Oznamy

- Deadline of HW2 extended until Dec. 7
- HW3 will be published next week
- Next Thursday Dec.2: lecture and tutorials cancelled
- Thursday Dec.9: lecture and tutorials online
- Thursday Dec.16:
  - optional presentations of journal club during lecture time
  - tutorial for comp.sci. will take place
  - tutorial for biologists possibly cancelled
- End of semester deadlines
  - HW3 Tuesday Dec. 14, journal club reports Friday Dec. 17
- On Thursday Dec. 9, we will discuss:
  - if you want to present journal club (discuss in the group)
  - date of the exam (bring dates of other exams)

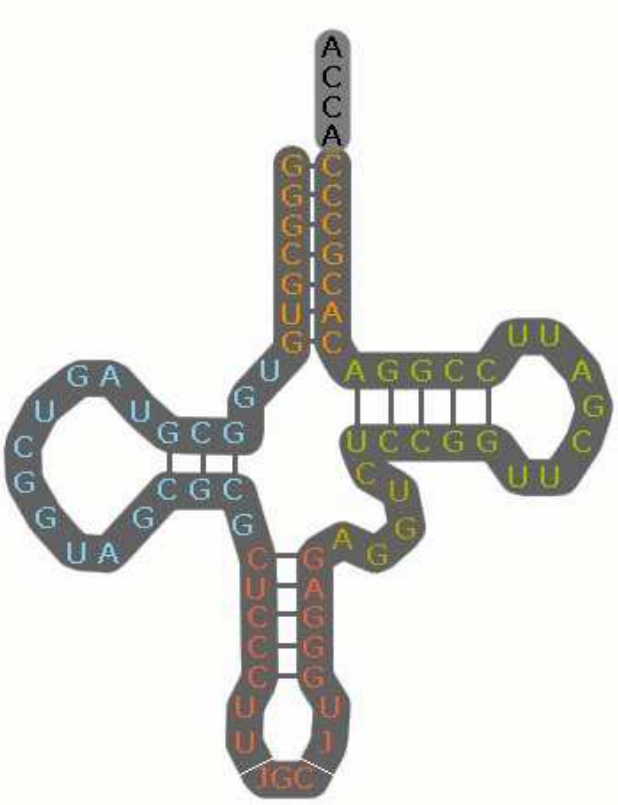
Recall: journal club report

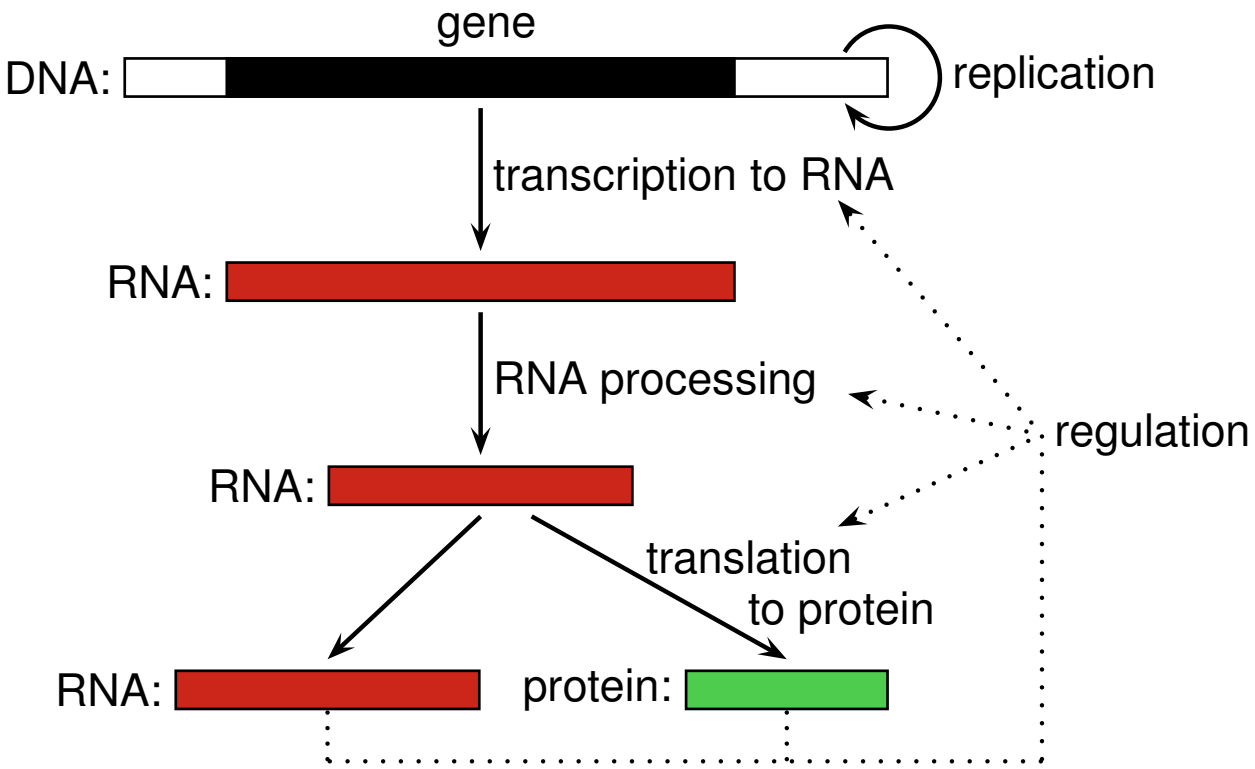
- The main methods and results of the article in your own words
- Understandable for students of this course (both comp.sci. and bio)
- You do not have to cover the entire content of the article in the report and, conversely, you can use other resources
- Try to express your own view of the topic, do not strictly follow the text of the article
- The recommended length is about 1-2 pages per person, one coherent text
- The report should list the members of the group who have actively participated. They will get the same points (the rest zero)
- Submit via Moodle, 1 pdf per group

RNA

Tomáš Vinař

Nov. 25, 2021




gene
DNA:
replication
transcription to RNA
RNA:
RNA processing
regulation
RNA:
translation
to protein
RNA:
protein:

Properties of RNA

Differences from DNA

- contains ribose instead of deoxyribose
- contains uracyl instead of thymine (bases A,C,G,U)
- single-stranded molecules, usually shorter
- complex secondary structure with paired complementary regions
- pairs A-U, C-G as well non-canonocal pairs e.g. G-U
- various functions in the cell:
  central role in gene expression (messanger RNA, transfer RNA, ribosomal RNA),
  regulation of expression,
  catalythic functions,
  transfer of genetic information for RNA viruses

RNA structure

Example: transfer RNA

Secondary structure:

