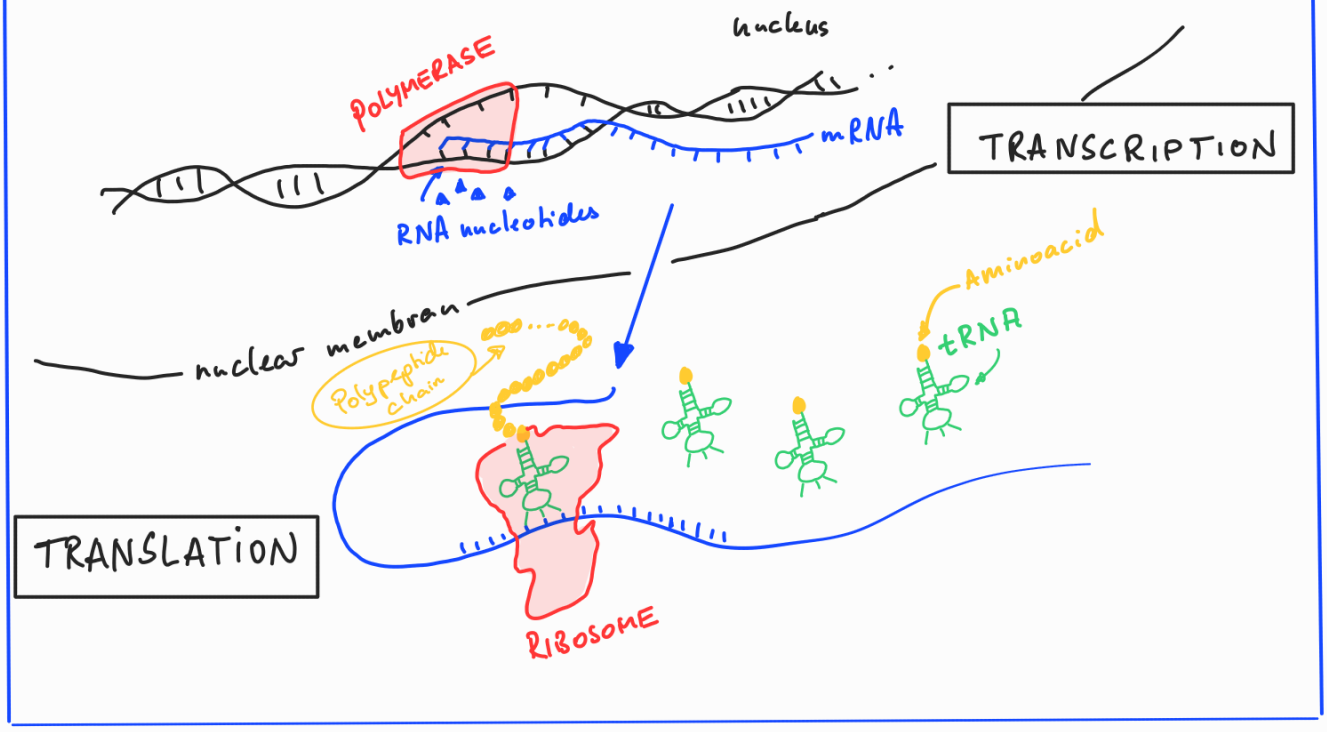
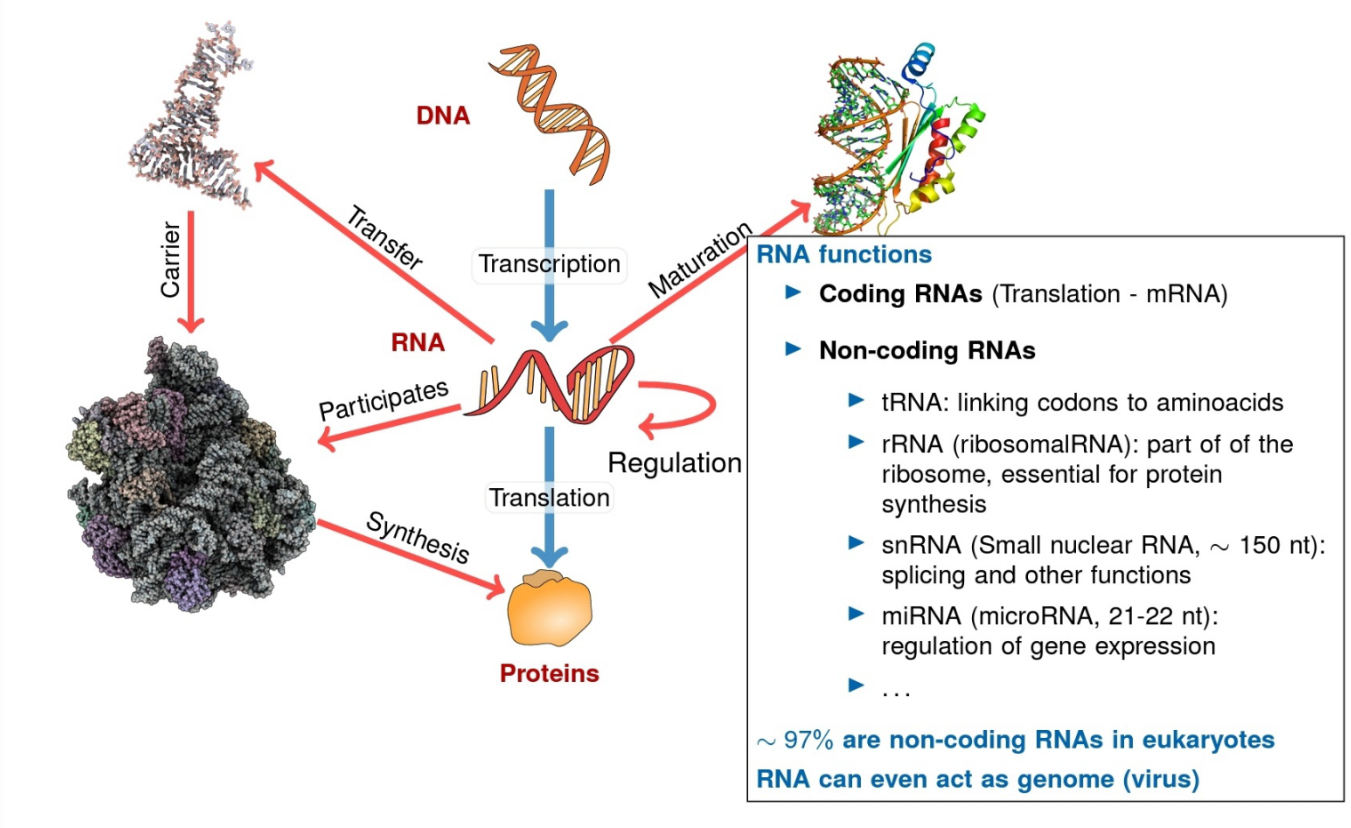
RNA (= Ribonucleic Acid)

