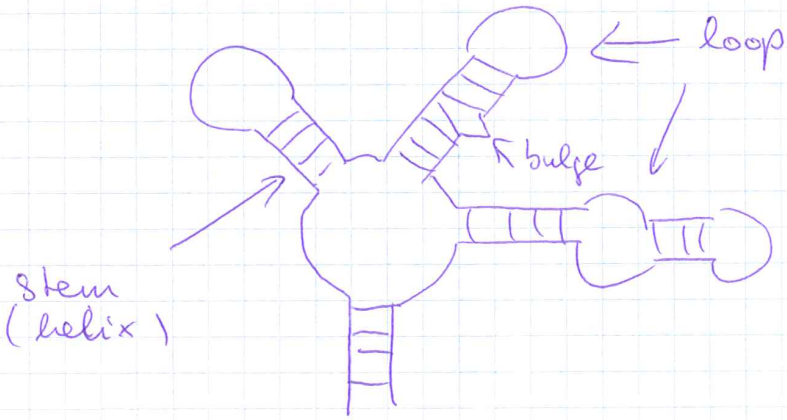
Watson Crick base pairs A - U

(Adenin - Uracil)

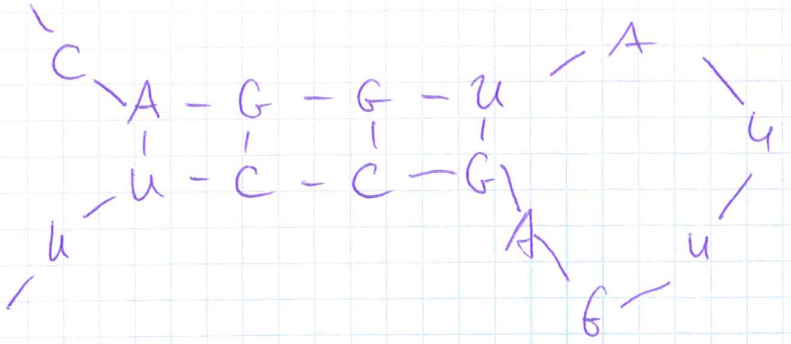
G - C

(Guanine - Cytozine)

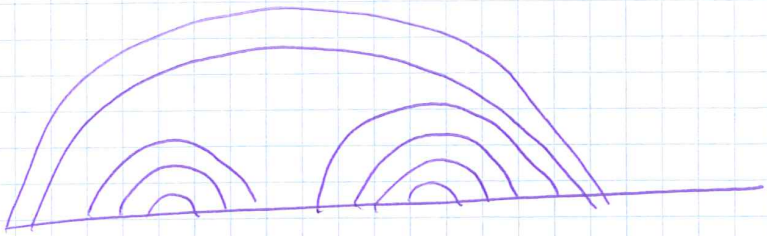
Wobble pairs G - U



Stem: → base pairing
→ complementary sequences



⇒ planar graphical structure



base pairs i, j

Secondary-structure prediction

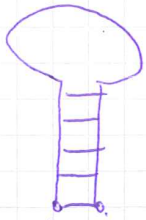
The Nussinov algorithm (1978)

⇒ maximizes number of base pairs

- dynamical programming (recursive)
- too simple to provide realistic 2D str., but algorithmic idea used also in more sophisticated algorithms.

Notation: subsequence (i, j)
→ from i to j : $a_i, a_{i+1}, \dots, a_j$

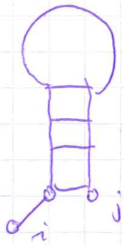
4 possibilities:



i,j paired

$\Downarrow$

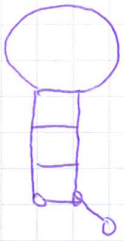
a_i-a_j pairs + optimal struct. for $(i+1, j-1)$



i unpaired

$\Downarrow$

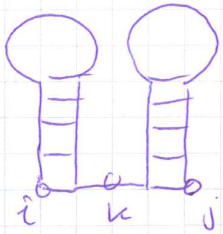
a_i added + opt. struct $(i+1, j)$



j unpaired

$\Downarrow$

a_j added + opt struct. $(i, j-1)$



bifurcation

$\Downarrow$

combine opt. struct. (i,k) and $(k+1, j)$

Notation: Sequence $a_1, \dots, a_L$

$$\Delta(i,j) = \begin{cases} 1 & \text{if } (a_i a_j) \text{ complementary} \\ 0 & \text{else} \end{cases}$$

Score: $\gamma(i,j)$ = maximal number of bp in seq. (i,j)

Nussinov

- Initialization:

$$\gamma(i, i-1) = 0 \quad \text{for} \quad i = 2, \dots, L$$
$$\gamma(i, i) = 0 \quad \text{for} \quad i = 1, \dots, L$$