pairing of nucleotides

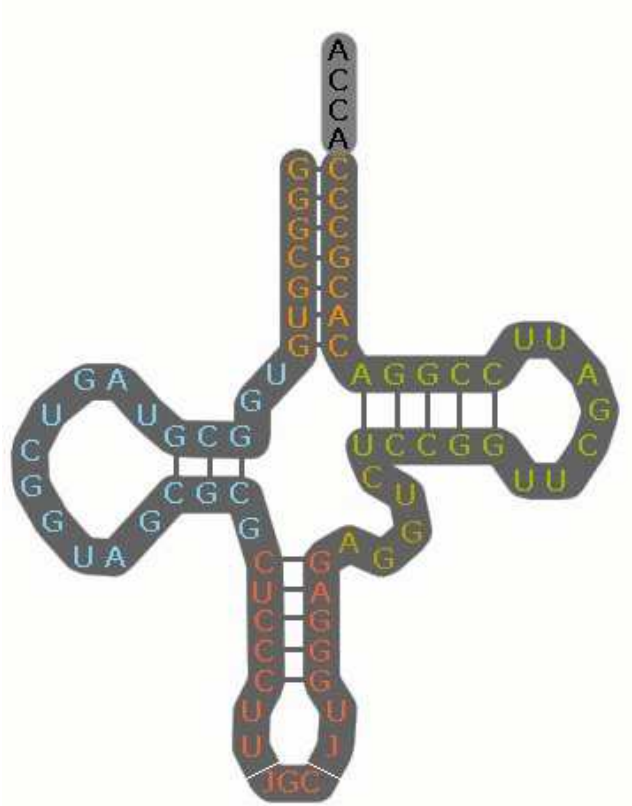
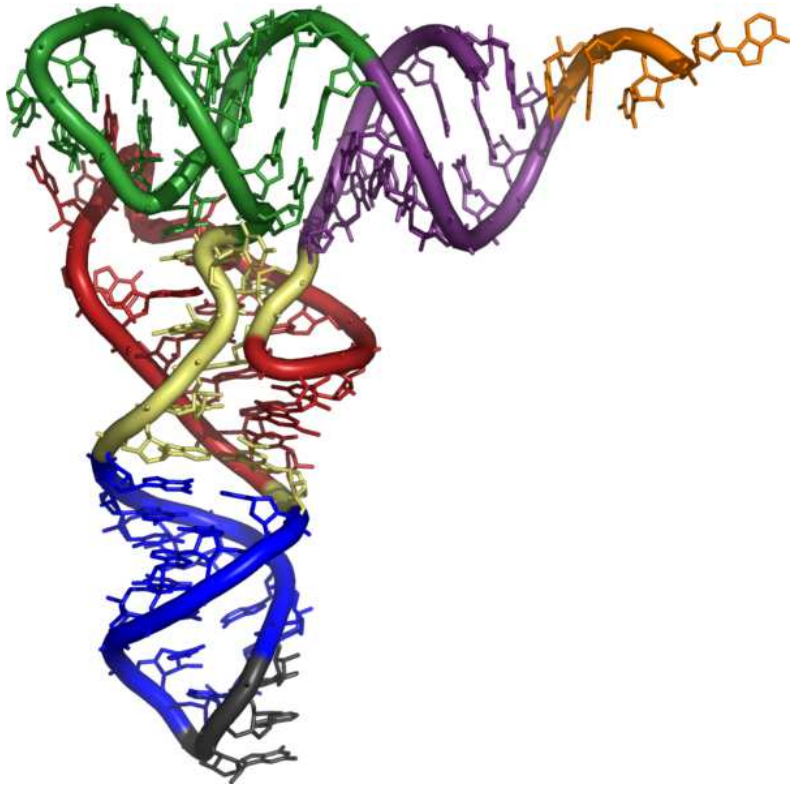


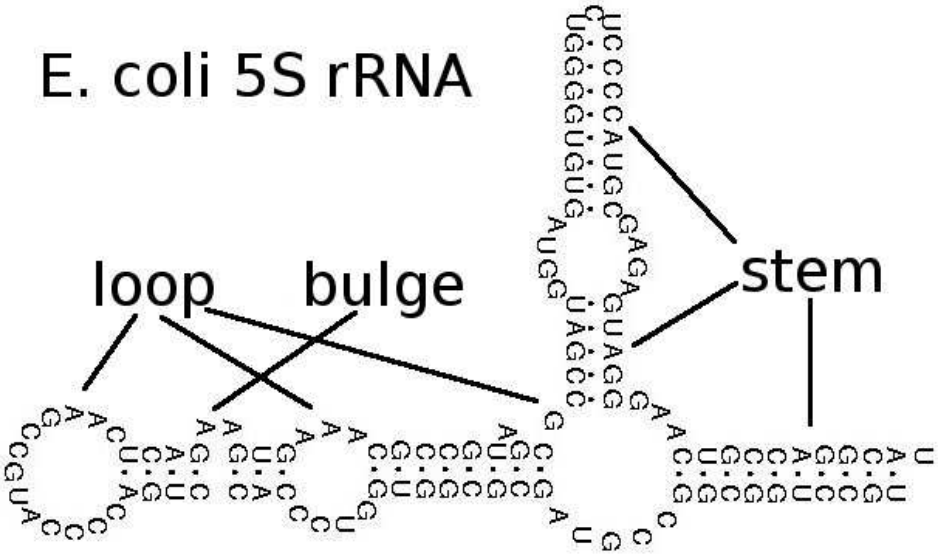
Figure source: Wikipedia

Tertiary structure:

