Free Energy Minimization

Idea:

- Overcome the main drawback of Nussinov's algorithm: non-realism of base pair maximization!
- Define an energy model for RNA that can be parameterized by experimentally measured energies
- Devise an algorithm that minimizes the free energy of RNA according to this model
- Algorithm (by Zuker) will be similar to Nussinov's algorithm

Gibbs Free Energy

Definition (Gibbs Free Energy)

The *Gibbs Free Energy G* of a system (e.g. dilution of RNAs) is

$$G = H - TS$$

where H is the enthalpy (potential to perform work), T the absolute temperature and S the entropy (measure of disorder).

Remarks:

- For RNA, we will compute the free energy of (a certain amount $N_A \approx 6 \cdot 10^{23}$ of molecules, a "mol") of a certain structure P. More precisely, we compute the *change of free energy* ΔE due to folding into P from $P_{\text{unfolded}} = \{\}$.
- The (change of) Gibbs free energy corresponding to P can be computed by summing free energy contributions from single "structural elements".
- Those contributions (for loops, stacks, ...) can be measured experimentally (Turner). They consist of enthalpic and entropic terms. Due to the latter, they depend on temperature.

Gibbs Free Energy

Definition (Gibbs Free Energy)

The *Gibbs Free Energy G* of a system (e.g. dilution of RNAs) is

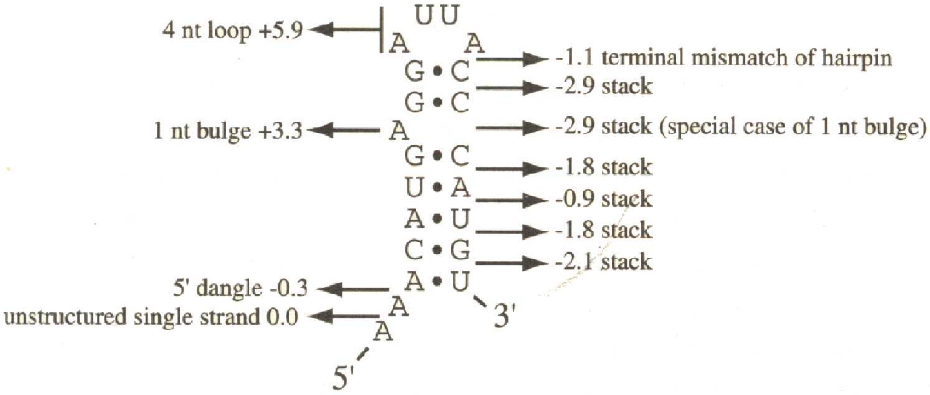
$$G = H - TS$$

where H is the enthalpy (potential to perform work), T the absolute temperature and S the entropy (measure of disorder).

Remarks:

- For RNA, we will compute the free energy of (a certain amount $N_A \approx 6 \cdot 10^{23}$ of molecules, a "mol") of a certain structure P. More precisely, we compute the *change of free energy* ΔE due to folding into P from $P_{\text{unfolded}} = \{\}$.
- The (change of) Gibbs free energy corresponding to P can be computed by summing free energy contributions from single "structural elements".
- Those contributions (for loops, stacks, ...) can be measured experimentally (Turner). They consist of enthalpic and entropic terms. Due to the latter, they depend on temperature.

Free Energy — Example



overall $\Delta G = -4.6$ kcal/mol

Free Energy Model of RNA — Definitions

Definition (Secondary structure elements/Loops)

Let S RNA sequence of length n, P RNA structure of S. Call $1 \leq i \leq n$ *unpaired in* P, iff there is no j, s.t. $(i,j) \in P$ or $(j,i) \in P$.

