Nussinov: Problem Definition

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 $|P|$ (= number of base pairs in P).

Remarks:

- We defined two variants of the problem. One with the addiditional requirement that structures are non-crossing and one without. Without this restriction the problem is NP-hard (at least for interesting scoring schemes) — with the restriction there will be an efficient algorithm for solving the problem.
- Maximizing base pairs will help to understand the more realistic case of minimizing energy.
- RNA structure prediction is often called *RNA folding* also we do not model the folding process but only predict the result.

Nussinov Algorithm — Matrix definition

Let S be an 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

Nussinov algorithm: determines the maximal number of bonds that a structure P for a sequence S can have

Definition (Nussinov matrix)

Let S be an RNA sequence.
The Nussinov matrix $(N_{i,j})_{\substack{1 \leq i \leq |S| \\ i-1 \leq j \leq |S| \wedge j > 1}}$ is defined by

$$N_{i,j} = \max \left\{ |P| \mid \begin{array}{l} \text{P is a nested structure} \\ \text{of the subsequence } S_i \dots S_j \end{array} \right\}$$

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 $1 \leq i < j \leq n$)

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

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.
- $S_{1,n}$ is the maximal $|P|$ of any P of S.
- Recursion furnishs a DP-Algorithm for computing the Nussinov matrix (including $S_{1,n}$) in $O(n^3)$ time and $O(n^2)$ space.
- How to restrict loop length?
- What happens without restriction non-crossing?

Nussinov Algorithm — Example

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

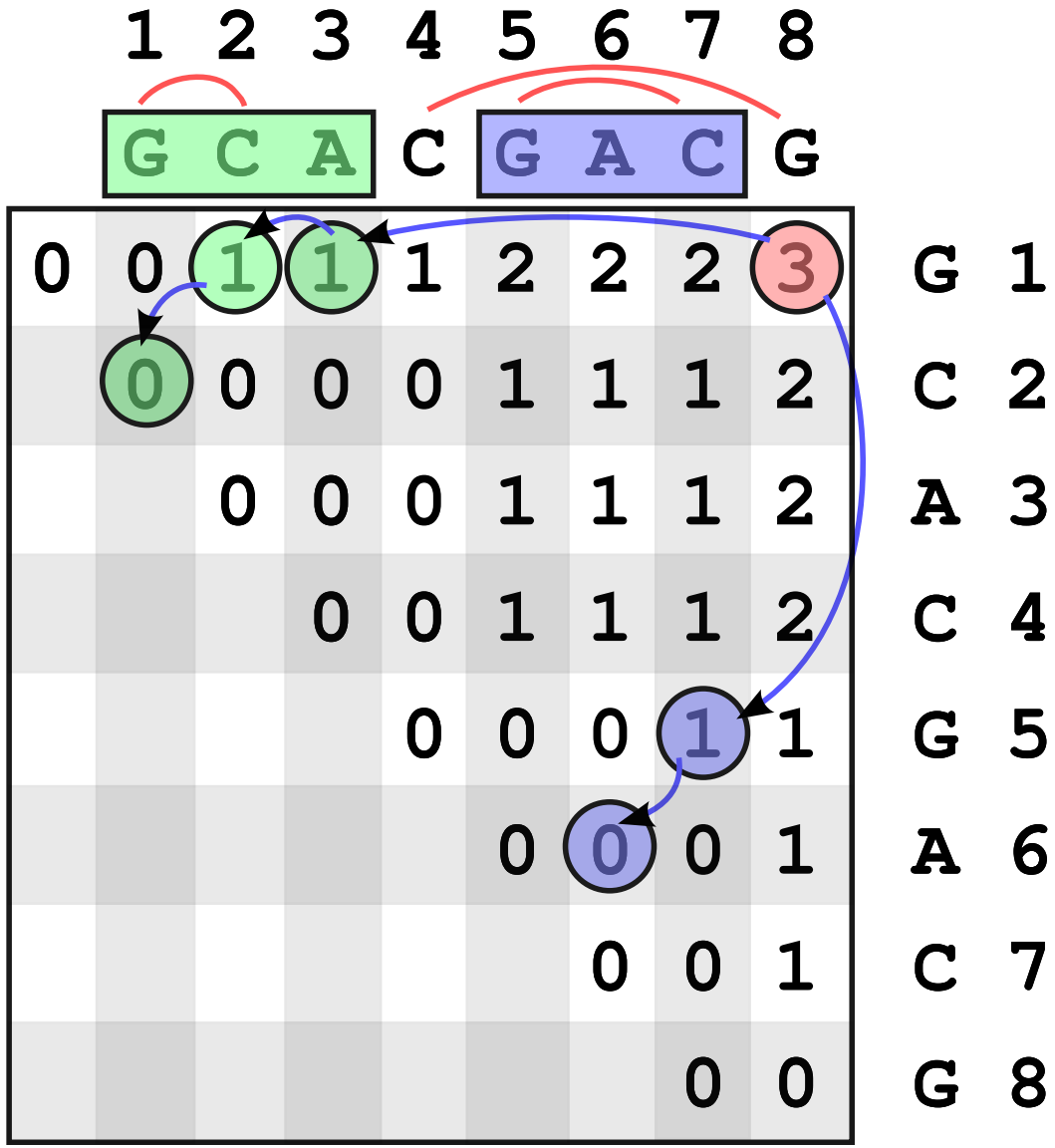
- start with entry at upper right corner N_{1n}
- determine recursion case (and the entries in N) that yield maximum for this entry
- trace back the entries where we recursed to