3D coordinates



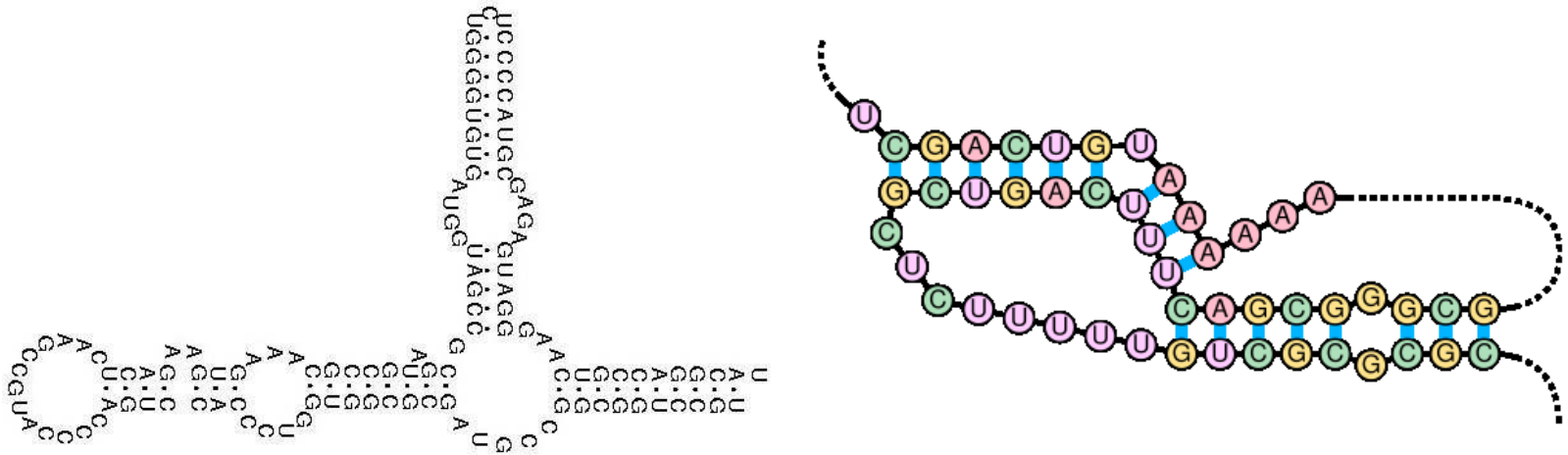
Sekundárna štruktúra RNA



Representation using well-parenthesized expression:

```
((((((((((.....(((   ).)).((   ))...)))))))))).
UGCCUGGCGGCCGUAGCG...UAGCGCC...GGGAACUGCCAGGCAU
```

Well-parenthesized expression vs. pseudoknots



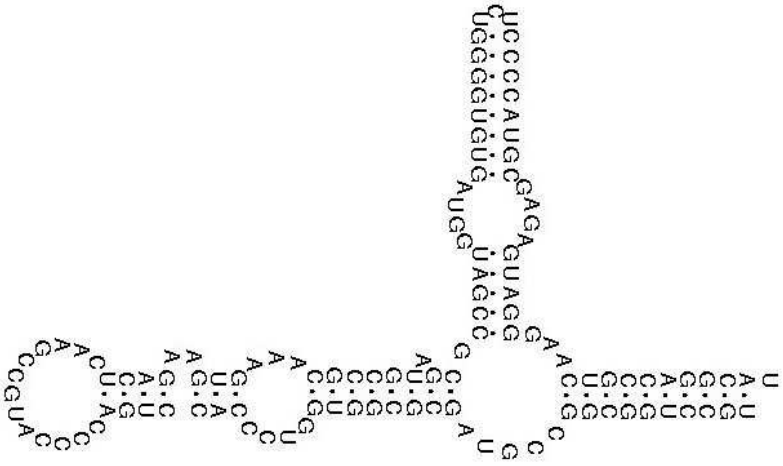
Left: can do well-parenthesized expression

```
((((((((((.....(((    ).)).((    )).....)))))))))).
UGCCUGGCGGCCGUAGCG...UAGCGCC...GGGAACUGCCAGGCAU
```

