Drawback of Nussinov

Nussinov does not calculate biologically relevant structures because:

- often there are various possibilities for basepairing (especially due to pseudoknots), the Nussinov algorithm detects mostly just one variant.

- stacking of basepairs is not considered $\Rightarrow$ differences in structure and stability of helices.

- Size of internal loops are not considered

Solution: minimizing the free energy

Definition (Free Energy)

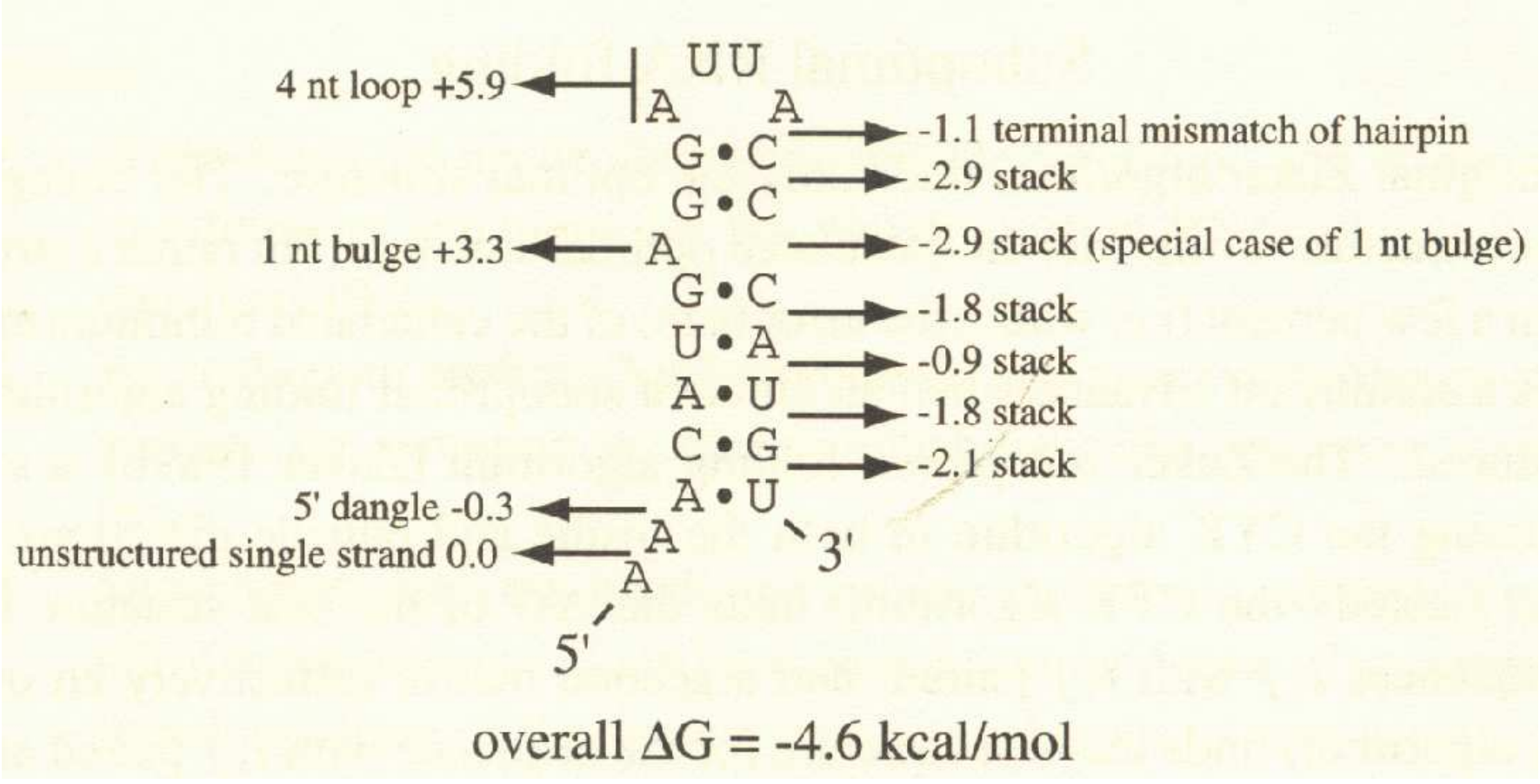
The *Gibbsian Free Energie* G in a system (e.g. of gas molecules in equilibrium or in a dilution of molecules) holds

$$G = H - TS \tag{1}$$

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

- enthalpy: by basepairs
- entropy: "disorder in unpaired regions"
- only possible to measure the difference
$$\Delta G = \Delta H - T\Delta S$$
- can be measured $\Rightarrow$ flexible rules for loops,stacks and further secondary structure elements.
- complete free energy: summation

Freier Rules



The Zuker Algorithm/Definitions

Definition (Secondary structure elements)

Let S be a fixed sequence. Further, let P be an RNA structure for S.