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

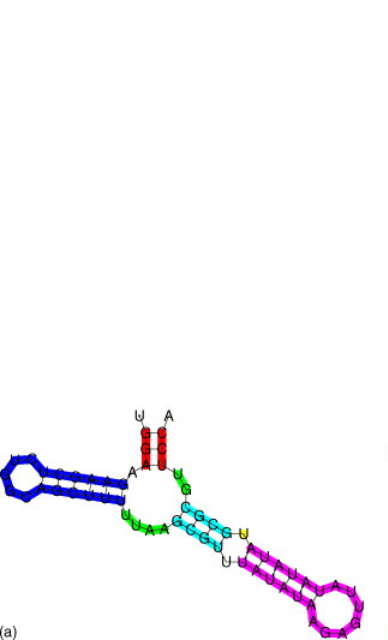
CALL: traceback$(1, n)$

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 nc structures?
- How to trace-back non-recursively?
- How to output / represent structures?
  - Dot-bracket
  - 2D-layout
  - Tree-like
- Why doesn't it work for crossing structure?

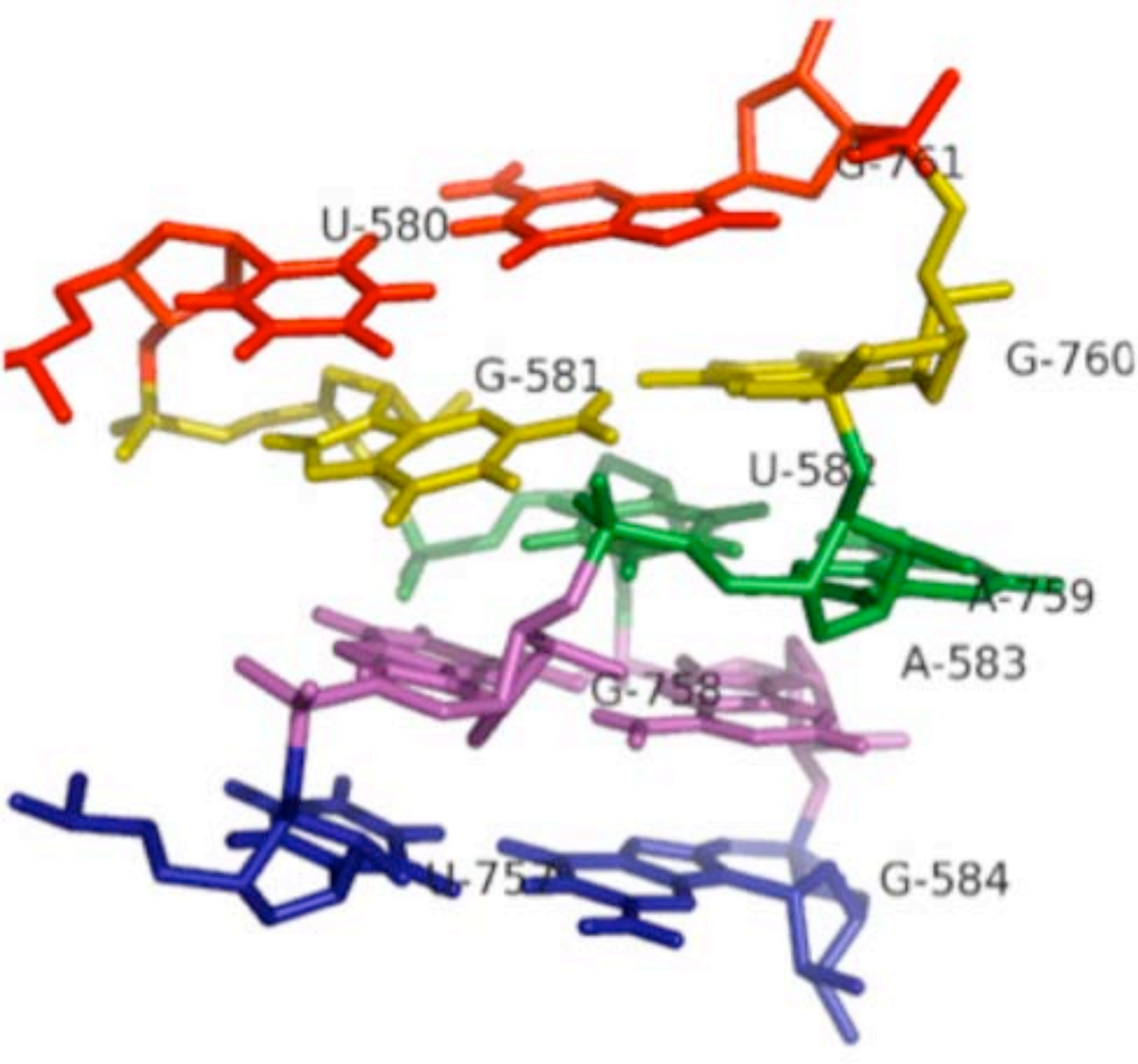

(a)
(b)

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



16S *T. thermophilus*

5‘ 3‘

580 U – G 761

581 G G 760

582 U A 759

583 A G 758

584 G – U 757

3‘ 5‘

Disadvantages of Nussinov

Nussinov doesn't determine biological relevant structures since:

- there are several possibilities to form base pairs, where Nussinov finds only one:

  A—U versus G—C

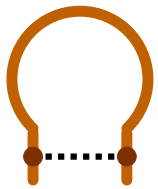
- Stacking of base pairs not considered $\Rightarrow$ difference in structure and stability of helices

  G—C and G—C
  C—G G—C

- size of intern loops not considered

instable

stable



instable