- $(i,j) \in P$ *closes a* **hairpin loop** iff all $k : i < k < j$ unpaired in P
- $(i,j) \in P$ *closes a* **stacking loop** iff $(i+1, j-1) \in P$
- $(i,j) \in P$ *and* $(i',j') \in P$ *form an* **internal loop** (i,j,i',j') iff
    - $i < i' < j' < j$
    - (i,j) does not close a stacking loop
    - all $i+1, \ldots, i'-1$ and $j'+1, \ldots, j-1$ unpaired in P

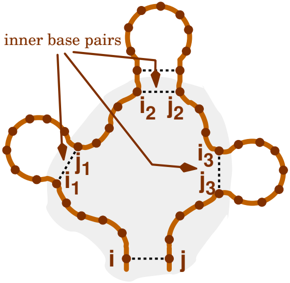
Free Energy Model of RNA — Definitions, ctd.

- An internal loop (i, j, i', j') is called **left** *(***right***)* **bulge**, iff $j = j' + 1$ ($i' = i + 1$), respectively.
- A ***k*-multiloop** consists of k base pairs $(i_1, j_1) \dots (i_k, j_k) \in P$ and a closing base pair $(i, j) \in P$ with the property that
    - $i < i_1 < j_1 < i_2 < j_2 < \dots < i_k < j_k < j$
    - $i + 1 \dots i_1 - 1; j_1 + 1 \dots i_2 - 1; \quad \dots \quad ; j_{k-1} + 1 \dots i_k - 1; j_k + 1 \dots j - 1$ unpaired in P

  $(i_1, j_1) \dots (i_k, j_k)$ close the **inner base pairs** *of the multiloop*.

Remarks

- *k*-**multiloop**



- Usually hairpin loops have minimal loop size of $m = 3$
  $\Rightarrow$ for all $(i, j) \in P$: $i < j - 3$.
- each secondary structure element is defined uniquely by its closing basepair
- for any basepair (i, j) we denote the corresponding secondary structure element with $Sec(i, j)$.

Energy of Secondary Structure Elements

Definition (Energy contribution of loops)

Energy contributions of the various structure elements:

- **hairpin loop** (i,j): $\mathrm{eH}(i,j)$
- **stacking** (i,j): $\mathrm{eS}(i,j)$
- **internal loop** (i,j,i,j'): $\mathrm{eL}(i,j,i',j')$
- **multiloop**: $\mathrm{eM}(i,j,i_1,j_1,\ldots,i_k,j_k)$

Remark

General multi loop contribution will be too expensive in prediction: exponential explosion!
$\Rightarrow$ Use a simplified contribution scheme.

Definition (Simplified energy contribution of multiloops)

- **multiloop** $\mathrm{eM}(i,j,k,k') = a + bk + ck'$

  $a, b, c =$ weights, $a =$ energy contribution for closing of loop
  $k =$ number of inner base pairs
  $k' =$ number of unpaired bases within loop

Loop Energy and Free Energy of an RNA

Definition (Free Energy of an RNA)

Given an RNA structure P of an RNA sequence S.

loop free energy: $E_{ij}^P :=$ energy contribution of $Sec(i,j)$

total free energy: $E(P) := \sum_{(i,j)\in P} E_{ij}^P$

Remark

more precisely we could write $E_S(P)$, since energy of P also depends on S

$\rightarrow$ we assume S is fix

Problem of Free Energy Minimization

Definition (RNA Structure Prediction by Energy Minimization)

- IN: RNA sequence S
- OUT: non-crossing RNA structure P of S, such that

$$E(P) = \min_{P' \text{ non-crossing RNA structure of } S} E(P')$$

Zuker's Algorithm for RNA Energy Minimization

Remarks

- Plan: the Zuker-Algorithm will be specified by defining matrix entries and giving recursion equations. Analogously to Nussinov, those recursions can be evaluated effictiently by DP. The optimal structure is obtained by Traceback.
- Do we need a *completely* new algorithm?

Definition (W-matrix)

For an RNA sequence S, define the Zuker-matrix W as a matrix of entries W_{ij} for $1 \leq i \leq j \leq n$ by

$$W_{ij} := \min\{E(P) \mid P \text{ non-crossing RNA } ij\text{-substructure of } S\}.$$

Remark

$E(P)$ can be used to evaluate a ij-substructure P, since P is still an RNA structure. Tacitely, we assume that sequence outside of base pairs does not contribute to the energy.