- a basepair $(i,j) \in P$ closes a **hairpin loop** if $\forall i < i' \leq j' < j : (i',j') \notin P$.
- a basepair $(i,j) \in P$ closes a **stacking** if $(i+1, j-1) \in P$.
- two basepairs $(i,j) \in P$ and $(i',j') \in P$ form an **internal loop** (i,j,i',j') if
  - $i < i' < j' < j$
  - $(i'-i)+(j-j') > 2$ (no stack)
  - there is no basepair (k,l) between (i,j) and (i',j').

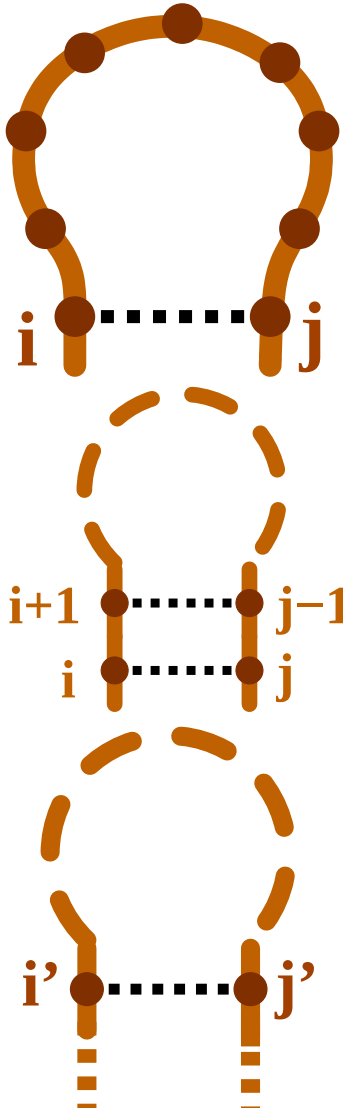
The Zuker Algorithm/Definitions

Definition (Secondary structure elements)

Let S be a fixed sequence. Further, let P be an RNA structure for S.

- a basepair $(i,j) \in P$ closes a **hairpin loop** if $\forall i < i' \leq j' < j : (i',j') \notin P$.
- a basepair $(i,j) \in P$ closes a **stacking** if $(i+1, j-1) \in P$.
- two basepairs $(i,j) \in P$ and $(i',j') \in P$ form an **internal loop** (i,j,i',j') if
  - $i < i' < j' < j$
  - $(i'-i)+(j-j') > 2$ (no stack)
  - there is no basepair (k,l) between (i,j) and (i',j').

- **hairpin loop**
- **stacking**
- **internal loop** (i,j,i',j')



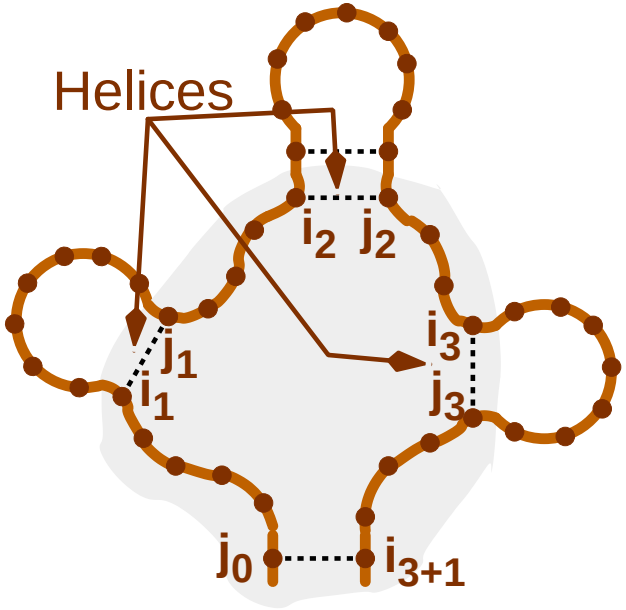
The Zuker Algorithm/Definitionen

- *An internal loop is called* **left** *(***right***, resp.)* **bulge**, *falls* $j = j' + 1$ *oder* $i' = i + 1$.

