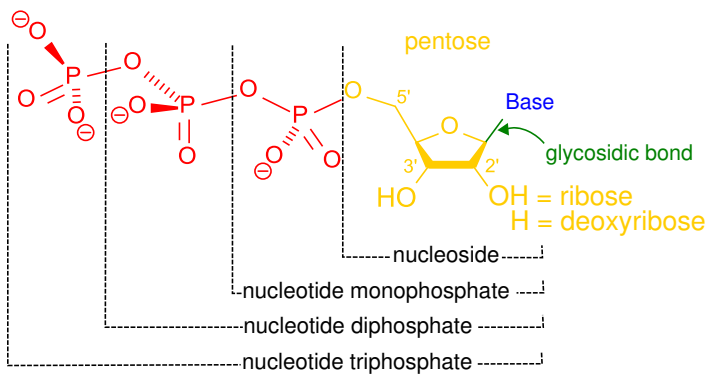
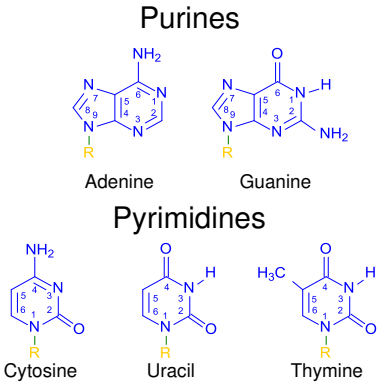
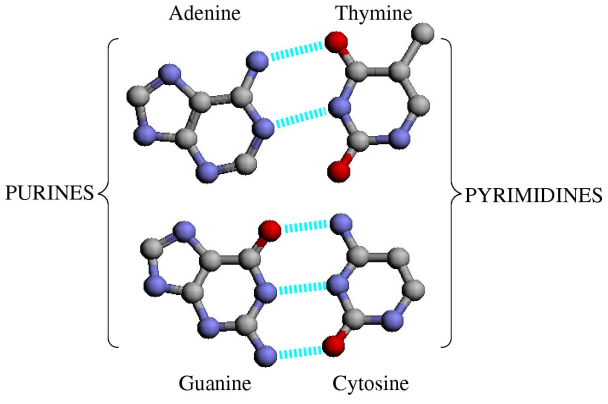
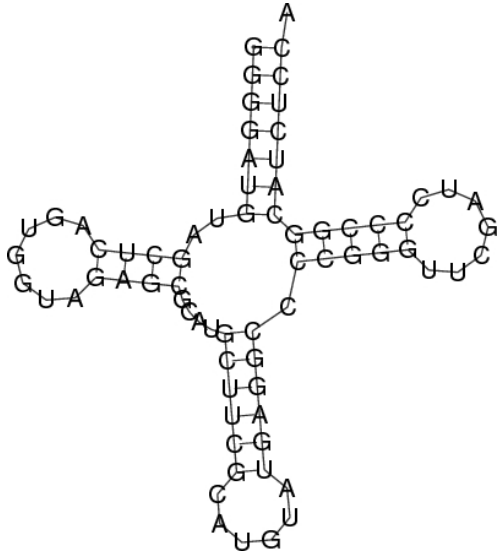
RNA Structure and RNA Structure Prediction









Definitions

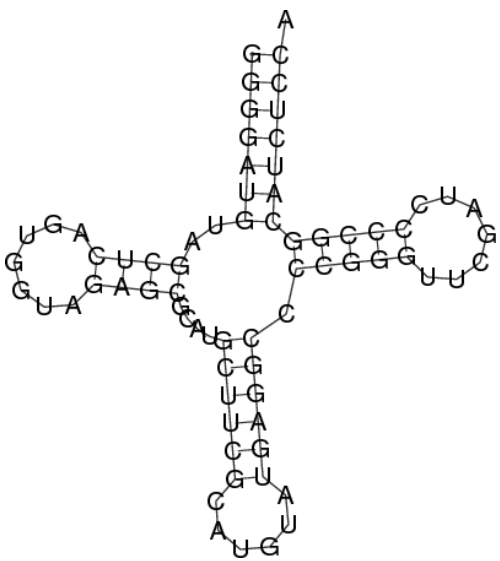
Definition (RNA Structure)

