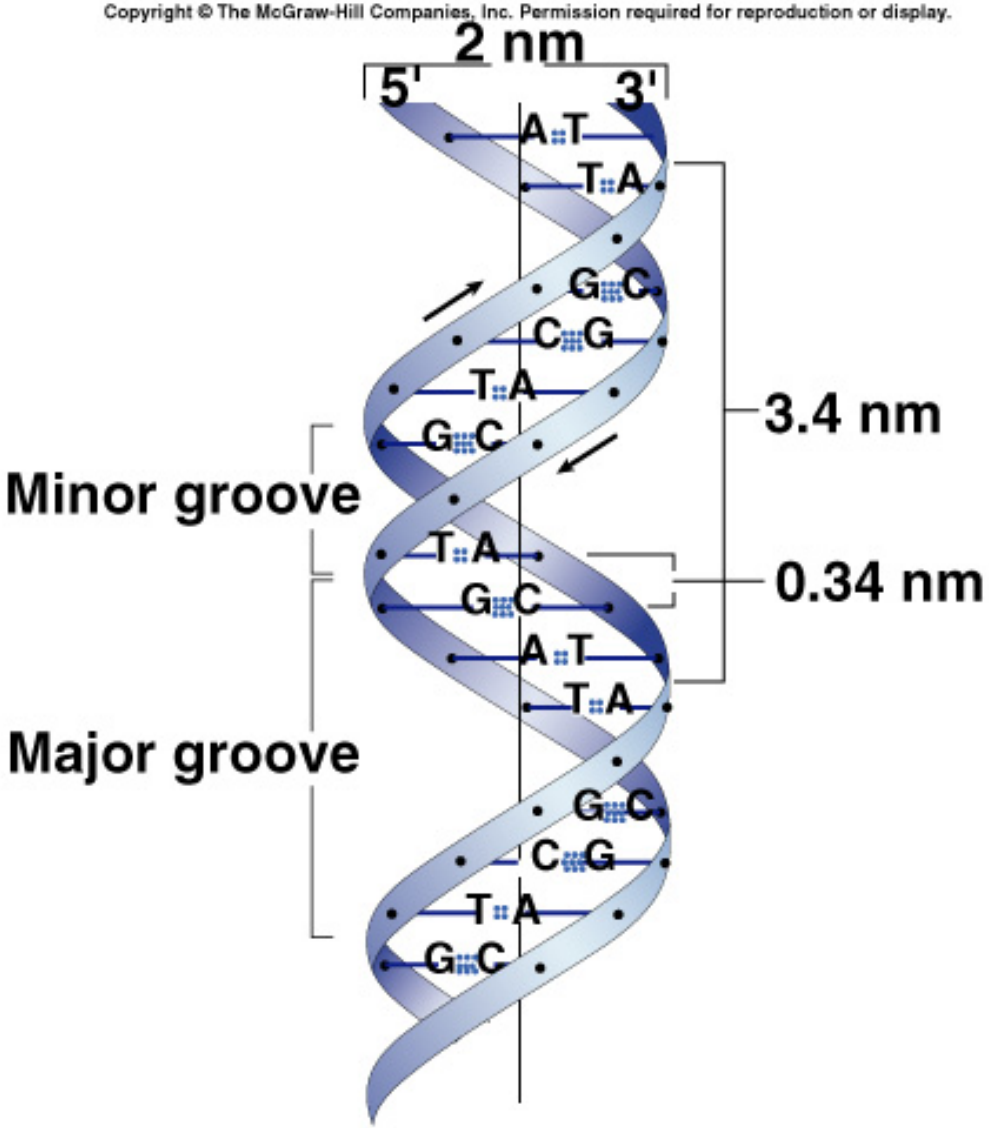
Exploring Life through Logic Programming

Alessandro Dal Palú Agostino Dovier
Andrea Formisano Enrico Pontelli

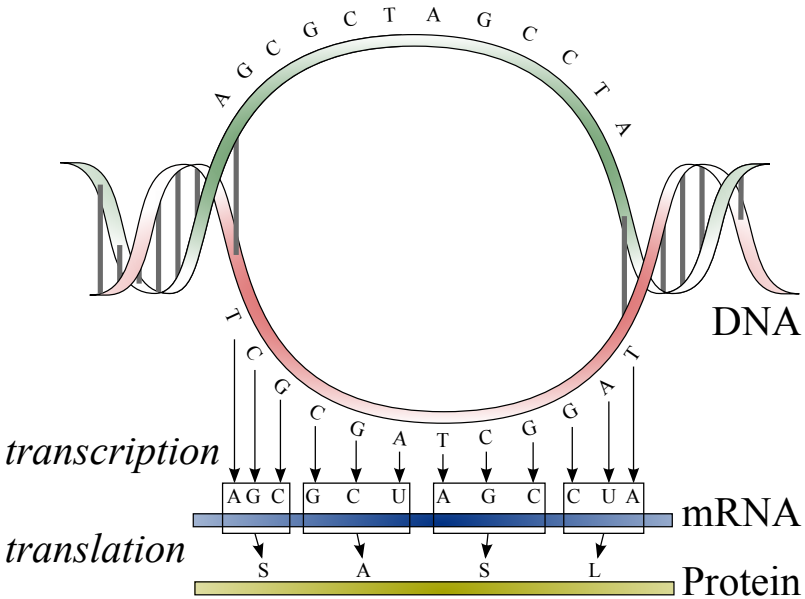
Dept. Computer Science, New Mexico State University, USA