1. Basics

Fundamental Dogma of molecular biology:

DNA —transcription→ RNA —translation→ protein





RNA functions

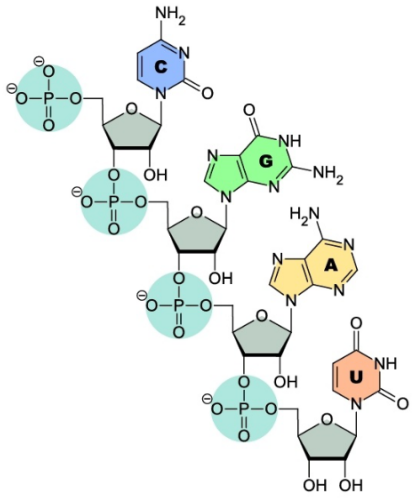
- **Coding RNAs** (Translation - mRNA)
- **Non-coding RNAs**
  - tRNA: linking codons to aminoacids
  - rRNA (ribosomalRNA): part of of the ribosome, essential for protein synthesis
  - snRNA (Small nuclear RNA, ~ 150 nt): splicing and other functions
  - miRNA (microRNA, 21-22 nt): regulation of gene expression
  - ...

~ 97% are non-coding RNAs in eukaryotes

RNA can even act as genome (virus)

⇒ **RNA world hypothesis:** self-replicating RNA's are precursors to all current life on earth

- single-stranded polymer
- polymer made of nucleotides+backbone
- nucleotides: guanine (G), adenine (A), uracil (U), cytosine (C)
- backbone: alternating sugar (ribose) and phospat groups (related to phosphoric acid) nucleotides are attached to sugar
- the nucleotides of polymer can bind (A-U, C-G, G-U) via hydrogen bonds, i.e., unlike DNA it is more often found in nature as a single-strand folded unto itself, rather than a paired double-strand.



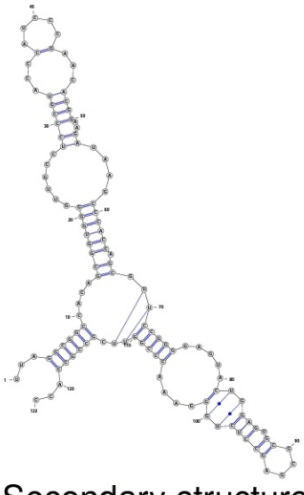
RNA can fold into complex 3D structures that are essential to its function(s).

Structure determines function!

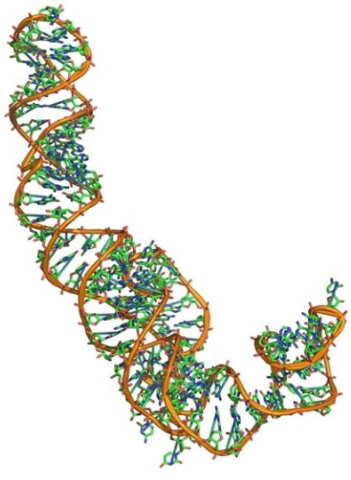
Three* levels of representation:

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary structure



Tertiary structure

Source: 5s rRNA (PDB 1K73:B)

≙ 3D structure, refering to locations of atoms in 3D-space.

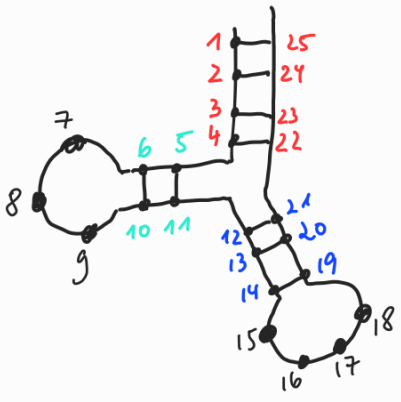
DEF:

$\mathbb{A} := \{A, C, G, U\}$ and $\mathbb{B} := \{AU, UA, GC, CG, GU, UG\}$

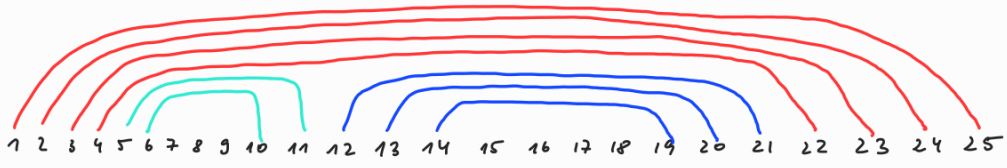
A **primary structure** (of length n) is a sequence $s = s_1 \dots s_n \in \mathbb{A}^n$.