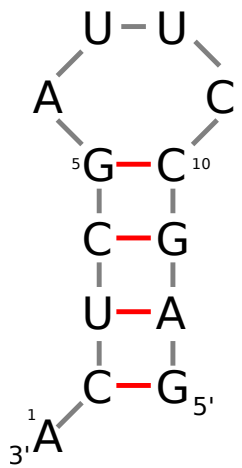
Let $S \in \{A, C, G, U\}^*$ be an *RNA sequence* of length $n = |S|$. An *RNA structure of S* is a set of *base pairs*

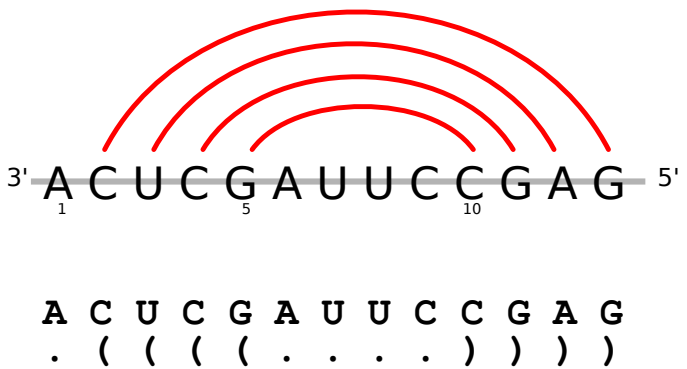
$$P \subseteq \{(i, j) \mid 1 \leq i < j \leq n, S_i \text{ and } S_j \text{ complementary}\}$$

such that the degree of P is at most one, i.e.

$$\text{for all } (i, j), (i', j') \in P : (i = i' \Leftrightarrow j = j') \text{ and } i \neq j'.$$







$$P = \{(2, 13), (3, 12), (4, 11), (5, 10)\}$$

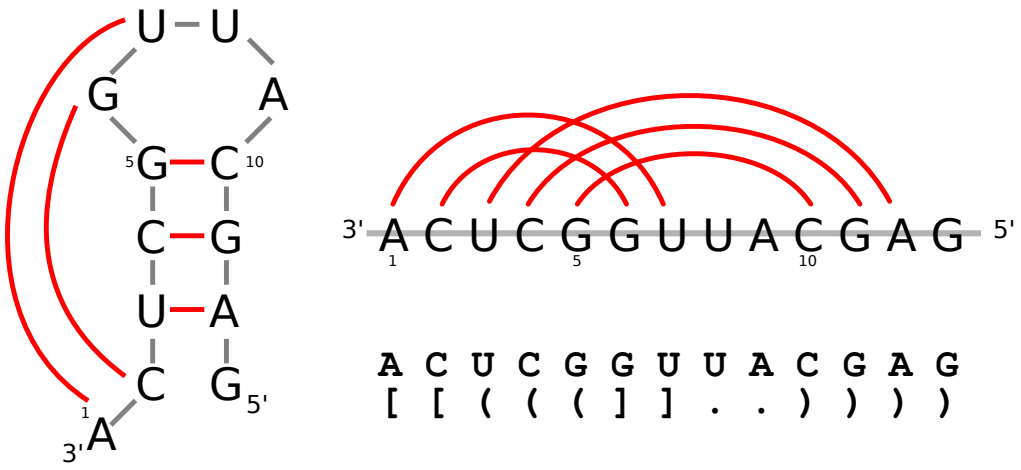
Definitions II

Definition (Crossing)