- *A* k-**multiloop** *consists of multiple base-pairs,* $(i_1, j_1) \dots (i_k, j_k) \in P$ *with a closing basepair* $(j_0, i_{k+1}) \in P$ *with the property that*
  - $\forall\ 0 \leq l \leq k : (j_l < i_{l+1})$
  - $\forall\ 0 \leq l, l' \leq k$ *is true that there is no basepair* $(i', j') \in P$ *with* $i' \in [j_l \dots i_{l+1}]$ *and* $j' \in [j_{l'} \dots i_{l'+1}]$.
- $(i_1, j_1) \dots (i_k, j_k)$ *close the* **helices** *of the multiloop.*

Remarks

- **k-multiloop**



Remark

- *Usually hairpin loops are constrained to a loop sequence of at least 3nt $\Rightarrow$ every hairpin loop $(i,j) \in P$ requires to hold the condition $i < j - 3$.*
- *each secundary 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 are as follows:

- **hairpin loop** (i, j): $eH(i, j)$
- **stacking** (i, j): $eS(i, j, i+1, j-1)$
- **internal loop** (i, j, i, j'): $eL(i, j, i', j')$
- **multiloop**: $eM(j_0, i_1, j_1, \ldots, i_k, j_k, i_{k+1})$

Remark

Multiloop contribution too expensive: exponential explosion!
For prediction use a simplified contribution.

Simplified Energy Contribution of Multiloops

Definition (simplified energy contribution for multiloop)

- **multiloop** $eM = a + bk + ck'$

  a, b, c = weights with

  a = energy contribution for closing the loop

  k = number of helices

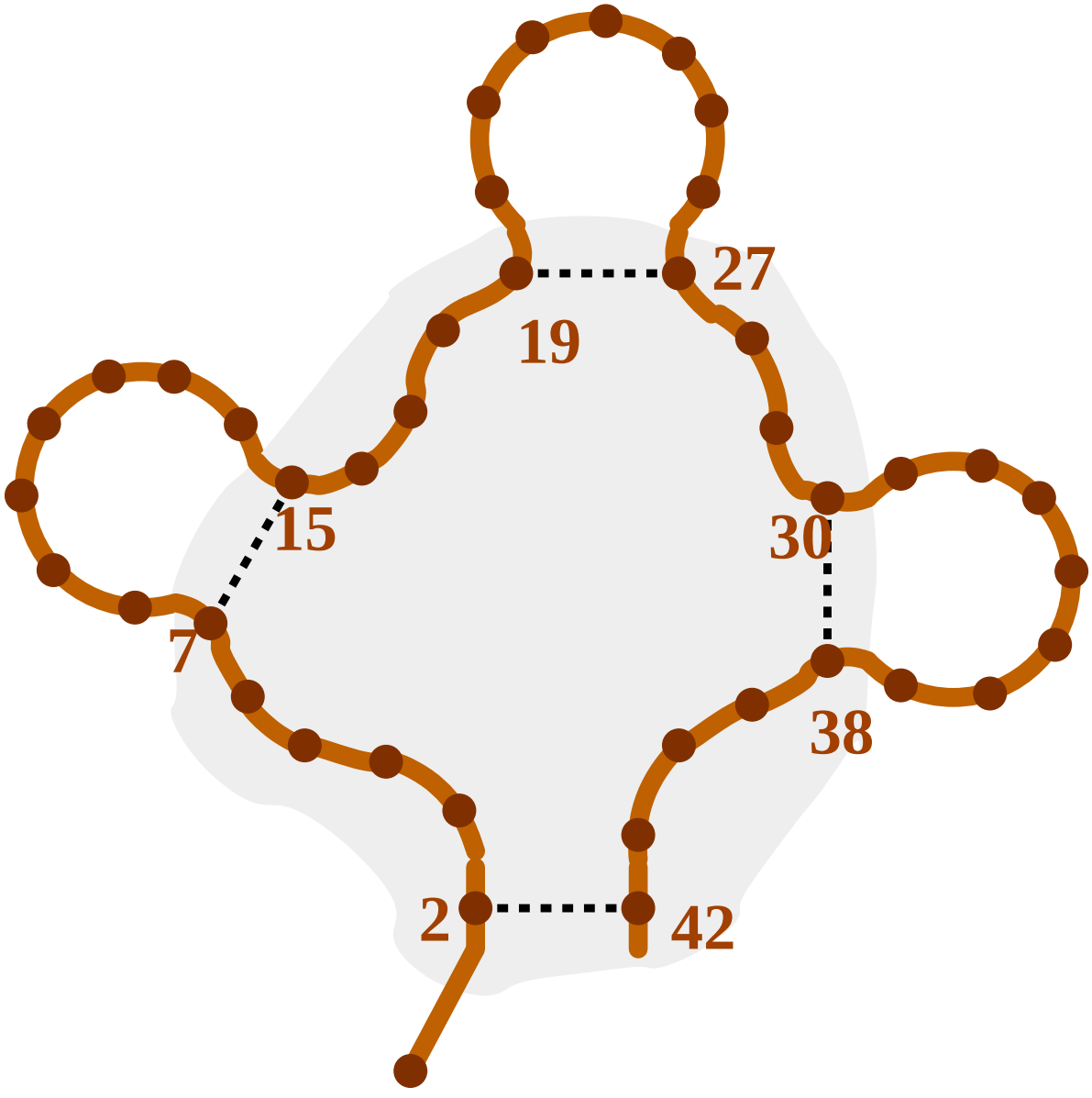
  k' = number of unpaired bases within the loop