Zuker Recursion, Take 1

Initialisation: (for $j - i \leq m$)

$$W_{ij} = 0$$

Recursion: (for $i < j - m$)

$$W_{ij} = \min \begin{cases} W_{ij-1} & \text{— } j \textit{ unpaired} \\ \min_{i \leq k < j-m} W_{ik-1} + W_{k+1j-1} + E(???) & \text{— } j \textit{ paired} \end{cases}$$

Zuker Recursion: W-Recursion and V-matrix

Initialisation: (for $j - i \leq m$)

$$W_{ij} = 0$$

Recursion: (for $i < j - m$)

$$W_{ij} = \min \begin{cases} W_{ij-1} & \text{— } j \text{ unpaired} \\ \min_{i \leq k < j-m} W_{ik-1} + \underset{V_{kj}}{\cancel{W_{k+1j-1} + E(???)}} & \text{— } j \text{ paired} \end{cases}$$

Definition (V-matrix)

For an RNA sequence S, define the Zuker-matrix V as a matrix of entries V_{ij} for $1 \leq i \leq j \leq n$ by

$$V_{ij} := \min \left\{ E(P) \;\middle|\; \begin{array}{l} P \text{ non-crossing RNA } ij\text{-substructure of } S, \\ \text{where } (i,j) \in P \end{array} \right\}.$$

"minimal energy of any closed ij-substructure of S"

V-Recursion, Take 1

Initialization: (for $j - i \leq m$)

$$V_{ij} = \infty$$

Recursion: (for $i < j - m$)

$$V_{ij} =$$

$$\min \begin{cases} \mathrm{eH}(i,j) & \text{— } \textit{hairpin loop} \\ V_{i+1,j-1} + \mathrm{eS}(i,j) & \text{— } \textit{stacking loop} \\ \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eL}(i,j,i',j') & \text{— } \textit{interior loop/bulge} \\ \min_{k,i<i_1<j_1<\dots<i_k<j_k<j} \mathrm{eM}(i,j,i_1,j_1,\dots,j_k,j_k) & \text{— } \textit{multi-loop} \\ \qquad + \sum_{1\leq k'\leq k} V_{i_{k'}j_{k'}} & \end{cases}$$

Remarks

- V-recursion for *general* multi-loop energy
- complexity: multi-loop case exponential
- now: optimize using simplified multi-loop energy

V-Recursion, Take 1

Initialization: (for $j - i \leq m$)

$$V_{ij} = \infty$$

Recursion: (for $i < j - m$)

$$V_{ij} =$$

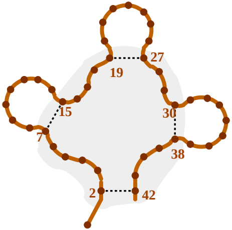
$$\min \begin{cases} \mathrm{eH}(i,j) & \text{— \textit{hairpin loop}} \\ V_{i+1,j-1} + \mathrm{eS}(i,j) & \text{— \textit{stacking loop}} \\ \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eL}(i,j,i',j') & \text{— \textit{interior loop/bulge}} \\ \min_{k,i<i_1<j_1<\cdots<i_k<j_k<j} \mathrm{eM}(i,j,i_1,j_1,\ldots,j_k,j_k) & \text{— \textit{multi-loop}} \\ \qquad + \sum_{1\leq k'\leq k} V_{i_{k'}j_{k'}} & \end{cases}$$

Remarks

- V-recursion for *general* multi-loop energy
- complexity: multi-loop case exponential
- now: optimize using simplified multi-loop energy

Simplified Multi-loop Energy — Example

- In general: multi-loop energy depends on everything: inner base pairs $(i_1, j_1) \dots (i_k, j_k)$, closing base pair (i, j), and sequence.
- Simplification: dependency only on number of inner base pairs k and number of unpaired bases k'.
- Example:



general: eM$(2, 42, 7, 15, 19, 27, 30, 38)$
simplified: eM$(2, 42, k, k') = a + bk + ck'$, where
$k = 3$: inner base pairs within loop
$k' = 12$: unpaired bases within multi-loop