- Recursion over length $(j-i)$ of sub-sequences

$$\gamma(ij) = \max \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + \Delta(ij) \\ \max\limits_{i<k<j} \left[\gamma(ik) + \gamma(k+1, j) \right] \end{cases}$$

$\Rightarrow \gamma(1, L)$ = max. number of bp in 2D structure

RNA 4

Traceback

- Initialization: $(1, L) \rightarrow$ stack
- Recursion:

While stack ≠ empty

- if $i \geq j$ continue
- else if $\gamma(i+1, j) = \gamma(i, j)$ push $(i+1, j)$
- else if $\gamma(i, j-1) = \gamma(i, j)$ push $(i, j-1)$
- else if $\gamma(i, j) = \gamma(i+1, j-1) + \Delta(i, j)$:
  - record base pair i, j
  - push $(i+1, j-1)$
- else for $k = i+1, \ldots, j-1$
  if $\gamma(i, j) = \gamma(i, k) + \gamma(k+1, j)$
  - push $(k+1, j)$
  - push (i, k)
  - break

---

algorithmic complexity:

time ~ $O(L^3)$

space ~ $O(L^2)$

The Zuker algorithm (Zuker, Stiegler '81)

⇒ energy minimization

Hypothesis: correct 2D structure

⇔ minimal equilibrium free energy (ΔG)

RNA5

Approximation :

ΔG = contribution from loops, stems

base pairing + stacking

+ dynamic programming
→ optimal 2D structure.

Mc Caskill '90

- probabilistic version
  $\Delta G \rightarrow$ Gibbs-Boltzmann weight
  $\sim \exp\{-\Delta G/kT\}$
- dynamic programming to calculate marginal probability that i & j are paired.
  (⇒ sum over all 2D structures where (i,j) is base paired)

Wuchty et al '99

⇒ all sub-optimal structures in given free-energy interval

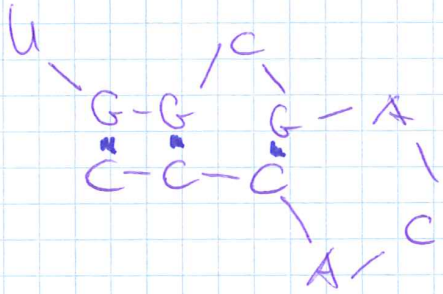
Implementation : Vienna RNA package

RNA evolution & structural constraints

exist: homologous RNA with
- same secondary structure
- very weak sequence similarity

ex:

human	U	G	G	C	G	A	C	A	C	C	C
mouse	A	C	A	C	C	A	A	A	G	U	G
worm	G	G	G	C	A	C	C	A	U	U	C
fly	U	U	G	C	U	A	C	C	A	U	A
orca	G	G	G	C	G	U	A	A	C	U	C



⇒ dramatic changes in sequence, but base pairing conserved!

⇒ covariation / coevolution

⇓

- single mutation breaks bp
- compensatory mutations re-establishes bp.

How to measure covariation?

$$f_i(a) = \frac{n_i(a)}{M}$$ ... empirical frequency single position

$$f_{ij}(a,b) = \frac{n_{ij}(a,b)}{M}$$ ... fraction of seqs. with nt a in pos. i, and nt. b in pos. j.

Mutual information

RN47

$$M_{ij} = \sum_{a,b \in \{A,G,C,U\}} f_{ij}(a,b) \log \frac{f_{ij}(a,b)}{f_i(a)\, f_j(b)}$$

$$= \begin{cases} 0 & \text{if pos. } i \text{ \& } j \text{ independent} \\ > 0 & \text{if correlated.} \end{cases}$$

[information about the identity on nt in position j provided by knowledge of nt in position i]

Observation: M_{ij} big for bp in secondary structure

⇒ possibility to predict 2D structure starting from MSA

Step 1: For each $1 \le i < j \le L$, calculate M_{ij}

Step 2: Apply modified Nussinov algo.

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + M_{ij} \\ \max_{i<k<j} \left[\gamma(i,k) + \gamma(k+1, j) \right] \end{cases}$$

⇒ This algorithm maximizes the likelihood of the data (MSA) under a covariance model

$$P(a_1,\dots,a_L \mid 2DS) = \underbrace{\prod_{i=1}^{L} f_i(a_i) \cdot \prod_{(i,j)\in 2DS} \frac{f_{ij}(a_i,a_j)}{f_i(a_i)\,f_j(a_j)}}$$

⇒ more complicated than PWM

⇒ long range correlations on 2D structure.

A big problem

- The secondary-structure prediction quality depends on the quality of the alignment
- RNA alignment is difficult due to low sequence similarities
  → substantially improved if based on secondary structure

⇒ iterative process

INFERNAL (S. Eddy)

= Inference of RNA alignments

(but technical details are involved...)