A **secondary structure** $\mathscr{S}$ is a collection of ordered pairs (i, j), where $1 \leq i < j \leq n$, s.t. the following properties hold:

1. If $(i, j), (k, l) \in \mathscr{S}$, then it is not the case that $i < k < j < l$.
2. If $(i, j), (k, l) \in \mathscr{S}$ and $i \in (k, l)$ implies that $i = k$ and $j = l$.
3. If $(i, j) \in \mathscr{S}$, then $j > i + \theta$, where θ is a fixed integer and usually taken to be 3.



$\mathcal{S} = \{(1,23), (2,22), (3,21), (4,22)$
$(5,11), (6,10)$
$(12,21), (13,20), (14,19)\}$



NOT allowed:

$i < k < j < \ell$ [Cond. 1]

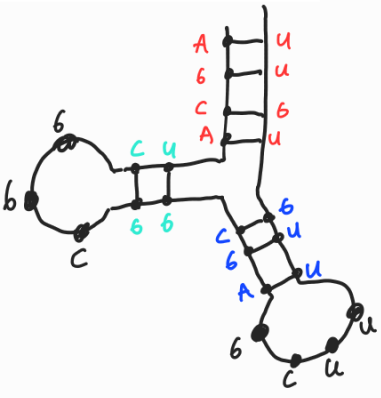
[Cond. 2]

DEF:

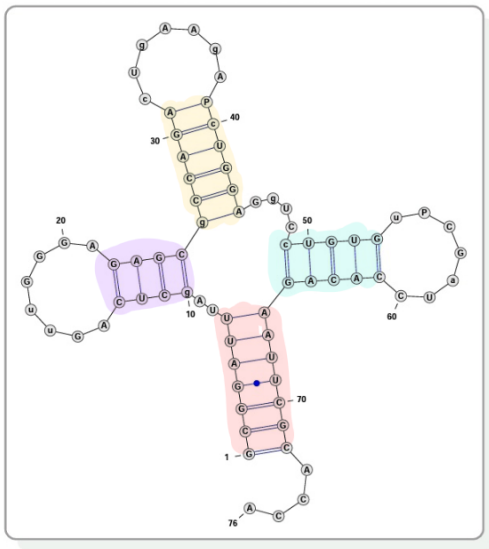
A secondary structure $\mathscr{S}$ for a given sequence $s = s_1 \dots s_n \in \mathbb{A}^n$ is a secondary structure fullfilling in addition

4. If $(i, j) \in \mathscr{S}$, then $s_i s_j \in \mathbb{B}$.

Structure $\mathcal{S}$:



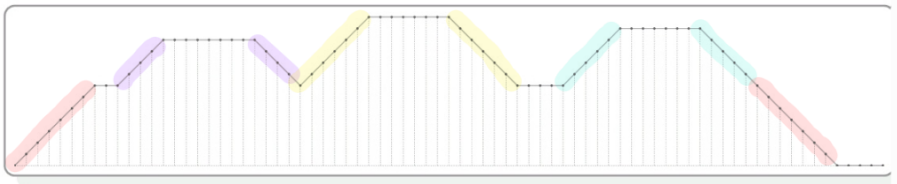
Different repr. of. Sec. Str.



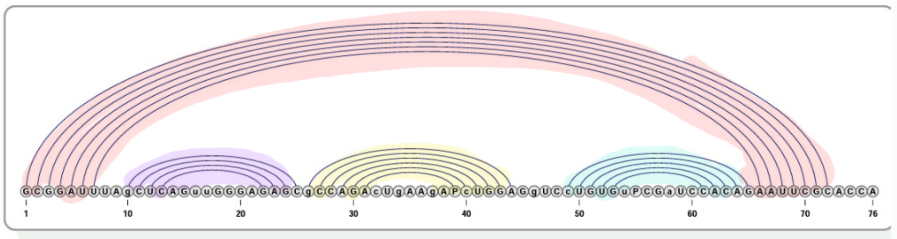
Outer-planar graphs

Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

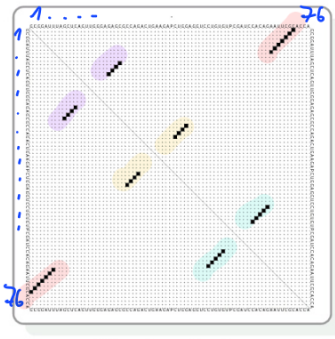
Motzkin words*



Positive 1D meanders* over $\mathcal{S} = \{+1, -1, 0\}$

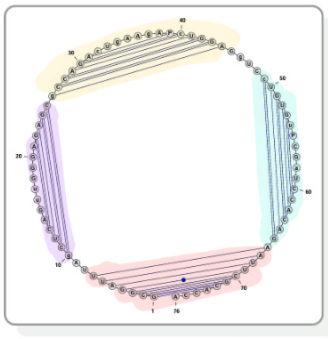


Non-crossing arc-annotated sequences*



Dot plots

Adjacency matrices*



Non-crossing arc diagrams*

Supporting intuitions

Different representations

Common combinatorial structure

*Additional steric constraints

2. Counting Sec. Structures

Thm: Let $S(n)$ denote the number of secondary structures of length n & where $\theta = 1$.

Then,

$S(0) = 0$ [no nucleotides, no structure]

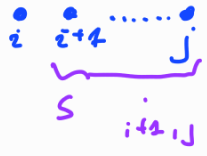
$S(1) = 1$ [$\hat{=}$ •]

$$\forall n \geq 1: \quad S(n+1) = S(n) + S(n-1) + \sum_{k=2}^{n-1} S(k-1)\,S(n-k)$$

proof:

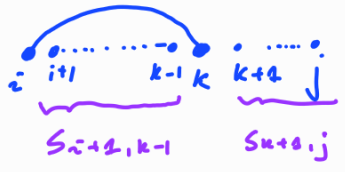
[general observation] Let S_{ij} = # of possible substructures on $i \cdots j$ $(i < j)$

2 cases: ▶ i unpaired with any $k \in \{i+2, \dots j\}$

$\Rightarrow$ 

$\Rightarrow S_{ij} = S_{i+1,j}$

▶ i paired with some $k \in \{i+2, \dots j\}$ $(\theta = 1)$

$\Rightarrow$ 

$\Rightarrow S_{ij} = S_{i+2,k-1} \cdot S_{k+1,j}$

[can be multiplied since „ ⌒ " forbidden]

$$\Rightarrow S_{ij} = S_{i+1,j} + \sum_{i+2}^{j} \left(S_{i+2,k-1} \cdot S_{k+1,j}\right).$$

how Induction on n:

Base Case: $n = 1$: $S(n+1) = S(1) + S(0) = 1$

$\hat{=}$ structure: • • & this is the only structure since $\theta = 1$.