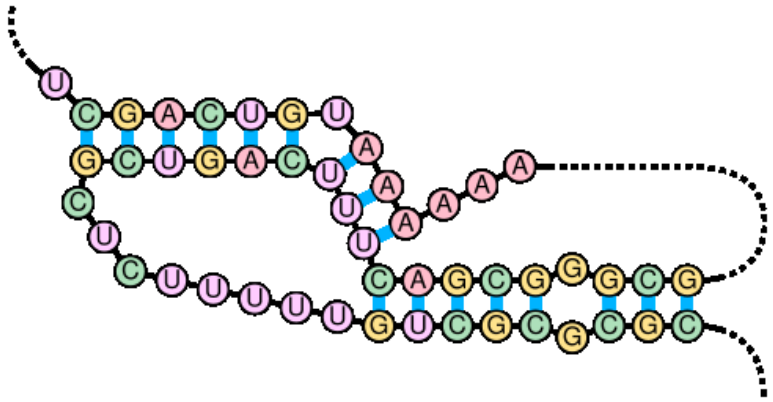
Right: pseudoknot (cannot do well-parenthesized expression)

```
.((((((.(((...[[[.[[[[[)))))))))).    .]]]]].]]]
UCGACUGUAAAAAAGCGGGCGACUUUCAGUCGC...UGUCGCGCGC
```