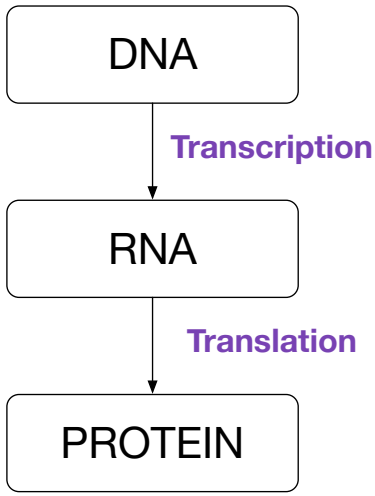
University of Udine, Italy
December 2015

RNA and Central Dogma



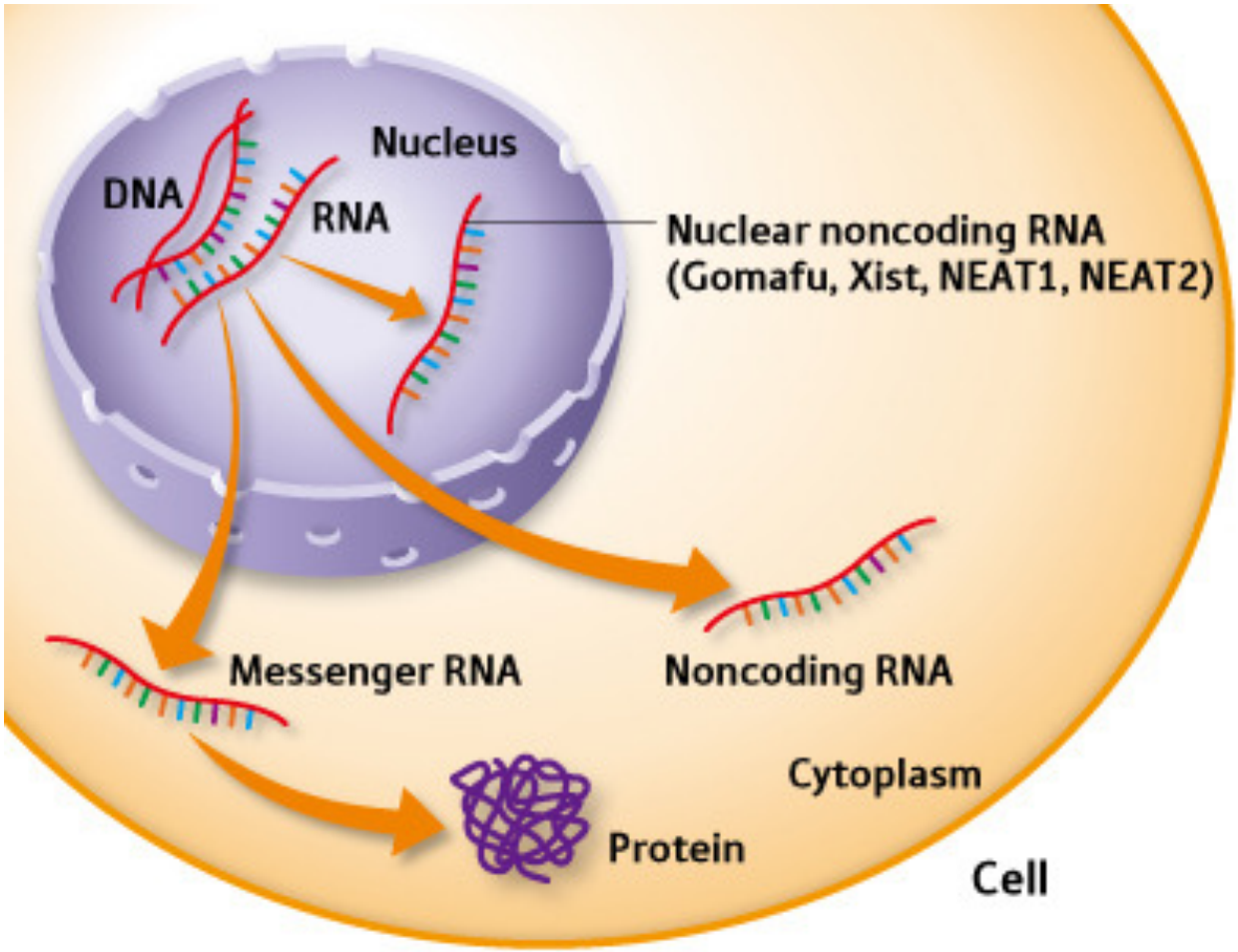
RNA and Central Dogma



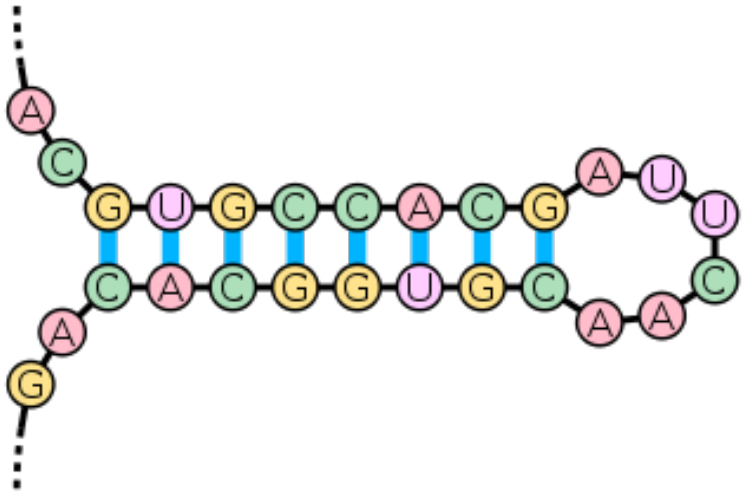


- RNA is a sequence of nucleotides (A,C,G,U) that (often) is just an intermediary between DNA and proteins
- DNA strands are transcribed to mRNA, in order to exit the cell's *nucleus*
- Nucleotides replacement: DNA T $\mapsto$ RNA U.

Central Dogma