✓

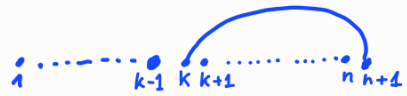
Ind. hyp.: $S(l)$ correct for all $l \in \{1 \dots n\}$.

Ind. step : $S(n+1)$

2 cases: ▶ pos. $n+1$ not paired with any k, $k \in \{1 \dots n-1\}$

$\Rightarrow S(n+1) = S(n)$

▶ pos. $n+1$ paired with some $k \in \{1 \dots n-2\}$



$\Rightarrow S(n+1) = S(k-1) \cdot S(n-k)$

$$\Rightarrow S(n+1) = S(n) + S(n-1) + \sum_{k=2}^{n-1} \left(S(k-1) \cdot S(n-k)\right)$$

special case $k=1$

□

Lemma $S(1) = 1$ & $\forall n \geq 2 : S(n) \geq 2^{n-2}$, $\theta = 1$

proof: by Induction on n

$n=2$: .. $S(2) = 1 \geq 2^{2-2} = 1$ ✓

$n=3$: ... , $S(3) = 2 = 2^{3-2}$ ✓

$n=4$: .... , , , $4 = 2^{4-2}$ ✓

$n=5$: ..... , , ,

 , ,

$8 = 2^{5-2}$ ✓

Assume statement true for $n \geq 5$

[Exercise]

□

$\Rightarrow$ There exist many possible second. structures!

Lemma: $S(n,k)$ = # sec. structures of length n with exactly k base pairs.

Then, $S(n,0) = 1 \quad \forall \; n$

$S(n,k) = 0 \quad \forall \; k \geq \frac{n}{2}$

& for all $n \geq 2$ holds:

$$S(n+1,k+1) = S(n,k+1) + \sum_{j=1}^{n-1} \left(\sum_{i=0}^{k} S(j-1,i)\, S(n-j,k-i) \right)$$

proof: Exercise

3. REALIZABILITY

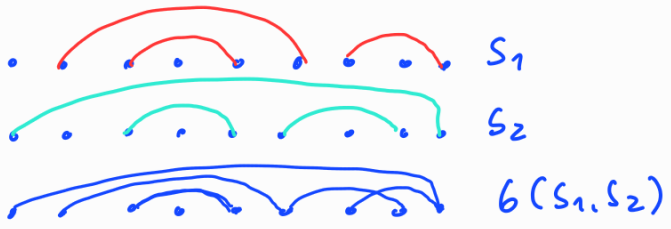
Def: $s \in A = \{A,U,G,C\}^n$ **realizes** secondary structure S & S is **compatible** with s of length n if, for all $(i,j) \in S$ it holds that $s_i s_j \in \mathcal{B} = \{AU, UA, GC, CG, GU, UG\}$.

$G(S_1 \dots S_k)$ with sec. structures $S_1 \dots S_k$ of length n.

↳ vertices: $1 \dots n$

edges: $\{i,j\}$ if exist (i,j) basepair in at least on S_ℓ, $1 \le \ell \le k$.

Exmpl:



Clearly, for all sec. structures S exists a sequence s that realizes S [since $G(S) = (\{1 \dots n\}, S)$ is bipartite]

Intersection Thm [Reidys 95]

Let $\mathcal{C}(S)$ = set of sequ. that realize sec. str. S

Then $\mathcal{C}(S_1) \cap \mathcal{C}(S_2) \neq \emptyset$ $\forall$ Sec.str. S_1, S_2.

proof: We show first that $G(S_1,S_2) = (V,E)$ is bipartite, i.e.,

$\exists$ bipartition of $V = \{1 \dots n\}$ into V_1, V_2

s.t. $(i,j) \in E$ implies $i \in V_1$ & $j \in V_2$ or vice versa.

To do so, it suffices to consider connected components of $G(S_1,S_2)$.

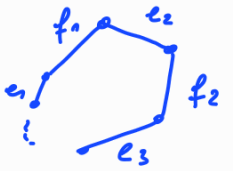
Note, in $G(S_i)$ each vertex has degree ≤ 1

$\Rightarrow$ in $G(S_1,S_2)$, each vertex has degree ≤ 2

$\Rightarrow$ connected components are single vertices [bipartite ✓]

paths [bip. ✓]

or cycles.

cycle: must look like 

where $e_i \in S_1$, $f_i \in S_2$

$\Rightarrow$ cycles are of even length

$\Rightarrow$ cycles bipartite

$\Rightarrow$ $\exists$ bip. V_1, V_2 of V st $\forall$ edges $i \text{—} j$: $i \in V_1$, $j \in V_2$ or vice versa

$\Rightarrow$ can take $s_i = u \quad \forall\, i \in V_1$

$s_j = b \quad \forall\, j \in V_2$

$\Rightarrow$ $s_1 \dots s_n$ realises both S_1 & S_2 /□

Generalized Intersection Thm [Flamm et al 2001]

Let $\mathcal{C}(S)$ = set of seqn. that realize sec. str. S

& $S_1 \dots S_k$ sec. structures of length n

Then,

$$\bigcap_{i=1}^{k} \mathcal{C}(S_i) \neq \emptyset \iff \mathcal{G}(S_1 \dots S_k) \text{ bipartite.}$$

proof: '$\Leftarrow$' $\mathcal{G}(S \dots S_k)$ bip $\Rightarrow$ $\exists$ partition V_1, V_2 of $\{1 \dots n\}$

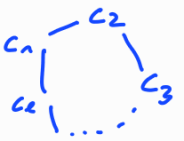
st $(i,j) \in S_\ell \iff i \in V_1, j \in V_2$ or vice versa.