Two base pairs (i, j) and (i', j') are *crossing* iff

$$i < i' < j < j' \qquad \text{or} \qquad i' < i < j' < j.$$

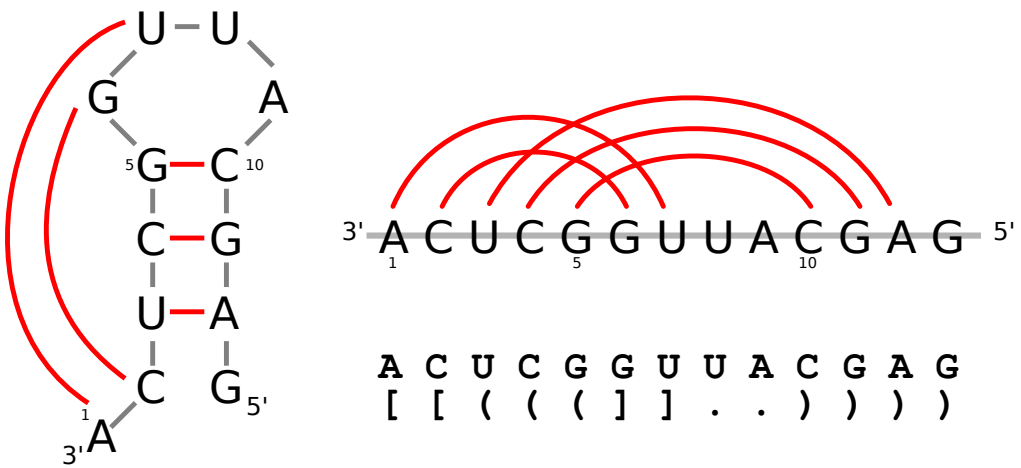
An RNA structure P (of an arbitary RNA sequence S) is *crossing* iff P contains (at least) two crossing base pairs. Otherwise, P is called *non-crossing* or *nested*.



$$P = \{(1,7),(2,6),(3,12),(4,11),(5,10)\}$$

Remarks

- Synonyms: $(i, j) \in P$ is a "base pair", "bond", "arc"
- Usually, assume minimal allowed size of base pair (aka loop length) m. Then: additional constraint $j - i > m$ in def of RNA structure.
- Crossing base pairs form "pseudoknots" — crossing structures contain pseudoknots. The terms *pseudoknot-free* and non-crossing are synonymous for RNA structures.
- As defined "RNA structure" describes the secondary structure of an RNA. We will look at tertiary structure only later.



$$P = \{(1,7),(2,6),(3,12),(4,11),(5,10)\}$$

Prediction of RNA (Secondary) Structure

Definition (Problem of RNA non-crossing Secondary Structure Prediction by Base Pair Maximization)

IN: RNA sequence S

OUT: a non-crossing RNA structure P of S that maximizes $|P|$ (i.e. the number of base pairs in P).

Remarks:

- By dropping the non-crossing condition, we can define the general base pair maximization problem. The general problem can be solved by maximum matching.
- Maximizing base pairs for non-crossing structures will help to understand the more realistic case of minimizing energy. For ernergy minimization, predicting general structures is NP-hard.
- RNA structure prediction is often (less precisely) called *RNA folding*.

Nussinov Algorithm — Matrix definition

Let S be and RNA sequence of length n.
The Nussinov Algorithm solves the problem of RNA non-crossing secondary structure prediction by base pair maximization with input S.

Definition (Nussinov Matrix)

The *Nussinov matrix* $N = (N_{ij})_{\substack{1 \le i \le n \\ i-1 \le j \le n}}$ *of* S is defined by

$$N_{ij} := \max\{|P| \mid P \text{ is non-crossing RNA } ij\text{-substructure of } S\}$$

where we use:

Definition (RNA Substructure)

An RNA structure P of S is called *ij-substructure of* S iff $P \subseteq \{i, \ldots, j\}^2$.

Nussinov Algorithm — Recursive computation of $N_{i,j}$

Init: (for $1 \leq i \leq n$)

$$N_{ii} = 0 \text{ and } N_{ii-1} = 0$$

Recursion: (for $1 \leq i < j \leq n$)

$$N_{ij} = \max \begin{cases} N_{ij-1} \\ \max\limits_{S_k, S_j \text{complementary}} \; \max\limits_{i \leq k < j} \; N_{ik-1} + N_{k+1j-1} + 1 \end{cases}$$

Remarks:

- case 2 of recursion covers base pair (i, j) for $k = i$; then: N_{ik-1} (initialized with 0!) is max. number of base pairs in empty sequence.
- solution is in $N_{1,n}$
- Recursion furnishs a DP-Algorithm for computing the Nussinov matrix (including $N_{1,n}$) in $O(n^3)$ time and $O(n^2)$ space.
- How to guarantee minimal loop length?
- What happens without restriction non-crossing?
- Are there other decompositions?