- We will use: New multi-loop energy is additive

Efficient V-Recursion and WM-matrix

Initialization: (for $j - i \leq m$) $V_{ij} = \infty$ "as before"
Recursion: (for $i < j - m$)

$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j) & \text{— \textit{hairpin loop}} \\ V_{i+1,j-1} + \mathrm{eS}(i,j) & \text{— \textit{stacking loop}} \\ \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eL}(i,j,i',j') & \text{— \textit{interior loop/bulge}} \\ \min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + a & \text{— \textit{multi-loop}} \end{cases}$$

Definition (WM-matrix)

For an RNA sequence S, the Zuker-matrix WM has entries WM_{ij} for $1 \leq i \leq j \leq n$:

$$WM_{ij} := \min \left\{ E^m_{ij}(P) \;\middle|\; \begin{array}{l} P \text{ non-crossing RNA } ij\text{-substructure of } S, \\ P \text{ not empty} \end{array} \right\},$$

where E^m_{ij} evaluates P as part of a multi-loop (i.e. including energy contributions b,c due to inner base pairs, unpaired bases).

Efficient V-Recursion and WM-matrix

Initialization: (for $j - i \leq m$) $V_{ij} = \infty$ "as before"
Recursion: (for $i < j - m$)

$$V_{ij} = \min \begin{cases} eH(i,j) & \text{— hairpin loop} \\ V_{i+1,j-1} + eS(i,j) & \text{— stacking loop} \\ \min_{i<i'<j'<j} V_{i',j'} + eL(i,j,i',j') & \text{— interior loop/bulge} \\ \min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + a & \text{— multi-loop} \end{cases}$$

Definition (WM-matrix)

For an RNA sequence S, the Zuker-matrix WM has entries WM_{ij} for $1 \leq i \leq j \leq n$:

$$WM_{ij} := \min \left\{ E^m_{ij}(P) \;\middle|\; \begin{array}{l} P \text{ non-crossing RNA } ij\text{-substructure of } S, \\ P \text{ not empty} \end{array} \right\},$$

where E^m_{ij} evaluates P as part of a multi-loop (i.e. including energy contributions b,c due to inner base pairs, unpaired bases).

Remarks to Definition of WM-matrix

we defined:

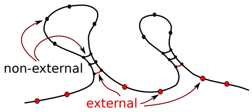
"$WM_{ij} := \min\{E^m_{ij}(P) \mid P$ RNA ij-substructure of S, P not empty$\}$, where E^m_{ij} evaluates P as part of a multi-loop"

Remarks

- "P not empty" ensures that the multi-loop case in the V-recursion cannot recurse to non-multiloops
- "$E^m_{ij}(P)$ evaluates P as part of a multi-loop" means that E^m_{ij} adds to $E(P)$ contributions c for unpaired bases (here we need i and j) and contributions b for inner base pairs of this part of a complete multi-loop. Define

$$E^m_{ij}(P) := E(P) + kb + k'c,$$

where k is the number of *external* base pairs and k' the number of *external* unpaired bases in P.



WM-Recursion

Initialization: (for $j - i \leq m$)

$$WM_{ij} = \infty \quad (ij\text{-substructure } P \text{ non-empty!})$$

Recursion: (for $i < j - m$)

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c & \text{— } \textit{j unpaired} \\ WM_{i+1j} + c & \text{— } \textit{i unpaired} \\ V_{ij} + b & \text{— } \textit{closed} \\ \min_{i<k<j} WM_{ik} + WM_{k+1j} & \text{— } \textit{non-closed} \end{cases}$$

Remark

decomposition complete — cases not distinct (which is ok for minimization!)