$\Rightarrow$ label $s_i = x$, $xy \in \mathcal{B}$

$s_j = y$

$\Rightarrow$ $\forall\, (i,j)$: $s_i s_j \in \mathcal{B} \Rightarrow s = s_1 \dots s_n$ realizes all $S_1 \dots S_k$.

'$\Rightarrow$' Assume, for contradiction, that $\mathcal{G}(S_1 \dots S_k)$ is not bipartite.

$\Rightarrow$ $\exists$ cycle $\mathcal{C}$ 

such that n is odd.

IF we want to find corresp. seqn. $\tilde{s}_1 \dots \tilde{s}_\ell \in s_1 \dots s_n$

then $\tilde{s}_i \in \{A, C, G, U\}$.

Consider the possible "bp-graph"

$A - U - 6 - C$

that is: $\tilde{s}_i = A \Rightarrow \tilde{s}_{i+2} = U$ ($\tilde{s}_i$ for c_i in ℓ; $\tilde{s}_{i+2}$ for c_{i+1} in ℓ)

$\tilde{s}_i = U \Rightarrow \tilde{s}_{i+2} \in \{A, 6\}$

$\tilde{s}_i = 6 \Rightarrow \tilde{s}_{i+1} \in \{U, C\}$

$\tilde{s}_i = C \Rightarrow \tilde{s}_{i+1} = G.$

in particular, to determine $\tilde{s}_{i+1}$ for given $\tilde{s}_i \in \{A, G, U\}$ we must "follow" the edges in $A - U - 6 - C$

Q: when can a letter chosen for $\tilde{s}_i$ occur a 2nd time?

eg: $\tilde{s}_i = A$

$A - U - 6 - C$

$\tilde{s}_i \rightleftarrows \tilde{s}_{i+1}$ [2 "steps"]

$\tilde{s}_i \to \tilde{s}_{i+1} \to \tilde{s}_{i+2}$

$\tilde{s}_{i+3} \to \tilde{s}_{i+4} \to \tilde{s}_{i+5}$ [8 steps]

$\tilde{s}_{i+8} \leftarrow \tilde{s}_{i+7} \leftarrow \tilde{s}_{i+6}$

$\Rightarrow$ A can only occur again after an even nr. of "step"

[same for U, 6, C]

$\Rightarrow$ $\tilde{s}_i \to \tilde{s}_{i+1} \to \tilde{s}_{i+2}$, ℓ

these "step" correspond to consecutive edges in ℓ

& in particular holds for the walk from $\tilde{s}_i$ to $\tilde{s}_i$ in ℓ

$\Rightarrow$ ℓ must be even. ↯

/17

4. Computing Sec. Str.

Aim: find "most likely" structure for given sequence.

"most likely" wrt. to what ??

In general one wants to find "most stable" structure.

A naïve approach, that laid foundation for many more sophisticated & realistic approaches: max # weighted basepairs

As more hydrogenbonds as more stable ≙ less free energy.

		weight w
Watson-Crick basepairs:	G≡C	3
	A=U	2
wobble basepairs:	G-U	1
	other	0

⊗

The Nussinov Algorithm [Ruth Nussinov]

Given a sequence $s \in \mathcal{A} = \{A, C, G, U\}^n$,
find a secondary structure S that has a maximum number of basepair among all structures

↳ ≙ $w(s_i, s_j) = 1$ iff $s_i s_j \in B$.

if needed modify to "Jacobson energy model" & use weights 3, 2, 1, 0 [according to table ⊗] (or other weights).

$$N_{i,t} = 0 \quad \forall\, t \in [i, i+\theta]$$

$$N_{i,j} = \max \begin{cases} N_{i+1,j} & \text{[case 1]} \\ \max\limits_{i+\theta+1 \le k \le j} w(s_i, s_k) + N_{i+1,k-1} + N_{k+1,j} & \text{[case 2]} \end{cases}$$

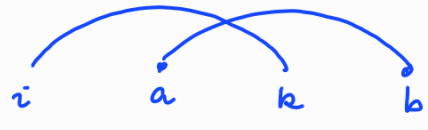
proof-sketch: by Induction show the opt. structure for all in interval $[i, j]$:

Assume property holds for all $[i', j']$ with $j' - i' < \ell$

consider $[i, j]$ with $j - i = \ell$

- i unpaired [case 1]
  - $\Rightarrow N_{i,j} = N_{i+1,j}$
- i paired with some k $\Rightarrow$ $k > i + \theta$
  - $\rightarrow N_{ij} = w(s_i s_k) + N_{i+1,k-1} + N_{k+1,j}$ [case 2]

in fact any bp (a, b) with $a \in [i+1, k-1]$, $b \in [k+1, j]$



would cross

□

FOR ($i = 0 \dots n-1$) : $N_{i,t} = N_{i+1,t} = 0 \quad \forall\, t \in [i, i+\theta]$ ($s_1 \dots s_n$)

" to cover special cases $i \;\; k=j$ "

FOR ($L = 0, \dots n-1$)

  FOR ($i = 0 \dots n-L-1$)

    $j = i + L + 1$

    $N_{i,j} = \max \begin{cases} N_{i+1,j} & \text{// } i \text{ unpaired} \\ \max\limits_{i+\theta+1 \le k \le j} \left(w(s_i, s_k) + N_{i+1,k-1} + N_{k+1,j} \right) & \text{// } i \dots j \end{cases}$

		0	1	2	3	4	5	6
		G	G	U	C	C	A	C
0	G	0	0	1	1	*		
1	G	0	0	0	1	1		
2	U		0	0	0	0	1	
3	C			0	0	0	0	0
4	C				0	0	0	0
5	A					0	0	0
6	C						0	0

hue.