Nussinov Algorithm — Example

	1	2	3	4	5	6	7	8		
	G	C	A	C	G	A	C	G		
0	0								G	1
	0	0							C	2
		0	0						A	3
			0	0					C	4
				0	0				G	5
					0	0			A	6
						0	0		C	7
							0	0	G	8

Note: example with minimal loop length 0.

Nussinov Algorithm — Example

	1	2	3	4	5	6	7	8		
	G	C	A	C	G	A	C	G		
0	0	1	1	1	2	2	2	3	G	1
	0	0	0	0	1	1	1	2	C	2
		0	0	0	1	1	1	2	A	3
			0	0	1	1	1	2	C	4
				0	0	0	1	1	G	5
					0	0	0	1	A	6
						0	0	1	C	7
							0	0	G	8

Note: example with minimal loop length 0.

Nussinov Algorithm — Traceback

Determine one non-crossing RNA structure P with maximal $|P|$.

pre: Nussinov matrix N of S:

		1	2	3	4	5	6	7	8		
		G	C	A	C	G	A	C	G		
0	0	1	1	1	2	2	2	3		G	1
	0	0	0	0	1	1	1	2		C	2
		0	0	0	1	1	1	2		A	3
			0	0	1	1	1	2		C	4
				0	0	0	1	1		G	5
					0	0	0	1		A	6
						0	0	1		C	7
							0	0		G	8

Idea:

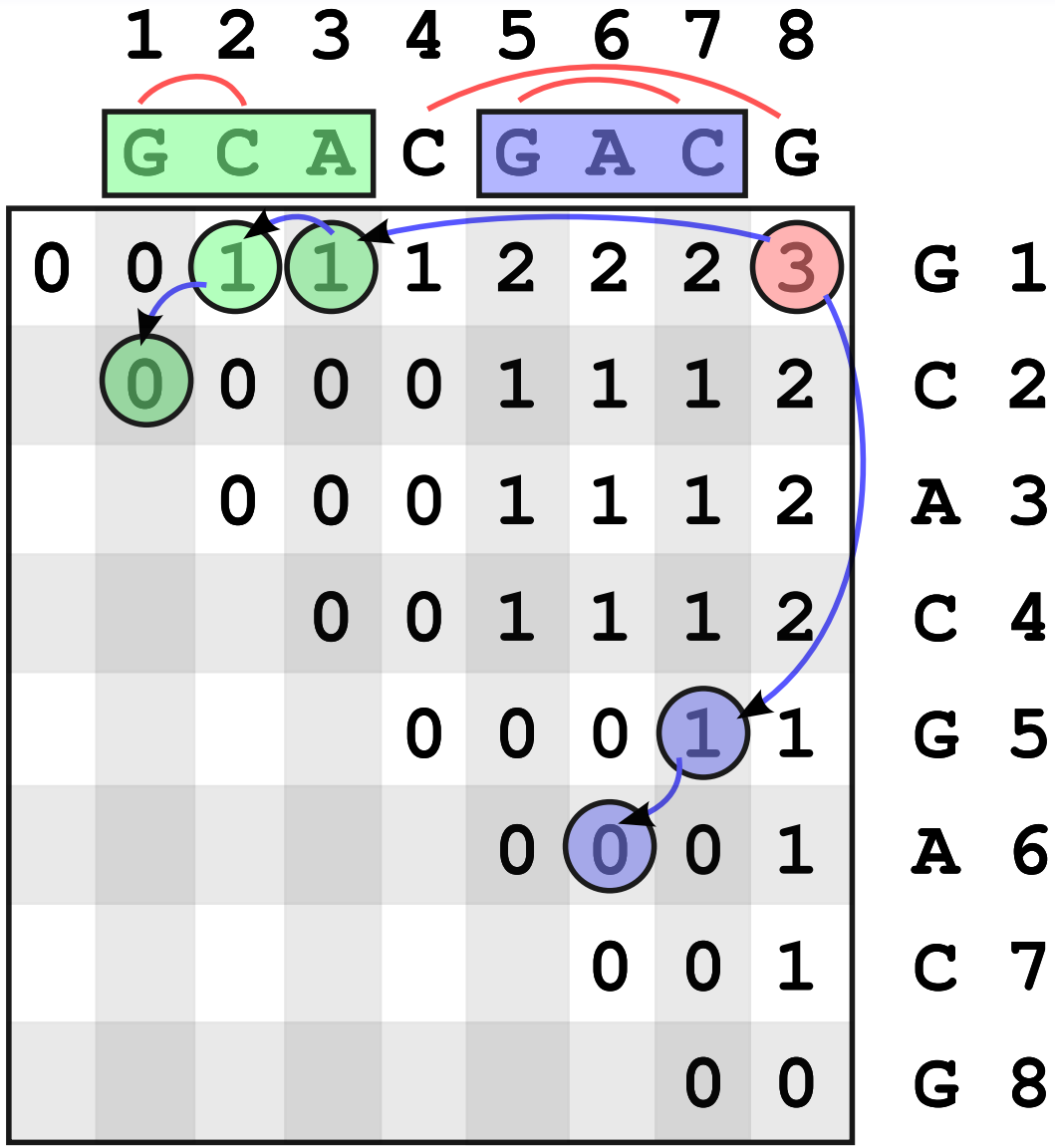
- start with entry at upper right corner N_{1n}
- determine recursion case (and the entries in N) that yield maximum for this entry
- trace back the entries where we recursed to