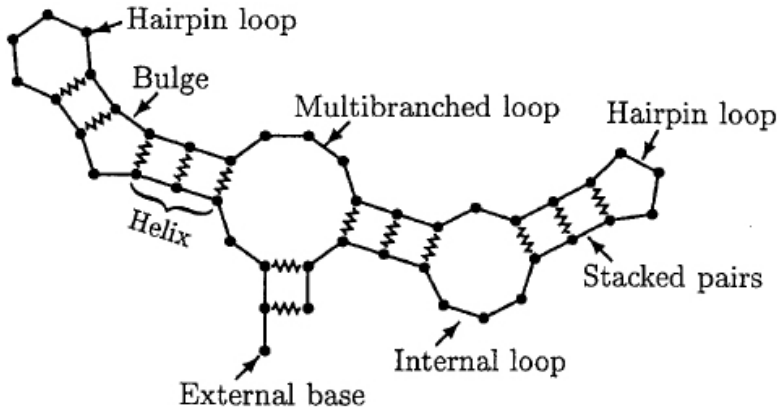


RNA Secondary Structure



- RNA folds according to favorable matchings (A-U, C-G, $\sim$ U-G)
- The secondary structure is the set of its base pairings
- Secondary structure determines the 3D properties

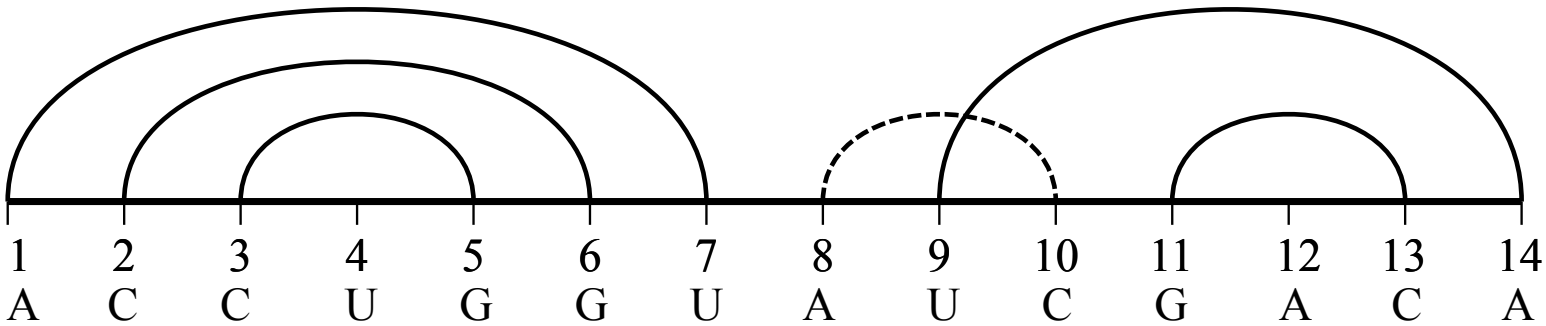
RNA Secondary Structure



- RNA folds according to favorable matchings (A-U, C-G, $\sim$ U-G)
- The secondary structure is the set of its base pairings
- Secondary structure determines the 3D properties