$w(s_i, s_j) = 1 \quad \forall\, s_i s_j \in \mathcal{B}$.

$\theta = 1$ init: $N_{ii} = 0$, $N_{i,i+1} = 0$

$L=1,\ i=0,\ j=2$: $N_{0,2} = \max\left(N_{1,2},\ \max_{0+1+1 \le k \le 2} \left(1 + N_{0,0} + N_{3,2} \right) \right)$

" special case i pairs with last "

$k = 2$

0 2 paired (possible $w(0,2) = 1$)

0 1 2) not paired, opt subst.

0 1 2 opt jump

$N_{1,2} = 0$ current opt.

$N_{0,2} = 1$

" ... and so on ... "

$L=3,\ i=0,\ j=4$: $\boxed{N_{04}} = \max\left(N_{14},\ \max_{2\le k\le 4}(\ldots)\right)$

$= \max\left(N_{14} \overset{!}{=} 1,\ \begin{cases} w_{02} + N_{1,1} + N_{3,4} = 1+0+0 \\ w_{03} + N_{1,2} + N_{4,4} = 1+0+0 \\ w_{04} + N_{1,3} + N_{5,4} = 1+1+0 \end{cases}\right)$

$i=0$ not paired

i - paired with k

$= \boxed{2}$

" 0 pairs with 4 while 1 pairs with 3 "

" and so on "

		0	1	2	3	4	5	6
		G	G	U	C	C	A	C
0	G	0	0	1	1	2	2	2
1	G	0	0	0	1	1	1	2
2	U		0	0	0	0	1	1
3	C			0	0	0	0	0
4	C				0	0	0	0
5	A					0	0	0
6	C						0	0

Final value = max # BP

$\Rightarrow$ max #BP of sequence GGUCCAC is 2

But how does such a structure look like?

$\Rightarrow$ TRACE BACK.

```
i = 0, j = SIZE - 1
TRACEBACK (N, i, j)
    IF ( N_{i,j} = N_{i+1,j} )          // i unpaired
        TRACEBACK (N, i+1, j)
    ELSE
        FOR ( k = i + θ + 1, ..., j )
            IF (N_{ij} = w_{ik} + N_{i+1,k-1} + N_{k+1,j} )   // (i,k) bp
                print ("bp i-k")
                TRACEBACK (N, i+1, k-1)
                TRACEBACK (N, k+1, j)
```

[just sketch! needed: mark i, j as part of bp & k+1 < size]

		0	1	2	3	4	5	6
		G	G	U	C	C	A	C
0	G	0	0	1	1	2	2	2
1	G	0	0	0	1	1	1	2
2	U		0	0	0	0	1	1
3	C			0	0	0	0	0
4	C				0	0	0	0
5	A					0	0	0
6	C						0	0

$i = 0,\ j = size - 1 = 7 - 1 = 6$ (init)

$N_{06} = N_{1,6}$ ✓ // 0 not paired & opt substructure $N_{1,6}$

⇒ TRACEBACK $(N, 1, 6)$

$N_{16} = N_{26}$? no!

⇒ 1 paired with some k.
find possible k.

$k = i + \theta + 1 \dots j$
(3 … 6)

$k = 3$ // $N_{ij} = w_{ik} + N_{i+1,k-1} + N_{k+1,j}$

$N_{16} = w_{13} + N_{22} + N_{46}$
$= 1 + 0 + 0$ no!

$k = 4$

$N_{16} = w_{14} + N_{23} + N_{56}$
$= 1 + 0 + 0$ no!

$k = 5$

$N_{16} = w_{15} + N_{24} + N_{6,6}$
$= 0 + 0 + 0$ no!

$k = 6$

$N_{16} = w_{16} + N_{25} + N_{7,6}$
$= 1$

in this case placeholder set to 0 $k+1 > j$

recurse on N_{25} (+ $N_{k+1,j}$ if $k+1 < j$)

i pairs with j = 6 (1 2 … 5 6)

⇒ get 

N_{25} : $N_{25} = N_{35}$?
" "
1 0 no.

$k = 4, 5$

$k = 4$ $(i = 2, j = 5)$

$N_{25} = W_{24} + N_{33} + N_{55}$
" " " "
1 0 0 0 N_0,

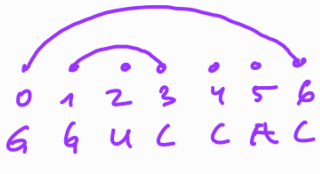
$k = 5$

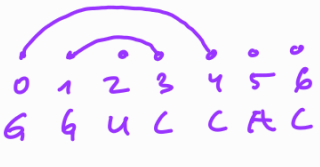
$N_{25} = W_{25} + N_{34} + \text{"}N_{65}\text{"}$
" " " "
1 1 0 0
yes ⇒ 2 5 pair.