Nussinov Algorithm — Traceback Example

		1	2	3	4	5	6	7	8		
		G	C	A	C	G	A	C	G		
0	0	1	1	1	2	2	2	3		G	1
	0	0	0	0	1	1	1	2		C	2
		0	0	0	1	1	1	2		A	3
			0	0	1	1	1	2		C	4
				0	0	0	1	1		G	5
					0	0	0	1		A	6
						0	0	1		C	7
							0	0		G	8

Recall: example with minimal loop length 0 and without G-U pairing.

Nussinov Algorithm — Traceback Example



Recall: example with minimal loop length 0 and without G-U pairing.

Nussinov Algorithm — Traceback Pseudo-Code

CALL: traceback$(1, n)$

Procedure traceback(i, j)

if $j \leq i$ **then**
return
else if $N_{ij} = N_{ij-1}$ **then**
traceback$(i, j-1)$;
return
else
for all $k : i \leq k < j$, S_k and S_j complementary **do**
if $N_{ij} = N_{i\,k-1} + N_{k+1\,j-1} + 1$ **then**
print (k,j);
traceback$(i, k-1)$; traceback$(k+1, j-1)$;
return
end if
end for
end if

Remarks

- Complexity of trace-back $O(n^2)$ time
- How to get all optimal non-crossing structures?
- How to trace-back non-recursively?
- How to output / represent structures?
  - Dot-bracket
  - 2D-layout
  - Tree-like

Limitations of the Nussinov Algorithm

- Base pair maximization does not yield biologically relevant structures:
  - no stacking of base pairs considered
  - loop sizes not distinguished
  - no special scoring of multi-loops
- only one structure predicted
  - base pair maximization can not differnciate structures sufficiently well: possibly many optima
  - no sub-optimal solutions
- crossing structures cannot be predicted

However:

- shows pattern of RNA structure prediction by DP (simple+instructive)
- energy minimization (Zuker) will have similar algorithmic structure
- "only one solution"-problem can be overcome (suboptimal: Wuchty)
- prediction of (restricted) crossing structure can be seen as extension

Limitations of the Nussinov Algorithm

- Base pair maximization does not yield biologically relevant structures:
  - no stacking of base pairs considered
  - loop sizes not distinguished
  - no special scoring of multi-loops
- only one structure predicted
  - base pair maximization can not differnciate structures sufficiently well: possibly many optima
  - no sub-optimal solutions
- crossing structures cannot be predicted

However:

- shows pattern of RNA structure prediction by DP (simple+instructive)
- energy minimization (Zuker) will have similar algorithmic structure
- "only one solution"-problem can be overcome (suboptimal: Wuchty)
- prediction of (restricted) crossing structure can be seen as extension