RNA Secondary Structure Prediction

Carlos Cotta

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga

http://www.lcc.uma.es/∼ccottap

Comput Eng, Softw Eng, Comput Sci & Math – 2021-2022

UNIVERSIDAD
DE MÁLAGA

Index

Biological Context

RNA molecules play a key role in many biological processes (as catalysts, messengers, gene expression regulators, etc.).

The structure of the molecule is essential to fulfill its mission. To carry out a certain regulatory function, a stable structure is typically needed.

The structure of the molecule arises from physical and chemical processes. Predicting this structure is important to understand the biological processes in which it is involved.

RNA Structure

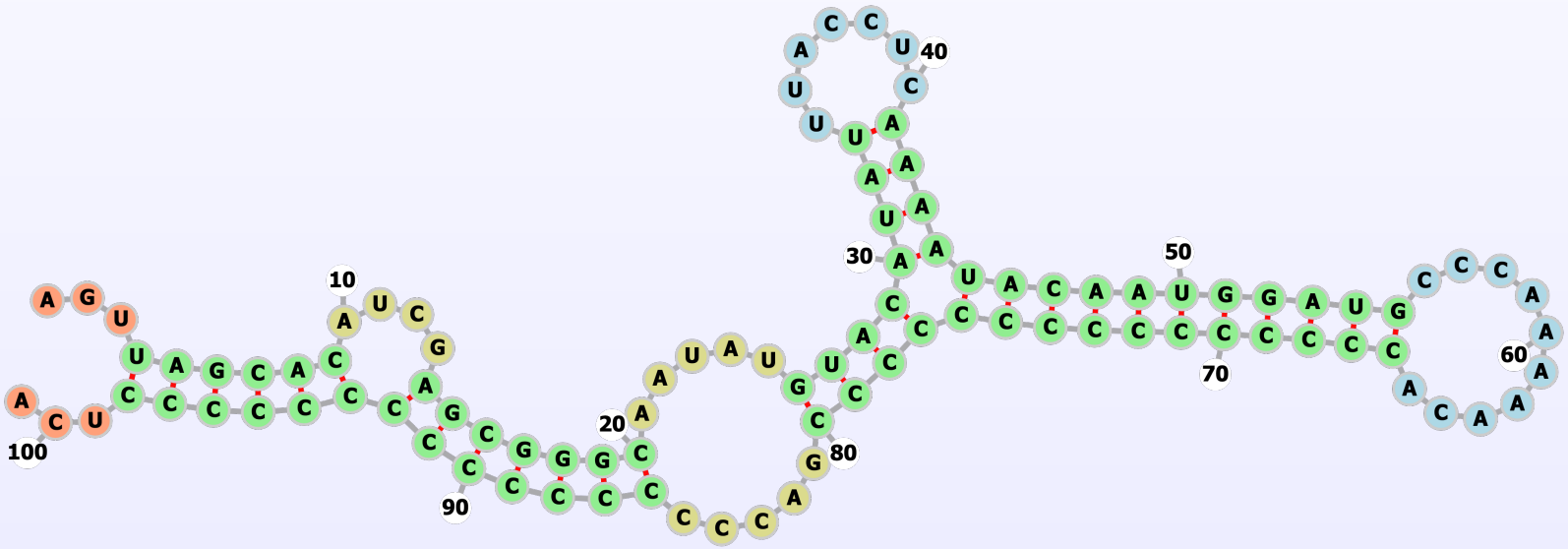
An RNA molecule is formed by a sequence of nucleotides: adenine (A), guanine (G), cytosine (C), and uracil (U). Hydrogen bonds between consecutive nucleotides constitutes the primary structure.

Pairing between non-adjacent bases gives rise to structural elements (bulges, stems, internal loops, hairpin loops, ...). These base pairings constitute the secondary structure.

The secondary structure provides the scaffolding for the 3D arrangement of the nucleotides in the molecule. This 3D folding produces the tertiary structure.

RNA Secondary Structure

We focus on predicting the secondary structure given the sequence of nucleotides in the RNA molecule (i.e., its primary structure).



The stability of the molecule is linked with the number of pairs in the secondary structure.

Problem Statement

RNA Secondary Structure Prediction

Given an RNA molecule $S = s_1 s_2 \dots s_n$ (where each nucleotide $s_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$, for $1 \leqslant i \leqslant n$), find a set of compatible base pairs $P \subset \{(i,j) \mid 1 \leqslant i < j \leqslant n\}$ of maximum cardinality, respecting any structural constraint that may exist.