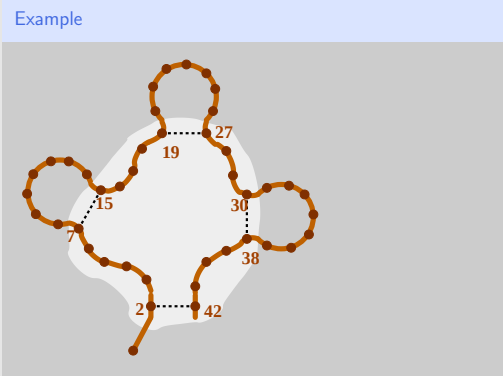
Definition (free energy)

loop free energy: $E_{i,j}^{P}$ = energy of the structure element $Sec(i, j)$

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

Example





- for multiloop: $k' = \sum_{l=0}^{k} i_{l+1} - j_l - 1$

example for simplified energy function of multiloops:

$eM(2, 7, 15, 19, 27, 30, 38, 42) = a + b \cdot 3 + c \cdot 12$
$k' = 12 \Rightarrow$ unpaired bases within big loop
$k = 3 \Rightarrow$ helices within loop
$E^P_{2,42} = eM(2, 7, 15, 19, 27, 30, 38, 42) + eH(7, 15) +$
$eH(19, 27) + eH(30, 38)$
Bem: $(2, 42) \Rightarrow$ external binding

Zuker's Free Energy Minimization Problem

Definition (RNA Structure Prediction (by Energy Minimization))

- IN: RNA sequence S
- OUT: non-crossing RNA structure of S

$$\text{argmin}_{P \text{ of } S} E(P)$$

Remark

- *actually $E_S(P)$: energy of P also depends on S.*
- $\rightarrow$ *assume S fix.*
- *efficient solution: again DP*
- $\rightarrow$ *necessary differences to Nussinov?*
- $\rightarrow$ *define DP by recursion equations*

Matrices for Zuker

In general: Each matrix entry contains the best free energy for subsequence $S_i \ldots S_j$

Definition

For minimizing the energy the following matrices are applied:

$$V_{i,j} = \min \left\{ \; E(P) \;\middle|\; \begin{array}{l} P \text{ structure of } S_i \ldots S_j \\ \text{and } (i,j) \in P \end{array} \right\}$$

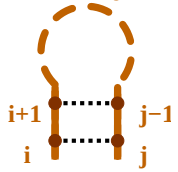
$$WM_{i,j} = \min \left\{ \; E(P) \;\middle|\; \begin{array}{l} P \text{ structure of } S_i \ldots S_j \\ \text{and } S_i \ldots S_j \text{ is real part} \\ \text{of a multiloop} \end{array} \right\}$$

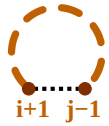
$$W_i = \min \left\{ \; E(P) \;\middle|\; P \text{ structure of } S_1 \ldots S_i \; \right\}$$

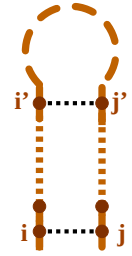
Recursion

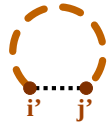
- 

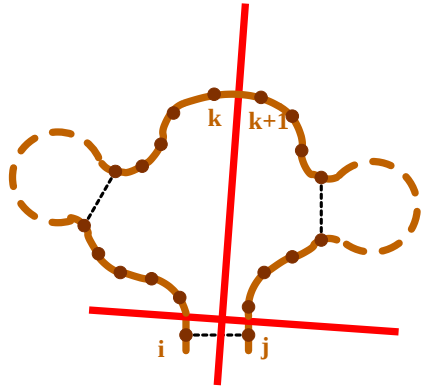
$\Rightarrow$ eH(i,j)