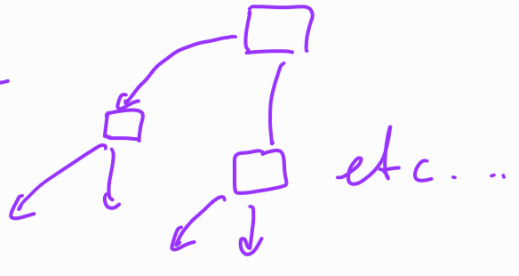
& so on get

get finally with THIS traceback 

other optimal structures:

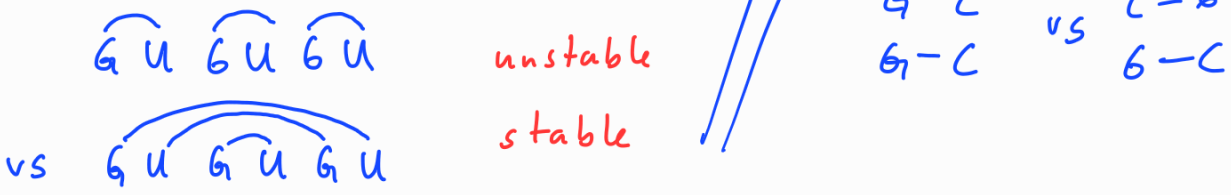




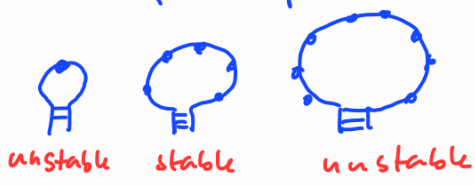
Traceback can in general look like 

Drawbacks of this approach: do not always find "biological relevant" structures.

- "stacking" not considered.



- size of "loops" not considered



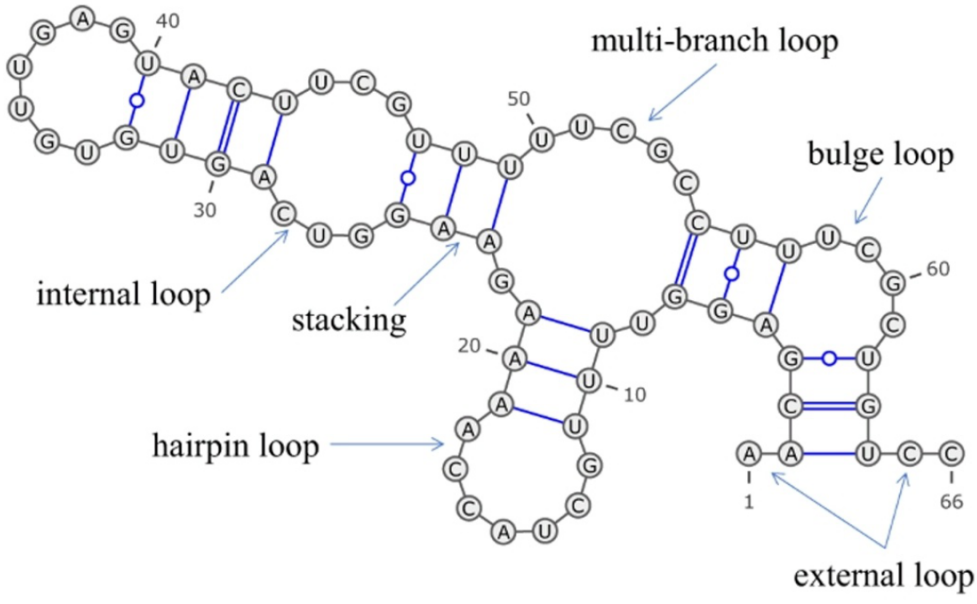
MFE-Folding

Stability of RNA structure ≙ thermodyn. stability

free-energy quantifies amount of free energy that is released by building base pairs.

AIM: Find structure for given sequence that minimizes free energy: minimum-free-energy (MFE)

Define energy model for RNA that takes into account local energy contributions from loop and stacking regions.



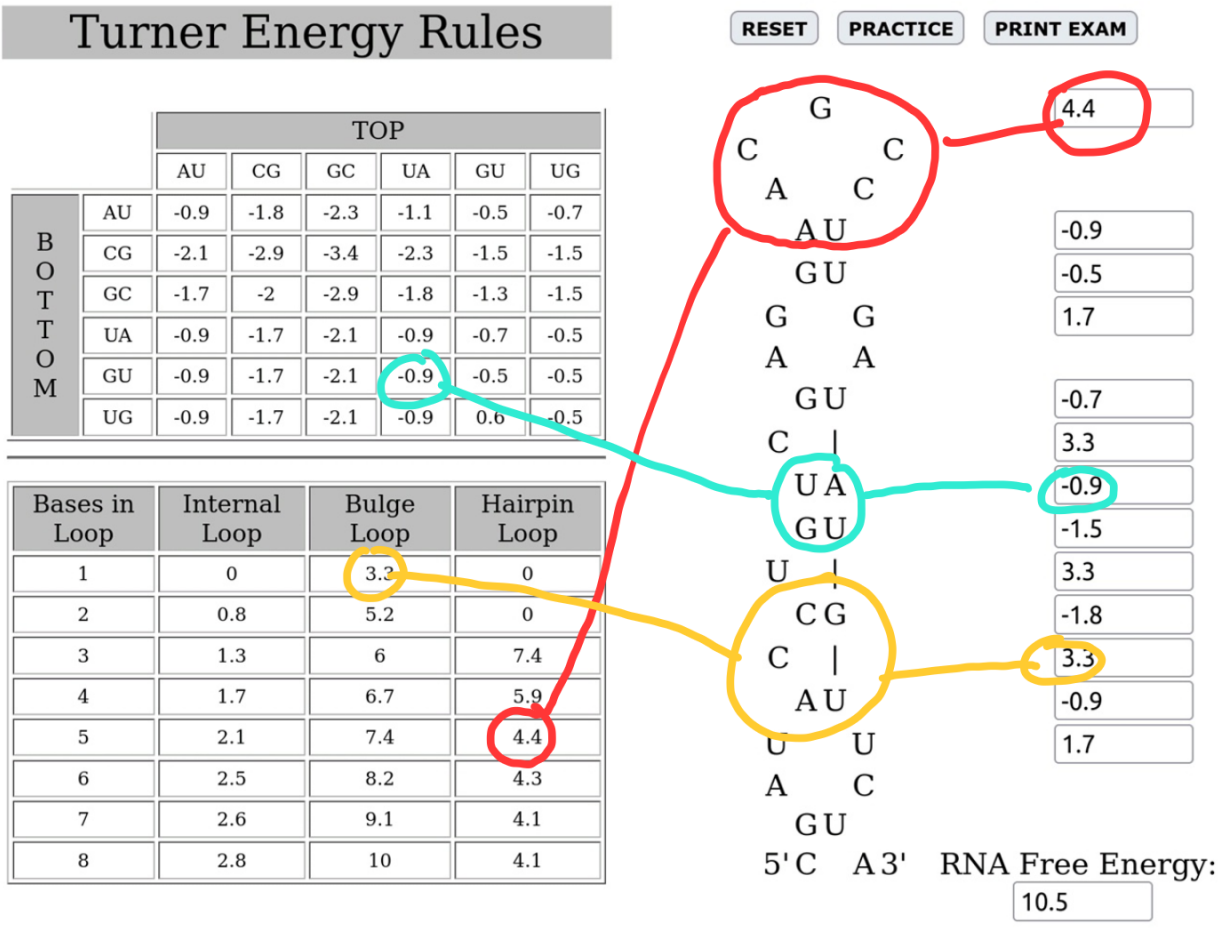
- More realistic: thermodynamics and statistical mechanics.
- Stability of an RNA sec.str. coincides with thermodynamic stability
- Quantified as the amount of free energy released/used by forming bp.

