Algorithms and Complexity

Dynamic programming

Julián Mestre

School of Information Technologies
The University of Sydney

THE UNIVERSITY OF SYDNEY

RNA

RNA is a close relative of DNA but has a single strand of genetic code

Unlike DNA, which folds into a double helix shape, RNA folds onto itself

The 3D structure of RNA influences it behavior and molecular function

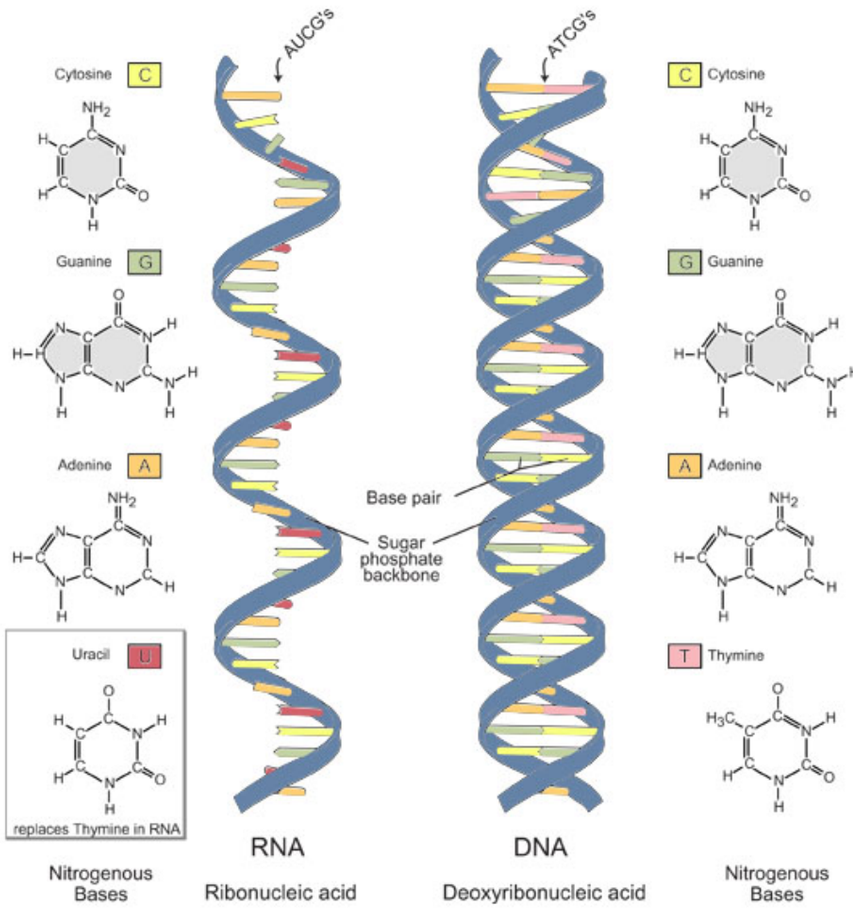
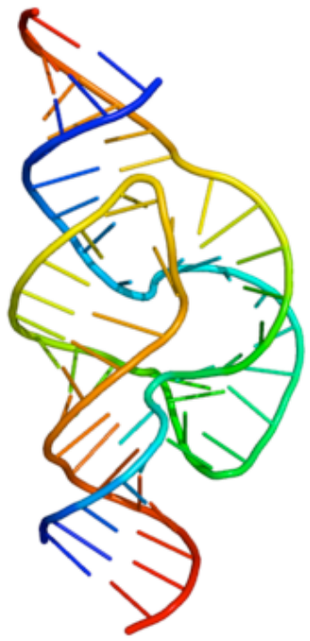