Compatible base pairs

1. Watson-Crick pairs: A–U, C–G.
2. Wobble base pairs: G–U

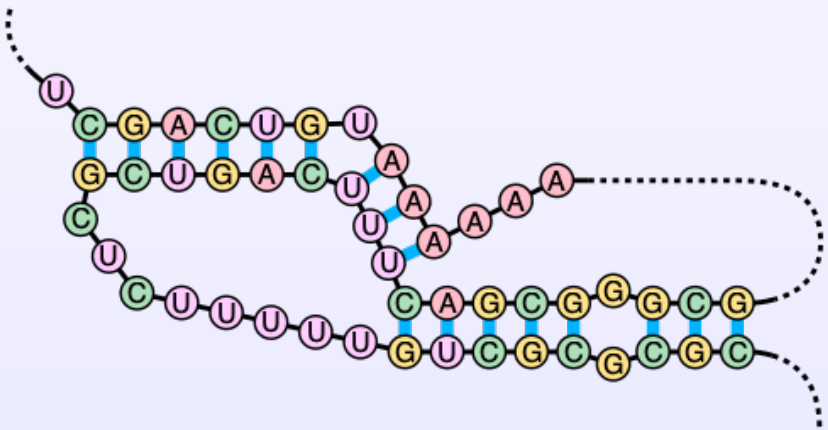
Any nucleotide in the sequence can only be paired once (if any).

Structural Constraints

No Pseudoknots

P is non-crossing. For any $(i,j),(i',j') \in P$, it cannot happen that

$$i < i' < j < j'$$



These structures are known as pseudoknots, and are rare.

If there are no crossings, the only possibilities are

$$i < j < i' < j'$$

or

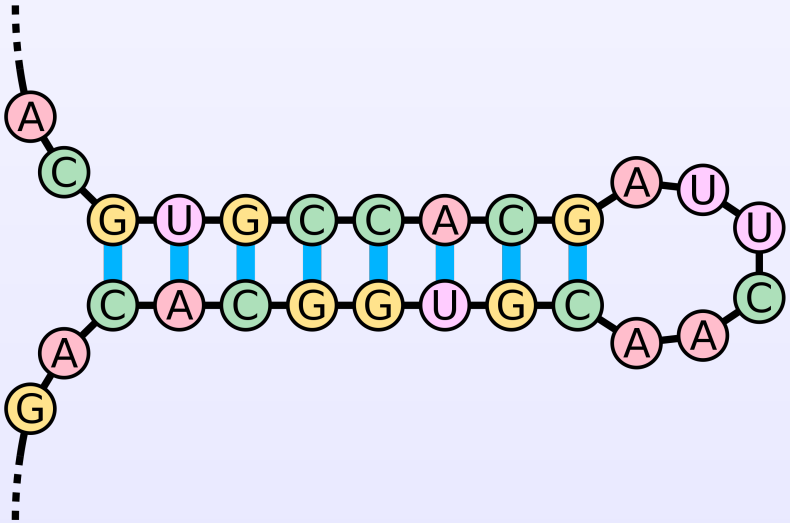
$$i < i' < j' < j,$$

i.e., paired segments are sequential or nested.

Structural Constraints

Loop size

For any $(i,j), \in P$, $j - i > L$, where L is the minimal stem-loop (also known as hairpin or hairpin-loop) size.



The length of the loop influences its stability. Chemically, it is not possible to have fewer than three bases in a loop. Typically, they have 4-8 bases.

Very long loops are also unstable, but we will not directly constraint the maximum size.

Nussinov Algorithm

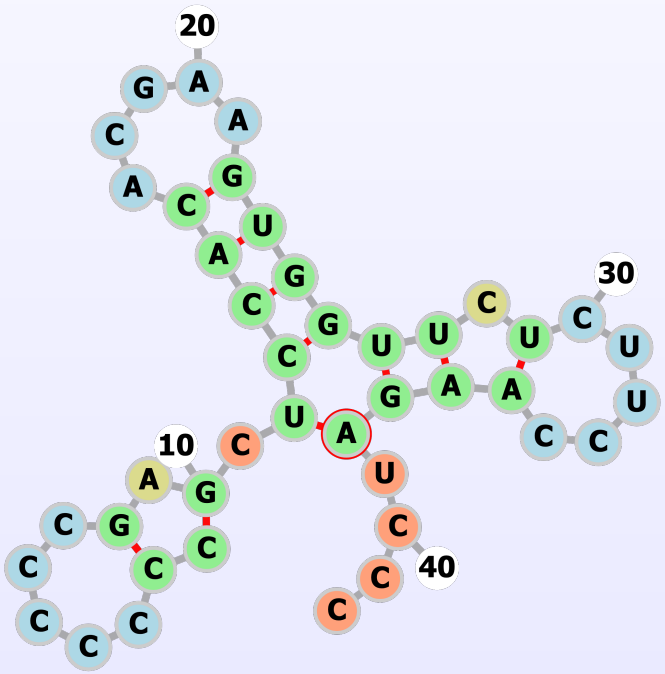
Solutions, decisions, and evaluation

1. A solution is a collection P of base pairs that fulfills the problem constraints.
2. We can construct this solution by deciding to which other nucleotide we are going to pair the nucleotide in a given position (including the possibility of not pairing it at all) .
3. The objective function (maximization) is $f(P) = |P|$.

Nussinov Algorithm

Optimal Substructure

Let P^* be the optimal solution, and let $(i, j) \in P^*$ be the pair with the largest value of j (in the example, this is the ($\mathtt{U}_{12}$, $\mathtt{A}_{38}$) pair).