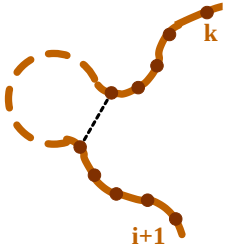
- 

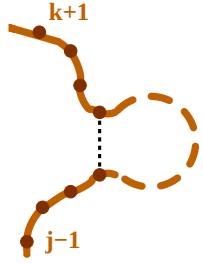
$\Rightarrow$ $eS(i,j,i+1,j-1)$ $+$ 

- 

$\Rightarrow$ $eL(i,j,i',j')$ $+$ 

- 

$\Rightarrow$ 

$+$ 

$+$ a

$\Rightarrow$ $WM_{i+1,k}$ $+$ $WM_{k+1,j-1}$ $+$ a

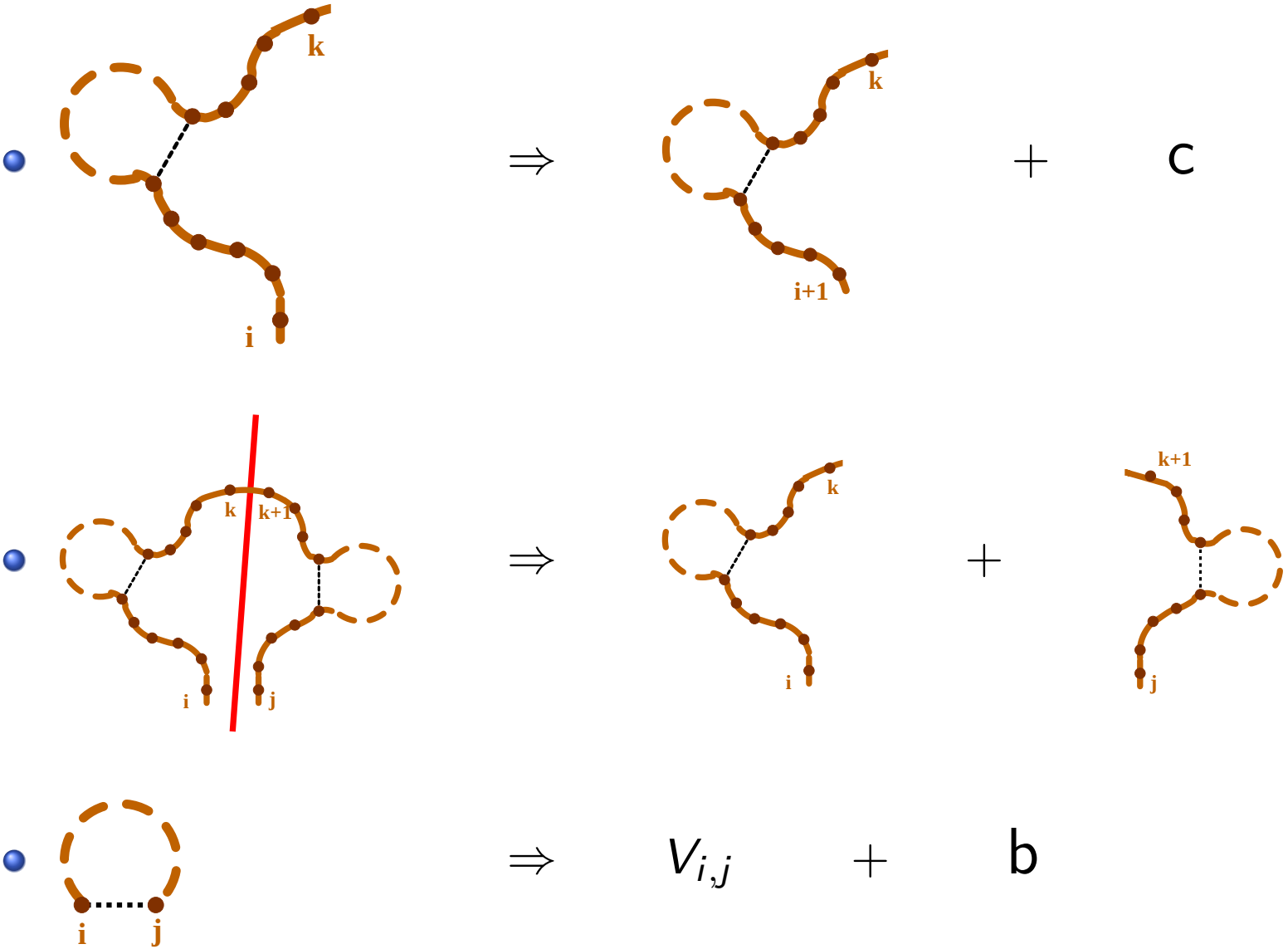
recursion for $V_{i,j}$

$V_{i,j}$: (i,j) closes either hairpin, stacked, internal loop or multiloop

Thus $V_{i,j}$ = minimum over

$$\left\{ \begin{array}{ll} eH(i,j) & \\ eS(i,j,i+1,j-1) + V_{i+1,j-1} & \\ \min\limits_{i<i'<j',\ i'-i+j-j'>2} & \left\{ eL(i,j,i',j') + V_{i',j'} \right\} \\ \min\limits_{i+1<k<j} & \left\{ WM_{i+1,k} + WM_{k+1,j-1} + a \right\} \end{array} \right\}$$