Well-parenthesized expression vs. pseudoknots



without pseudoknots

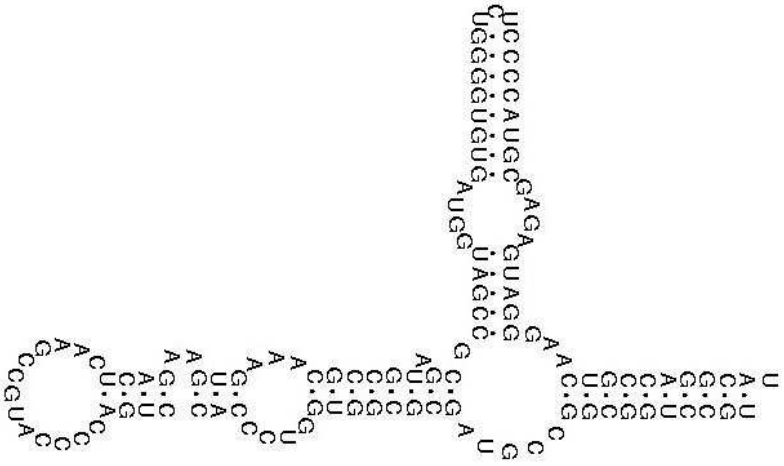


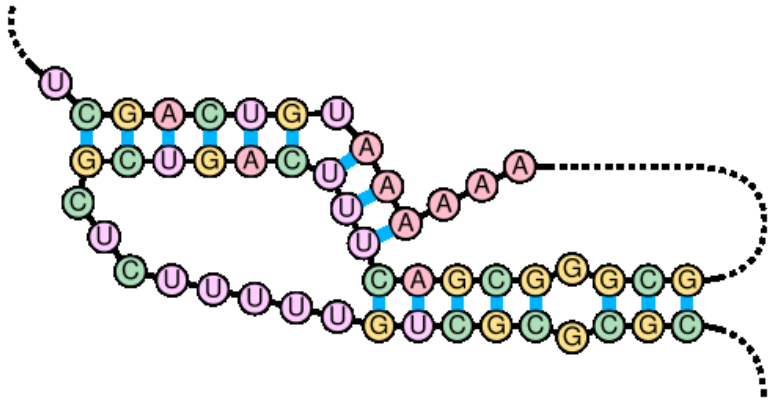
pseudoknot

Approx. 1.4% of paired RNA bases involved in pseudoknots

Yet many algorithms **ignore pseudoknots**

Well-parenthesized expression vs. pseudoknots

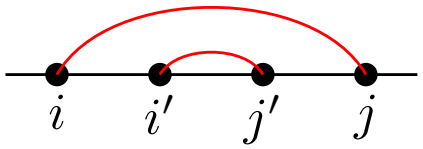




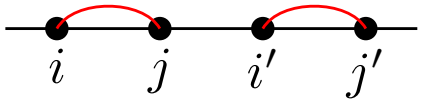
Mathematical structure of secondary structure w/o pseudoknots:

If position i is paired with j and position i' with j'
where $i < i'$ then either $i < i' < j' < j$ or $i < j < i' < j'$.

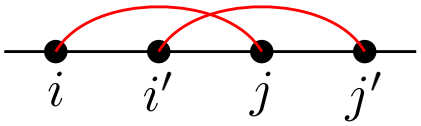
good:



good:



bad:



Problem: determining secondary RNA structure

Input: RNA sequence

Goal: find which bases are paired

Simplified formulation: find well-parenthesized expression corresponding to the structure with the highest number of complementary pairs A-U, C-G.

Example:

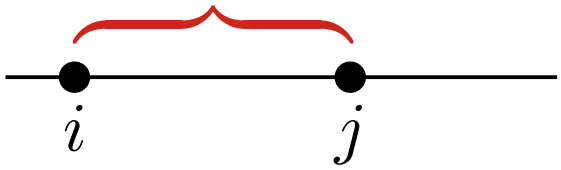
```
Input:  GAACACAUGUAAAAUUUGUC
Output: ((.((((()))(((.))))))
```

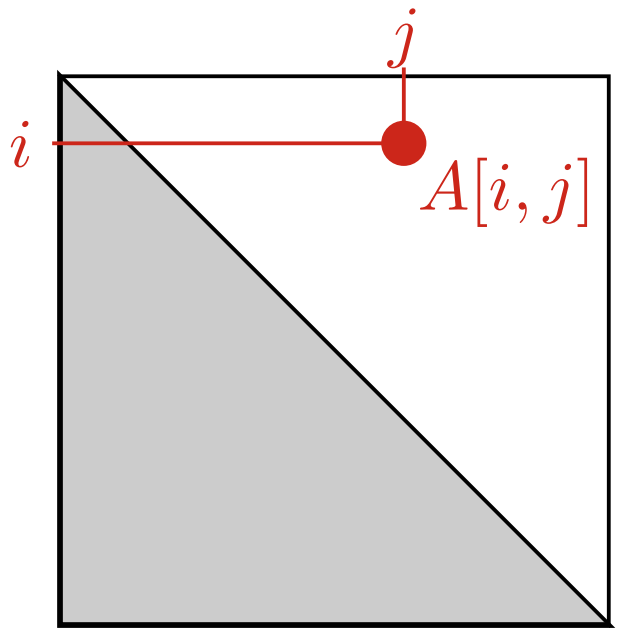
Nussinov algorithm

Dynamic programming:

Given RNA $x_1, \ldots, x_n$.

$A[i, j] =$ the maximum number of matched pairs in $x_i, x_{i+1}, \ldots, x_j$





Nussinov algorithm

Dynamic programming:

Given RNA $x_1, \ldots, x_n$.