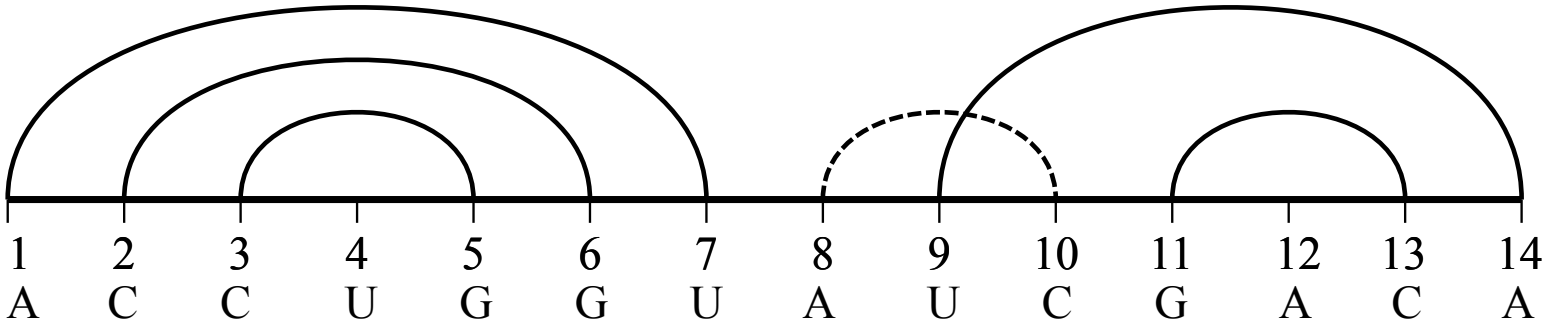
Mathematically

- A RNA sequence $\vec{s} = s_1 s_2 \cdots s_n$ is a string in $\{A, C, G, U\}^*$
- Structure described by set of pairs of interacting bases
- A RNA secondary structure is a (partial) injective function $P \subseteq \{1, \ldots, n\}^2$ such that
  - $(i, j) \in P \leftrightarrow (j, i) \in P$
  - $(i, j) \in P$ only if $(s_i, s_j) \in \{(A, U), (U, A), (C, G), (G, C), (U, G), (G, U)\}$