The Turner rules are a set of experimentally determined parameters which allow us to predict the stability of RNA secondary structures.

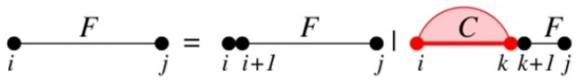
Turner Energy Rules

BOTTOM \ TOP	AU	CG	GC	UA	GU	UG
AU	-0.9	-1.8	-2.3	-1.1	-0.5	-0.7
CG	-2.1	-2.9	-3.4	-2.3	-1.5	-1.5
GC	-1.7	-2	-2.9	-1.8	-1.3	-1.5
UA	-0.9	-1.7	-2.1	-0.9	-0.7	-0.5
GU	-0.9	-1.7	-2.1	-0.9	-0.5	-0.5
UG	-0.9	-1.7	-2.1	-0.9	0.6	-0.5

Bases in Loop	Internal Loop	Bulge Loop	Hairpin Loop
1	0	3.3	0
2	0.8	5.2	0
3	1.3	6	7.4
4	1.7	6.7	5.9
5	2.1	7.4	4.4
6	2.5	8.2	4.3
7	2.6	9.1	4.1
8	2.8	10	4.1

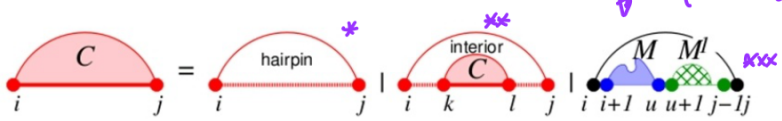


https://www.kelleybioinfo.org/



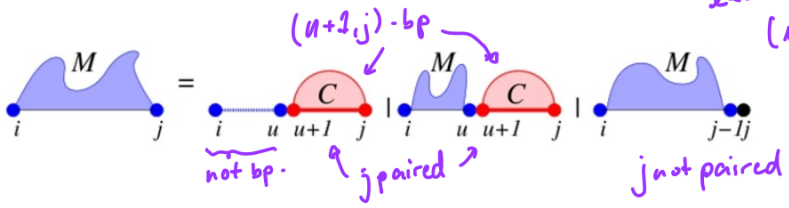
$$F_{i,j} = \min \left\{ \begin{array}{l} F_{i+1,j}, \\ \min\limits_{i<k\leq j} C_{i,k} + F_{k+1,j} \end{array} \right\}$$

multi-loop decomp. in left & right part.



$$C_{i,j} = \min \left\{ \begin{array}{l} \mathcal{H}(i,j), \\ \min\limits_{i<k<l<j} C_{k,l} + \mathcal{I}(i,j;k,l), \\ \min\limits_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a \end{array} \right\}$$

contains exactly 1 bp (u+1, k)



$$M_{i,j} = \min \left\{ \begin{array}{l} \min\limits_{i<u<j} (u-i+1)c + C_{u+1,j} + b, \\ \min\limits_{i<u<j} M_{i,u} + C_{u+1,j} + b, \\ M_{i,j-1} + c \end{array} \right\}$$

$$M^1_{i,j} = \min \left\{ \begin{array}{l} M^1_{i,j-1} + c, \\ C_{i,j} + b \end{array} \right\}$$

penalty as initialization terms (init loop, unpaired in loop)

First proposed by Zuker et al.

init: $F_{i,j} = 0$, $C_{i,j} = M_{i,j} = M^1_{i,j} = \infty$.

- $F_{1,n}$ stores the energy value of the thermodynamically most stable structure, its Minimum Free Energy (MFE).
- traceback structure

- $F_{i,j}$: free energy of the opt. sub-struct. on the sub-seq. $s_i \ldots s_j$.
- $C_{i,j}$: free energy of the opt. sub-struct. on the sub-seq. $s_i \ldots s_j$ given that i and j form a base pair.
- $M_{i,j}$: free energy of the opt. sub-struct. on the sub-seq. $s_i \ldots s_j$ given that $s_i \ldots s_j$ is part of a multi-loop and has at least one "component".
- $M^1_{i,j}$: free energy of the opt. sub-struct. on the sub-seq. $s_i \ldots s_j$ given that $s_i \ldots s_j$ is part of a multi-loop and has exactly one component which has the closing pair (i, h) for some h satisfying $i < h \leq j$.

RNAfold - webserver:

```
http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi

 RNAfold < trna.fa
>AF041468
GGGGGUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGGUUCGAGUCCCCUUACCUCCA
(((((((..((((........)))).((((.......)))))).....((((.......))))))))))).
-31.10 kcal/mol
```