$A[i,j]$ = the maximum number of matched pairs in $x_i, x_{i+1}, \ldots, x_j$

Recurrence:

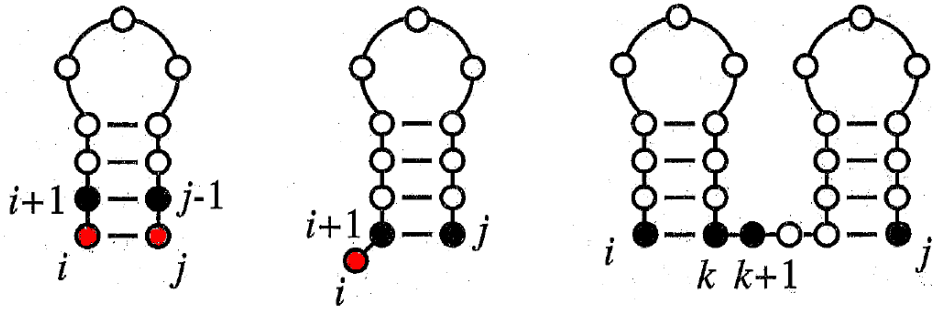
Substrings of length 1: no pairs possible $\Rightarrow A[i,i] = 0$

Longer substrings:

– x_i not involved in a pair: $A[i,j] = A[i+1,j]$

– x_i paired with x_j: $A[i,j] = A[i+1,j-1] + c(x_i, x_j)$

– x_i paired with x_k $(k < j)$: $A[i,j] = A[i,k] + A[k+1,j]$

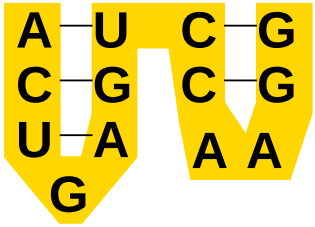


Rekurencia: $A[i,j] = \max \begin{cases} A[i+1,j], \\ A[i+1,j-1] + c(x_i, x_j), \\ \max_{k=i+1\ldots j-1}\{A[i,k] + A[k+1,j]\} \end{cases}$

	A	C	U	G	A	G	U	C	C	A	A	G	G
A	0	0	1	1	1	2	3	3	3	3	3	4	5
C		0	0	1	1	2	2	2	2	3	3	4	4
U			0	0	1	1	1	2	2	3	3	3	3
G				0	0	0	1	2	2	2	2	3	3
A					0	0	1	1	1	1	1	2	3
G						0	0	1	1	1	1	2	2
U							0	0	0	1	1	1	2
C								0	0	0	0	1	2
C									0	0	0	1	1
A										0	0	0	0
A											0	0	0
G												0	0
G													0

$$c(x_i, x_j) = \begin{cases} 1 & \text{if } x_i\text{-}x_j \text{ is A-U or C-G pair} \\ 0 & \text{otherwise} \end{cases}$$

$A[i,j] = 0$ for $i \geq j$



Complexity:

$O(n^3)$ time

$O(n^2)$ memory

Minimum free energy (MFE) folding

More realistic formulation

Assumption: the molecule in the state of equilibrium with minimum Gibbs free energy.

Energies for modules measured experimentally.

Nearest neighbor model: parameters = energies for neighbouring pairs in helixes, lengths of loops, etc.

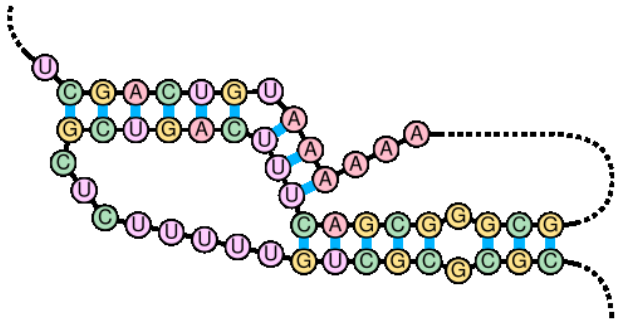
Derived from experimental measurements.