$WM_{i,j}$



Recursion for $WM_{i,j}$

- $WM_{i,j} \Rightarrow S_i \ldots S_j$ is **part** of a multiloop ((i,j) no external basepairing!)
- multiloop must be split at least once, otherwise simple internal loop
- **Idea** cut parts of multiloop until only helices are left over

  $\Rightarrow WM_{i,j} =$ minimum over

$$\left\{ \begin{array}{l} WM_{i+1,j} + c \\ WM_{i,j-1} + c \\ \\ \min\limits_{i<k\leq j} \left\{ WM_{i,k} + WM_{k+1,j} \right\} \\ \\ V_{i,j} + b \\ \\ \end{array} \right\}$$

Remark

- Is it guaranteed that recursion

$$V_{i,j} \Rightarrow \min_{i+1<k<j} \{WM_{i,k} + WM_{k+1,j} + a\}$$

  produces at least 2 helices?
- for that: new WM recursion:

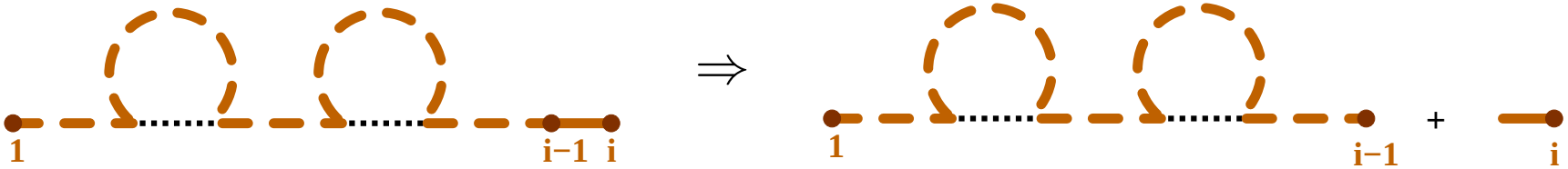
$$WM_{i,j} = \min \left\{ \begin{array}{l} WM_{i+1,j} + c \\ WM_{i,j-1} + c \\ \min\limits_{i<k<j} \{WM_{i,k} + V_{k+1,j}\} \\ V_{i,j} + b \end{array} \right\}$$

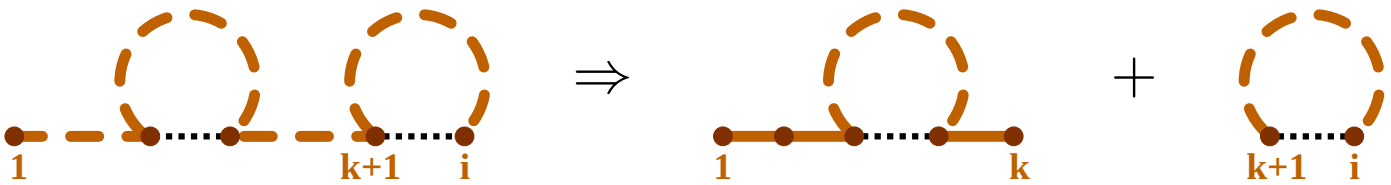
- Other possibilities?

Recursive Calculation of W-Matrix

- so far: several parallel helices only in multiloops

  $\Rightarrow$ here case of no closing base pair

- 

- 

- **Rekursionsgleichung für W_i:**

$$W_i = \min \begin{cases} W_{i-1} \\ \min\limits_{0 \le k < i} \{W_k + V_{k+1,i}\} \end{cases}$$

Complexity

	space	time
W-matrix	$O(n)$	$O(n^2)$
WM-matrix	$O(n^2)$	$O(n^3)$
V-matrix	$O(n^2)$	$O(n^4)$

Total complexity?

Remark

- *reason for* n^4*:*

  number of runs $1 \leq i' < j' \leq n$ *is given by:*
  for $i' = 1$ *there are* $n-1$ *values for* j'
  for $i' = 2$ *there are* $n-2$ *values for* j'
  $\#(i', j') = \sum_{j'=1}^{n'} j' = \frac{n'(n'-1)}{2}$, *where* $n' = n-1$.