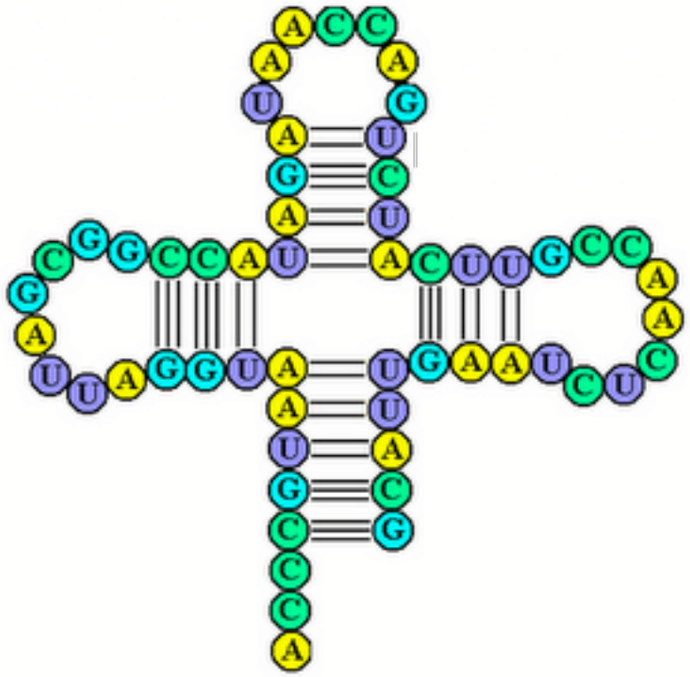
Image adapted from: National Human Genome Research Institute. Talking Glossary of Genetic Terms. Available at: www.genome.gov/Pages/Hyperion//DIR/VIP/Glossary/Illustration/rna.shtml.

RNA's secondary structure

A RNA molecule can be represented as a string over the alphabet {A, C, G, U}; each character is called a base.

Some bases pair up, causing the strand to fold onto itself. This pairing is called RNA secondary structure





Matching bases

Let S be the set of paired positions. If (i,j) in S then we say that position i is matched to position j

The pairing S must obey some properties

- *[Matching]* Each position is matched at most once
- *[Compatible matches]* Only matches A-U and C-G are allowed
- *[No sharp turns]* If (i,j) is a match then $|i-j| > 4$
- *[Non crossing]* If (i,j) and (k,l) are matched then intervals are either disjoint or nested in one another.

Nature prefers low energy configurations. In our case this means maximizing the number of matched positions

RNA secondary structure prediction

Input:

- String over the alphabet {A, C, G, U}

Task:

- Find a matching S obeying properties (i)-(iv) that maximizes |S|

Properties:

- *(i) [Matching]* Each position is matched at most once
- (ii) *[Compatible matches]* Only matches A-U and C-G are allowed
- (iii) *[No sharp turns]* If (i,j) is a match then $|i-j| > 4$
- (iv) *[Non crossing]* If (i,j) and (k,l) are matched then intervals are either disjoint or nested in one another.

DNA sequence alignment

A fundamental task in biology is to determine how closely related two organisms are. For this we compare their DNA sequences.

DNA sequence is string over the alphabet {A, G, C, T}

Similar sequences:

- AGCCTATGCAT
- AGCTTATAAGCAT

Dissimilar sequences

- AGCCTATGCAT
- GGTATGCAACT

DNA sequences drift with time

Over time, mutations introduce changes in the DNA sequence of a species

- A single position may change to a different letter
- A letter (or several) may be deleted
- A letter (or several) may be added

Each of these can happen with some low probability. Over evolutionary time, these low probability events accumulate.

The longer two species have diverged, the more differences we should see.

Sequence alignment

Let $X=x_1, x_2, ..., x_n$ and $Y= y_1, y_2, ..., y_m$ be two string.

An alignment is a matching M between positions of the two string such that there are no crossings:

if (i,j) in M and (i', j') in M then if $i < i'$ then $j < j'$

Associated with a matching there is a cost:

- for each (i,j) in M such that $x_i \neq y_j$ we pay a mismatch cost $\alpha(x_i,y_j)$
- for each position in X or Y left unmatched we pay a gap penalty δ

Traveling Salesman Problem

Input:

- G=(V,E) undirected graph
- $\ell : E \to R$ edge lengths
- X : set of k nodes to visit

Output:

- $x_1, x_2, \ldots, x_k$ a sequencing of X

Objective:

- minimize $dist(x_1, x_2) + dist(x_2, x_3) + \ldots + dist(x_{k-1}, x_k) + dist(x_k, x_1)$