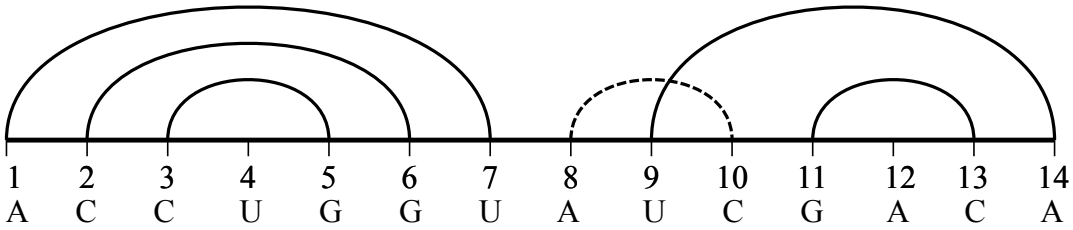


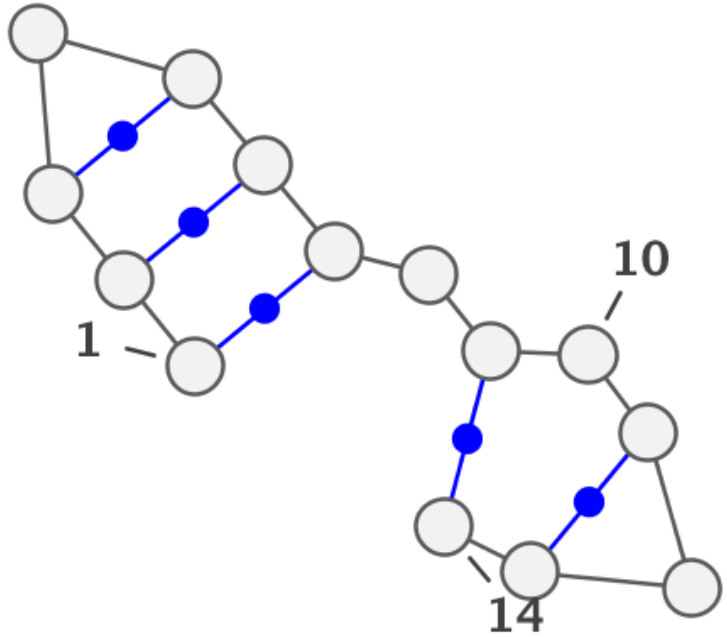
Mathematically

- A RNA sequence $\vec{s} = s_1 s_2 \cdots s_n$ is a string in $\{A, C, G, U\}^*$
- Structure described by set of pairs of interacting bases
- A RNA secondary structure is a (partial) injective function $P \subseteq \{1, \ldots, n\}^2$ such that
  - $(i, j) \in P \leftrightarrow (j, i) \in P$
  - $(i, j) \in P$ only if $(s_i, s_j) \in \{(A, U), (U, A), (C, G), (G, C), (U, G), (G, U)\}$
- We are interested in a solution with maximal pairings (and/or minimizing a more complex energy function)



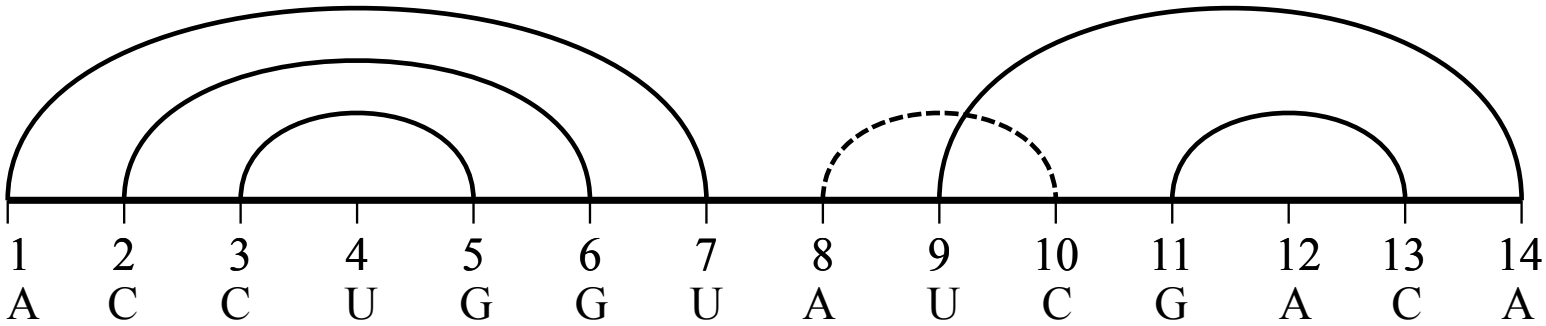
Hypothetical Arrangement





Complexity

- The general problem is NP-complete [Lyngsø and Pedersen 2000].
- A large sub-class has *polynomial time* complexity:
  - the absence of pseudo-knots.



Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop

CGUUGUGUACACGAUAGUACAU

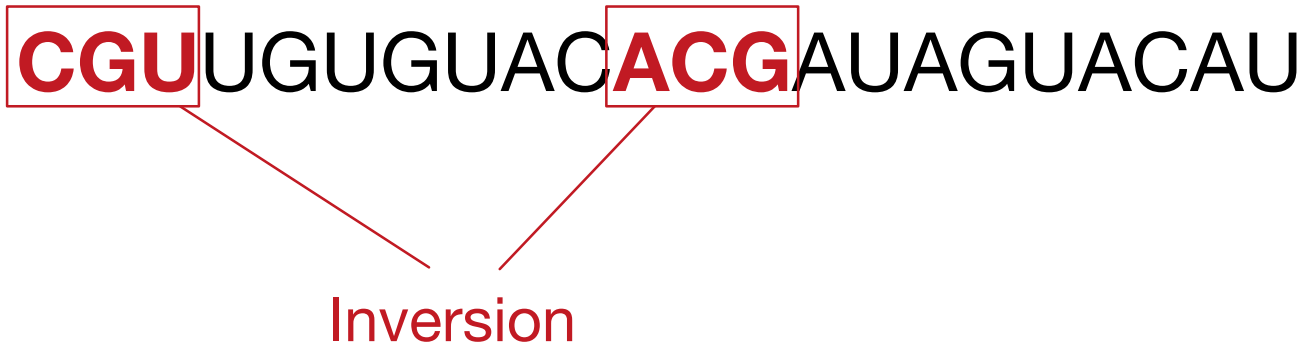
Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop

CGUUGUGUAC**ACG**AUAGUACAU

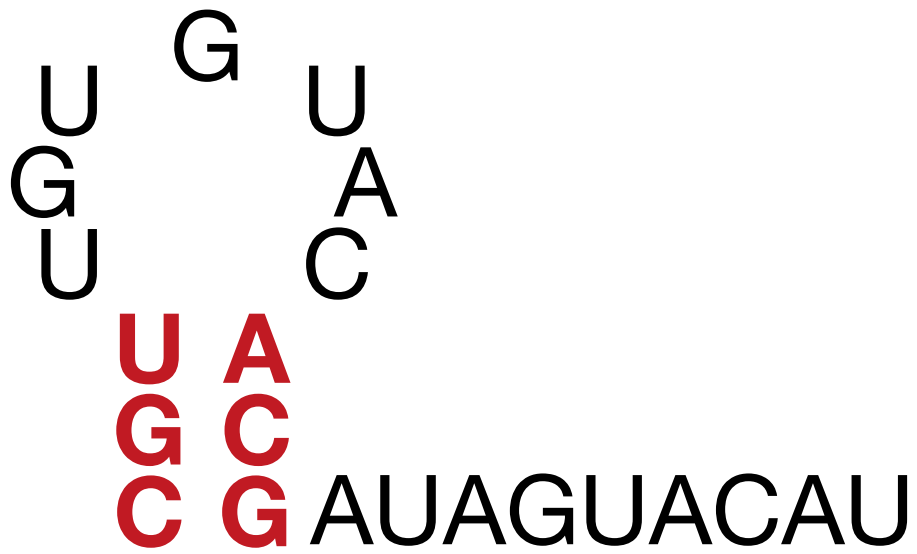
Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop



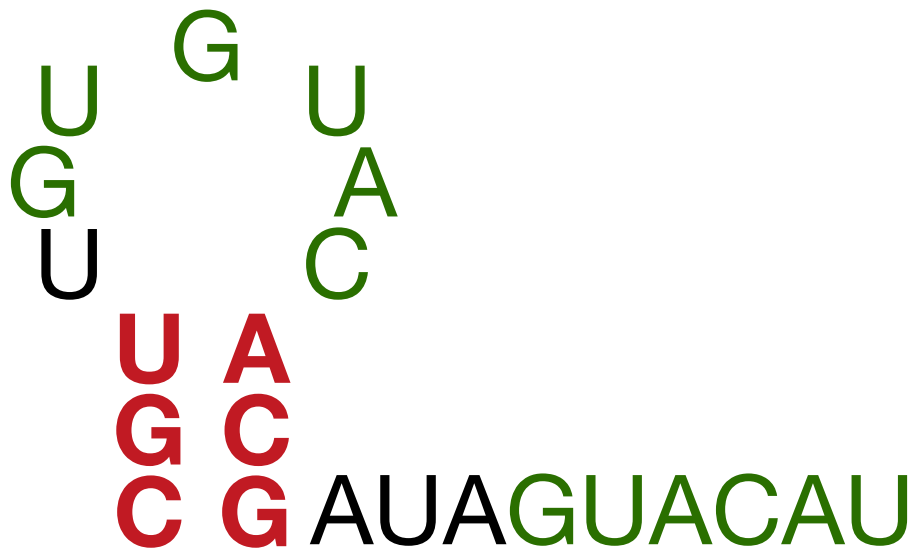
Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop



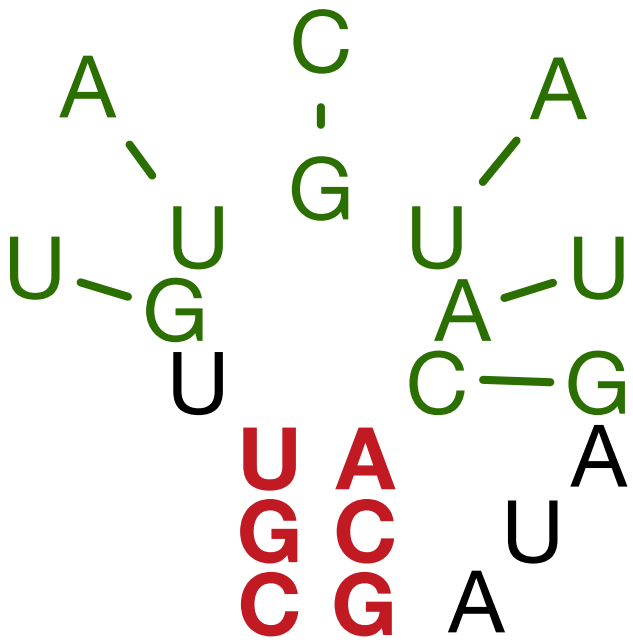
Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop



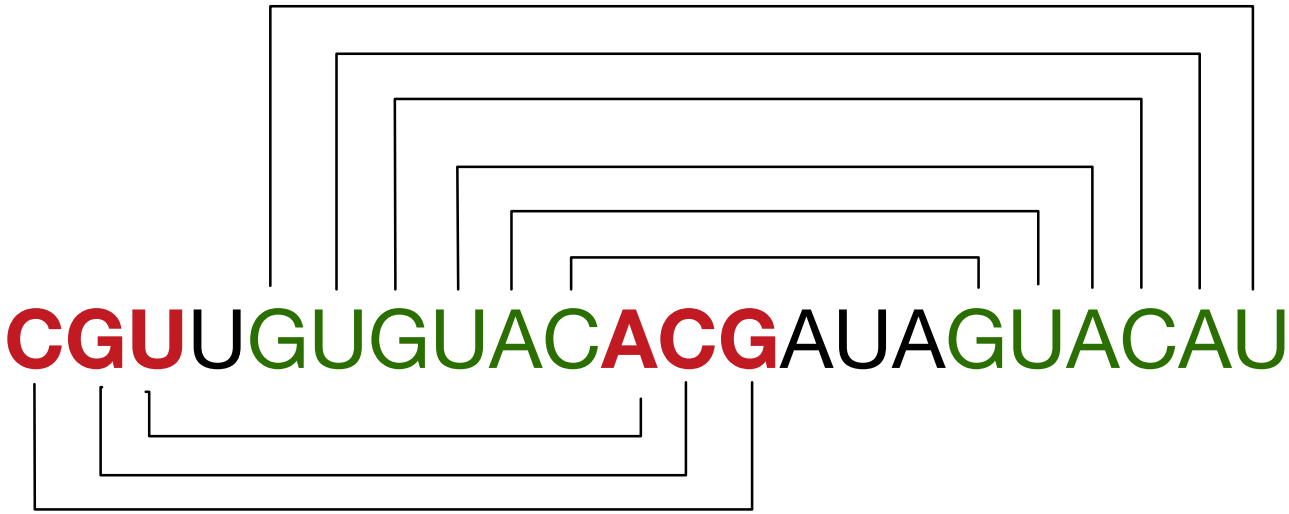
Pseudo-knots

- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop

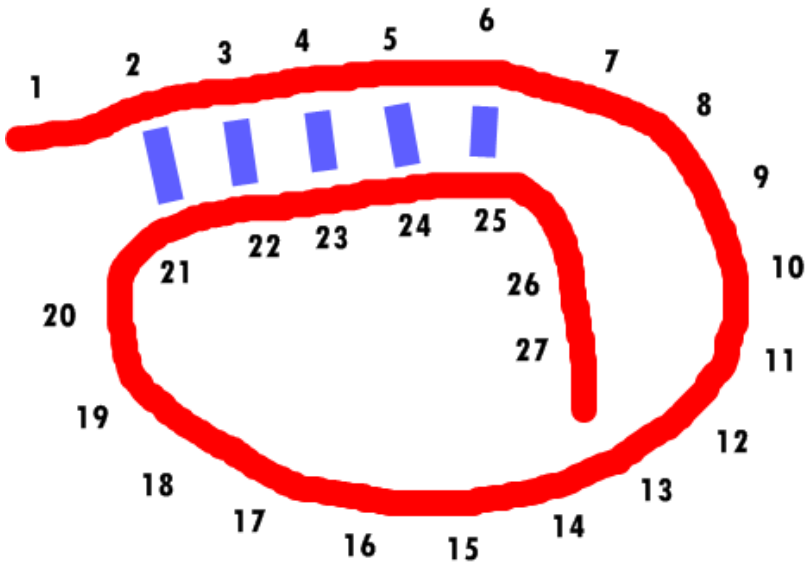


Pseudo-knots

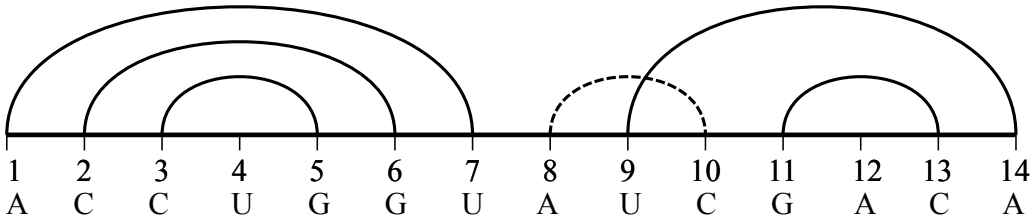
- Pseudo-knot: secondary structure where a loop is paired with a region outside of the stem flanking the loop



Pseudo-knots



To avoid pseudo-knots, we impose a constraint:
If $i < \ell < j$ and $(i, j) \in P$, and $((\ell, k) \in P$ or $(k, \ell) \in P)$, then $i < k < j$.