- *in practice:* $O(n^4)$ *too expensive: restrict loop size.*
  Which loop? Consequence?

How to get MFE structure?

As usual: by Traceback.

- Traceback similar to Nussinov, however 3 matrices/states
- use stack of entries (i,j,s)
- meaning of stack entry: determine base pairs in structure of $S_i \ldots S_j$, when starting from matrix/state s

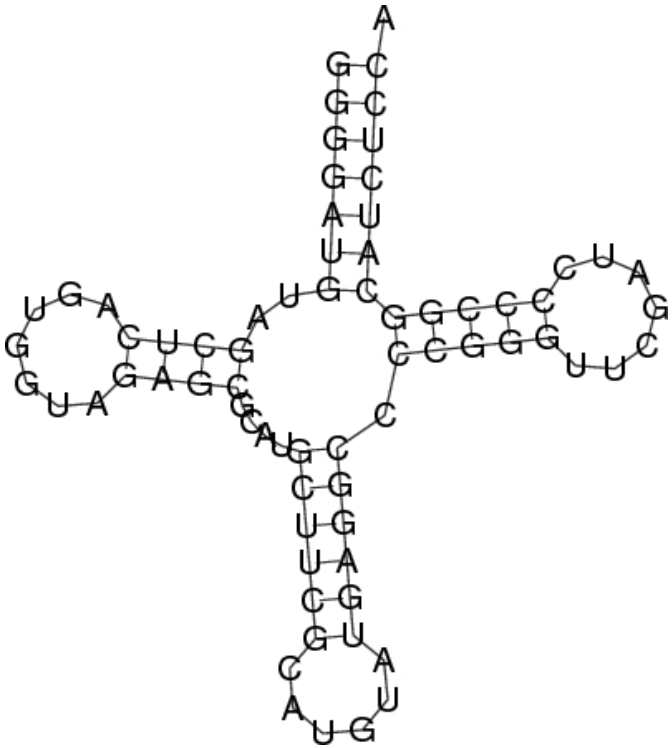
Programs:

- Zukers Mfold
- Vienna RNA Package
- Example:

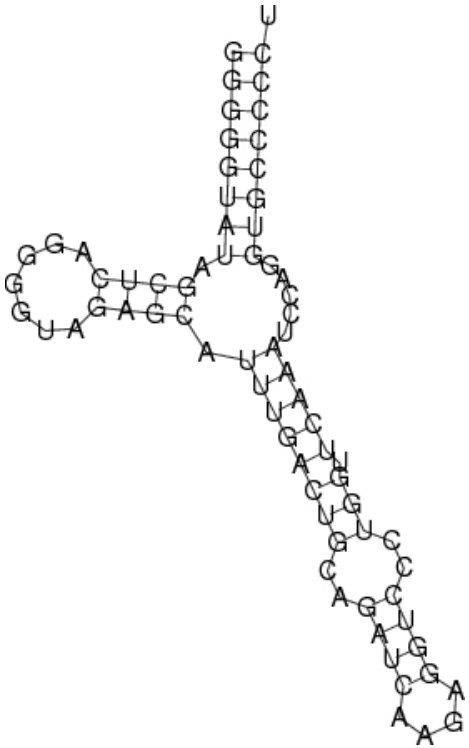
```
(95) rb@setepenra  > RNAfold

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
GGGGGTATAGCTCAGGGGTAGAGCATTTGACTGCAGATCAAGAGGTCCCTGGTTCAAATCCAGGTGCCCCCT
length = 72
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
(((((((((.((((.......)))).(((((((((..((((....))))..))))))))).....)))))))).
minimum free energy = -26.70 kcal/mol
```

Example: tRNAs

- Mouse tRNA-ALA:



- Mouse tRNA-CYS:



Problem of suboptimal Structures

- RNA structures variable

$\Rightarrow$ energy minimization does not always give correct result

$\Rightarrow$ structure with minimal energy only a high probable one

- How probable is a structure?
- in the following: $\mathcal{P} = \{P_1, \ldots, P_m\}$ set of all RNA structures with given sequence S
- $E(P_i)$ is free energy of structure P_i
- wanted:
  - probability p_i that S is structure of P_i
  - **Problem:** How to measure quality of distribution?
  - **Wanted:** distribution, which makes least number of *unproven* assumption
    $\Rightarrow$ requires measurement for the information content of a distribution
  - **Solution:** maximum entropy