We can observe that:

1. $\{(i', j') \in P^* \mid j' < i\}$ is the optimal pairing for $s_1 \dots s_{i-1}$.
2. $\{(i', j') \in P^* \mid i < i' < j' < j\}$ is the optimal pairing for $s_{i+1} \dots s_{j-1}$.
3. $\emptyset$ is the optimal pairing for $s_{j+1} \dots s_n$.

It is easy to prove these claims by contradiction.

Nussinov Algorithm

Subproblems and Bellman Equation

Subproblems considered

$N_{i,j}$ = maximum number of base pairs in $s_i \dots s_j$

To design the Bellman equation consider that:

- if $j - i \leqslant L$, no pairing is possible.
- if $j - i > L$, s_j can be paired with any compatible nucleotide in $s_i \dots s_{j-L-1}$, or it can be left unpaired.
  - if it is left unpaired, the problem reduces to finding the best pairing in $s_i \dots s_{j-1}$.
  - if it is paired with some s_k, we need to subsequently find the best pairing in $s_i \dots s_{k-1}$ and in $s_{k+1} \dots s_{j-1}$.

Nussinov Algorithm

Subproblems and Bellman Equation

The Bellman equation would thus be:

$$N_{i,j} = \begin{cases} 0 & j - i \leqslant L \\ \max\left(N_{i,j-1}, \max\limits_{\substack{i \leqslant k < j-L \\ \text{pair}(s_k, s_j)}} (N_{i,k-1} + N_{k+1,j-1} + 1) \right) & j - i > L \end{cases}$$

Nussinov Algorithm

Time and Space complexity

We need a table for $N_{i,j}$ (and another one for storing the optimal decisions associated to each subproblem), where $1 \leqslant i \leqslant n$ and $0 \leqslant j \leqslant n$. The space complexity is hence $\Theta(n^2)$.

Computing an entry in the table has complexity $\Theta(n)$. Therefore, the total time complexity is $O(n^3)$.

Nussinov Algorithm

Reconstruction of the Optimal Solution

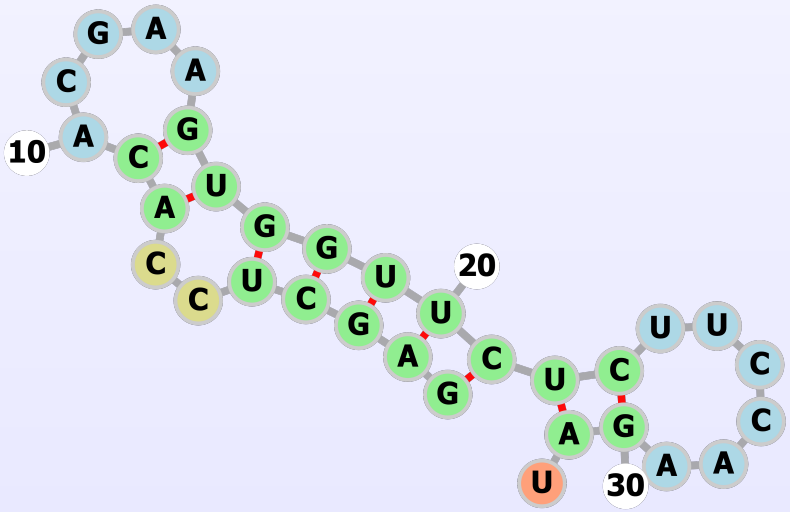
We have a table $D_{i,j}$ with the optimal decisions for each subproblem. We start at $D_{1,n}$:

1. If $D_{i,j} < 0$, the j-th nucleotide is unpaired. We go to $D_{i,j-1}$.
2. If $D_{i,j} = k > 0$, the pair (k, j) belongs to the solution, together with those we get from $D_{i,k-1}$ and those from $D_{k+1,j-1}$.
3. If $j - i \leqslant L$, there are no further pairs.

Dot-Bracket Notation

We will typically use dot-bracket notation to represent (non-crossing) solutions P. We use a string σ of length n in which:

- For any $(i, j) \in P$, $\sigma_i =$ '(' and $\sigma_j =$ ')'
- For any k not appearing in any pair, $\sigma_k =$ '.'




```
GAGCUCCACACGAAGUGGUUCUCUUCCAAGAU
(((((..((.....)))))))((......)).
```

Complementary Bibliography


P. Clote, R. Backofen,
Computational Molecular Biology: An Introduction,
John Wiley & Sons, Chichester, 2000


R. Nussinov, A.B. Jacobson,
"Fast algorithm for predicting the secondary structure of single-stranded RNA",
PNAS 77(11):6309–6313, 1980.

Image Credits

- RNA "Braveheart" molecule: Sanbonmatsu et al., *Nature Communications*, doi:10.1038/s41467-019-13942-4
- A pseudoknot structure in human telomerase RNA: Sakurambo - Own work; redraw of PMID 15849264, PMC 1149427 – CC BY 2.5
- An example of an RNA stem-loop: Sakurambo - Own work – CC BY 2.5