ASP Encoding

RNA sequence

Each sequence is encoded with *n* facts, e.g.:

```
seq(1,a).  seq(2,g).  seq(3,u).
seq(4,c).  seq(5,c).  seq(6,a).
```

The main predicate is `pairing/2` which is a partial function.

ASP Encoding

Pairings

```
(1)   sequence_index(X)   :- seq(X,_).
(2)   sequence_base(B)    :- seq(_,B).
(3)   0 {pairing(X,Y):sequence_index(Y)} 1 :-
                sequence_index(X).
```

- (1-2) collect domains for indexes and bases
- (3) defines the pairing: a partial function (at most one association for X)

ASP Encoding

Pairings

```
(1)   sequence_index(X)  :- seq(X,_).
(2)   sequence_base(B)   :- seq(_,B).
(3)   0 {pairing(X,Y):sequence_index(Y)} 1 :-
                sequence_index(X).

(4)   :-sequence_index(X1), sequence_index(X2),
        sequence_index(Y),X1<X2,
        pairing(X1,Y), pairing(X2,Y).
(5)   pairing(B,A):- sequence_index(A), sequence_index(B),
        pairing(A,B).
```

- (4) Injective constraint: the pairing can't cover a base two times
- (5) The pairs are symmetric

ASP Encoding

Pairings

```
(1)   sequence_index(X)  :- seq(X,_).
(2)   sequence_base(B)   :- seq(_,B).
(3)   0 {pairing(X,Y):sequence_index(Y)} 1 :-
                sequence_index(X).

(4)   :-sequence_index(X1), sequence_index(X2),
        sequence_index(Y),X1<X2,
        pairing(X1,Y), pairing(X2,Y).
(5)   pairing(B,A):- sequence_index(A), sequence_index(B),
        pairing(A,B).

(6)   wrong(X,X):- sequence_base(X).
(7)   wrong(a,c).  wrong(a,g). wrong(c,u).
(8)   :-wrong(B1,B2),seq(X1,B1),seq(X2,B2),pairing(X1,X2).
```

- Discarded associations
- Can't pair two discarded associations

ASP Encoding

Pairings

```
(9)   :- sequence_index(X1), sequence_index(X2), X1=X2+1,
         pairing(X1,X2).
```

- Cannot pair consecutive basis—chemical constraint

ASP Encoding

Pairings

```
(9)   :- sequence_index(X1), sequence_index(X2), X1=X2+1,
         pairing(X1,X2).

(10)  :- sequence_index(X1), squence_index(X2),
         sequence_index(X3), sequence_index(X4),
          X1<X3,X3<X2,X2<X4,
          pairing(X1,X2),pairing(X3,X4).
```

- No Pseudo-knots (optional)

ASP Encoding

Pairings

```
(9)   :- sequence_index(X1), sequence_index(X2), X1=X2+1,
         pairing(X1,X2).

(10)  :- sequence_index(X1), squence_index(X2),
         sequence_index(X3), sequence_index(X4),
          X1<X3,X3<X2,X2<X4,
          pairing(X1,X2),pairing(X3,X4).

(11)  contacts(C):- C = #count{ (A,B):pairing(A,B) }.
(12)  #maximize { C:contacts(C) }.
```

- Maximize the number of pairings (Nussinov Energy Function)

ASP Encoding

- Alternative energy functions are possible
- Statistics: 35% AU, 53% CG, 12% GU
- $NC = n - \#(contacts)$
- minimize:

$$c_1 \frac{NC}{n} + c_2 \frac{\#(AU) - 0.35(n - NC)}{n} + c_3 \frac{\#(CG) - 0.53(n - NC)}{n}$$

ASP Encoding

```
energy(E) :- total(N), contacts(C), au(AU),
     cg(CG),
     E = c1*(N-C) + c2*(100*AU-35*C) +
          c3*(100*CG - 53*C).
#minimize{E:energy(E)}.
```

(Some) References

- M. Zucker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleid Acid Research, 9(1):133–148, 2981.
- R.B. Lyngsø and C.N.S Pedersen. RNA Pseudoknot prediction in Energy-Based Models. J. of Computational Biology 7(3/4), 2000.
- G. Blin, G. Fertin, I. Rusu, and C. Sinoquet. Extending the hardness of RNA secondary structure comparison. LNCS 4614, pp. 140–151, 2007.
- M. Bauer, G.W. Klau, and K. Reinert. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics, 8, 2007.
- M. Bavarian and V. Dahl. Constraint Based Methods for Biological Sequence Analysis. J. Universal Computer Science 12(11):1500–1520, 2006 (also in WCB 05).
- A. Dal Palù, M. Möhl, S. Will. A Propagator for Maximum Weight String Alignment with Arbitrary Pairwise Dependencies. CP 2010: 167-175 (also in WCB 10)