Zuker-Algorithm: Summary

- 3 matrices:
  W — minimal energy of general substructure $i \dots j$
  V — minimal energy of closed substructure $i \dots j$
  WM — minimal energy of true part of a multi-loop $i \dots j$
- recursions equations

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \le k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j),\ V_{i+1,j-1} + \mathrm{eS}(i,j) \\ \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eL}(i,j,i',j') \\ \min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + a \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c,\ WM_{i+1j} + c,\ V_{ij} + b \\ \min_{i<k<j} WM_{ik} + WM_{k+1j} \end{cases}$$

immediate complexity: $O(n^4)$ time, $O(n^2)$ space

Complexity Revisited

$O(n^2)$ matrix entries
Multi-loop branching: “only” $O(n)$
Interior loop: $O(n^2)$ **limiting!**
Trick: reduce complexity of limiting case.
simplest: bound maximal interior loop size (e.g. 30)

Theorem. (Zuker)
Given an RNA sequence S, Zuker's algorithm predicts the non-crossing, minimal energy structure P of S in $O(n^3)$ **time and** $O(n^2)$ **space.**

Remarks

- Minimal free energy in W_{1n}
- We assume traceback is done analogously to Nussinov-Traceback. Same reduced complexity. Only extension: trace through three matrices, i.e. keep track of matrix.

Implementations

- Michael Zuker's `Mfold` / `Unafold`
- Ivo Hofacker's Vienna RNA Package: `RNAfold`
- David Mathew's `RNAstructure`
- Example:

```
ivo@tbi: $ RNAfold

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
length = 72
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
(((((((.((((.......)))).(((((((..((((....))))..))).)))))).....))))))).
minimum free energy = -26.70 kcal/mol
```

additionally: produces file rna.ps

Implementations

- Michael Zuker's `Mfold` / `Unafold`
- Ivo Hofacker's Vienna RNA Package: `RNAfold`
- David Mathew's `RNAstructure`
- Example:

```
ivo@tbi: $ RNAfold

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
length = 72
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
(((((((.((((.......)))).(((((((..((((....))))..))).)))))......))))))).
minimum free energy = -26.70 kcal/mol
```

additionally: produces file rna.ps

Implementations

- Michael Zuker's `Mfold` / `Unafold`
- Ivo Hofacker's Vienna RNA Package: `RNAfold`
- David Mathew's `RNAstructure`
- Example:

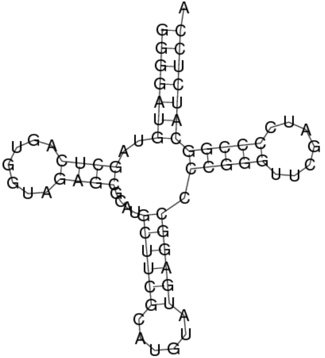
```
ivo@tbi: $ RNAfold

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
length = 72
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
((((((((.((((.......)))).((((((((..((((....))))..))).))))).....)))))))).
minimum free energy = -26.70 kcal/mol
```

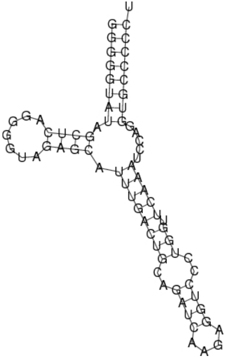
additionally: produces file rna.ps

Example: tRNAs

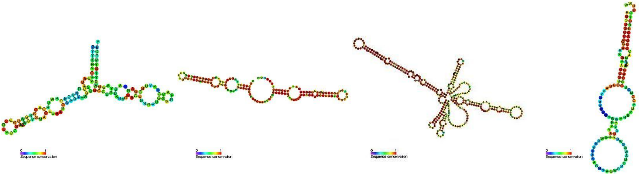
- Mouse tRNA-ALA:



- Mouse tRNA-CYS:



- A biologist finds new RNA (i.e. usually only RNA sequence!)
  - get (first idea of) structure by using RNAfold
  - see whether similarities to known structures exist. Can we guess the RNA family by characteristic shape?



5S rRNA H/ACA snoRNA 7SK Y RNA

recommended: browse Rfam, e.g.
`http://rfam.sanger.ac.uk/family/browse/top20`

- Biologist has several RNAs. Are they similar by structure?
- We have a sequence: could it be structural RNA?