Excursion: Entropy

- given probability space $\Omega = \{e_1, \ldots, e_n\}$
  p_i is probability for event e_i
- 2 persons A,B A knows which event Ω occured B used yes/no question to determine this event
- entropy measures the complexity of this situation, i.e., how much the information of A worth is

- $H_0(\vec{p})$ = number of yes/no questions for A, which B must use to determine the real event, given *known* (by B) distribution $\vec{p}$

Problem

- all possible distributions for Ω of size 2 have two possible entropy values:
  $\Rightarrow$ not fine enough

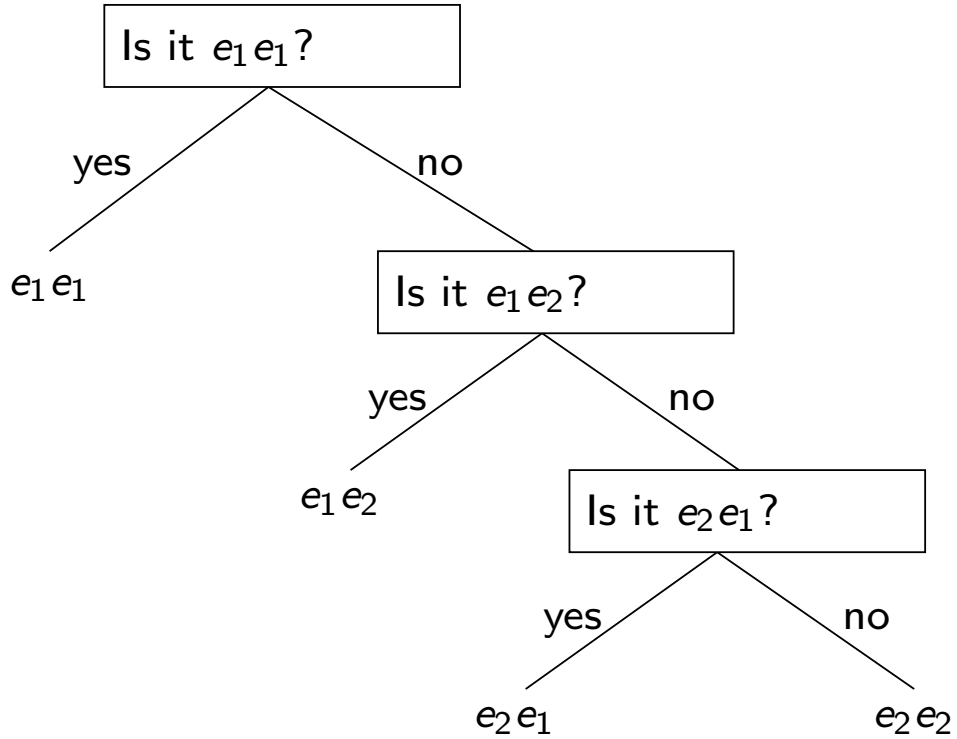
Example

Consider $\Omega = \{e_1, e_2\}$ and the following 2 distributions:

- $\vec{p_1} = (1, 0) \Rightarrow B$ doesn't need to ask, i.e. $H_0(\vec{p_1}) = 0$
- $\vec{p_2} = (\frac{1}{2}, \frac{1}{2}) \Rightarrow B$ asks: "Is it e_1?", i.e. $H_0(\vec{p_2}) = 1$
- $\vec{p_3} = (0.9999, 0.0001 \Rightarrow H_0(\vec{p_3}) = 1$
  $\Rightarrow$ B cannot benefit from his knowledge about the distribution

Now: k-times Independent Iteration

- for $\vec{p_3} = (0.9999, 0.00001)$: possible advantage from fact that the result is nearly always e_1:



- then the expected number of questions for two consecutive events is:

$$E[Quest.] = 0.9999^2 \cdot 1 + 0.9999 \cdot 0.0001 \cdot 2 + 0.0001 \cdot 0.9999 \cdot 3 + 0.0001^2 \cdot 3 = 1.0003$$

Real and Ideal Entropy

Definition (Real Entropy)

The *real entropy* $H_0^k(\vec{p})$ for k-times iteration of $\vec{p}$ is defined as the minimal number of questions that have to be asked on average to determine the sequence of k events (given an *optimal strategy* for performing the questions), divided by k

Definition (Ideal Entropy)

The ideal entropy of a distribution $\vec{p} = (p_1, \ldots, p_m)$ is defined by

$$H(\vec{p}) = -\sum_i p_i \log p_i$$

Theorem

(Shannon)

$$H(\vec{p}) = \lim_{k \to \infty} H_0^k(\vec{p})$$