Example:

```
                          y:  A    C    G    U
5'  Cx  3'                    --------------------
3'  Gy  5'              x:A | .    .    .  -2.1
                            C | .    .  -3.3  .
                            G | .  -2.4  .  -1.4
                            U | -2.1  .  -2.1  .
```

Algorithms similar to the Nussinov algorithm [Zuker and Stiegler, 1981].

Algorithms allowing pseudoknots



NP-hard in general [Lyngso and Pedersen, 2000].

Slow dynamic programming $O(n^4)$ – $O(n^6)$ for certain pseudoknot types [Rivas and Eddy, 1999].

Or use heuristics [Ren et al., 2005] (repeated greedy formation of strong helixes).

Probabilistic models for RNA secondary structure prediction

Want: Generative model for pairs sequence, secondary structure
Use: For a given sequence, find most probable structure

HMMs are **not** suitable: cannot capture dependencies between distant pairs

Solution: **Stochastic context-free grammars (SCFGs)**

- extension of context-free grammars
- individual rules will get probabilities

Stochastic context-free grammars (SCFGs)

non-terminals (upper-case) similar to states in HMMs

terminals (lower-case) represent nucleotides

rules rewrite non-terminals to strings of terminal and non-terminals

each rule has assigned probability

Example: single non-terminal, 14 rules (ϵ =empty string)

$$S \to \overbrace{aSu}^{0.1} \mid \overbrace{uSa}^{0.1} \mid \overbrace{cSg}^{0.1} \mid \overbrace{gSc}^{0.1} \mid$$

$$\overbrace{aS}^{0.05} \mid \overbrace{cS}^{0.05} \mid \overbrace{gS}^{0.05} \mid \overbrace{uS}^{0.05} \mid \overbrace{Sa}^{0.05} \mid \overbrace{Sc}^{0.05} \mid \overbrace{Sg}^{0.05} \mid \overbrace{Su}^{0.05} \mid \overbrace{SS}^{0.1} \mid \overbrace{\epsilon}^{0.1}$$

In each step choose the left-most non-terminal

rewrite with a randomly chosen rule:

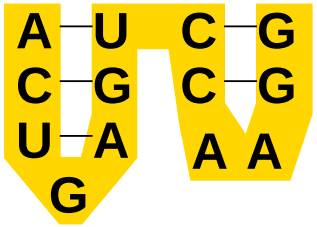
$$S \to SS \to aSuS \to acSguS \to acuSaguS \to acugSaguS \to$$
$$acugaguS \to acugagucSg \to acugagucc Sgg \to acugagucc Sagg \to$$
$$acugagucc aSagg \to acuguguccaagg$$

Stochastic context-free grammars

Example:

$S \to aSu|uSa|cSg|gSc|aS|cS|gS|uS|Sa|Sc|Sg|Su|SS|\epsilon$

$S \to SS \to aSuS \to acSguS \to acuSaguS \to acugSaguS \to$
$acugaguS \to acugagucSg \to acugagucgScg \to acugagucgSacg \to$
$acugagucgaSacg \to acugugucgaacg$



Problem: Find most probable derivation of given RNA

Bases generated in a single rule represent **paired bases**

Solution: Dynamic programming, algorithm CYK, $O(n^3)$ time

Training: Probabilities trained from known RNA structures

Grammars vs. energy minization

Grammar advantages:

- parameters can be trained automatically, no expensive experiments
- can be extended to multiple sequences

Grammar disadvantages:

- simple grammars do not capture full complexity of the problem
- lower accuracy

RNA sequence evolution

Often correlation between mutations in paired bases
e.g. C changes to A, paired G changes to U simultaneously

Example: several sequences from t-RNA D-arm

```
((((..............))))
GCUCAGCC.CGGG...AGAGC
GCCUAGCC.UGGUCA.AGGGC
GUCUAGC...GGA...AGGAU
GAGCAGUU.CGGU...AGCUC
GUUCAAUC..GGU...AGAAC
```

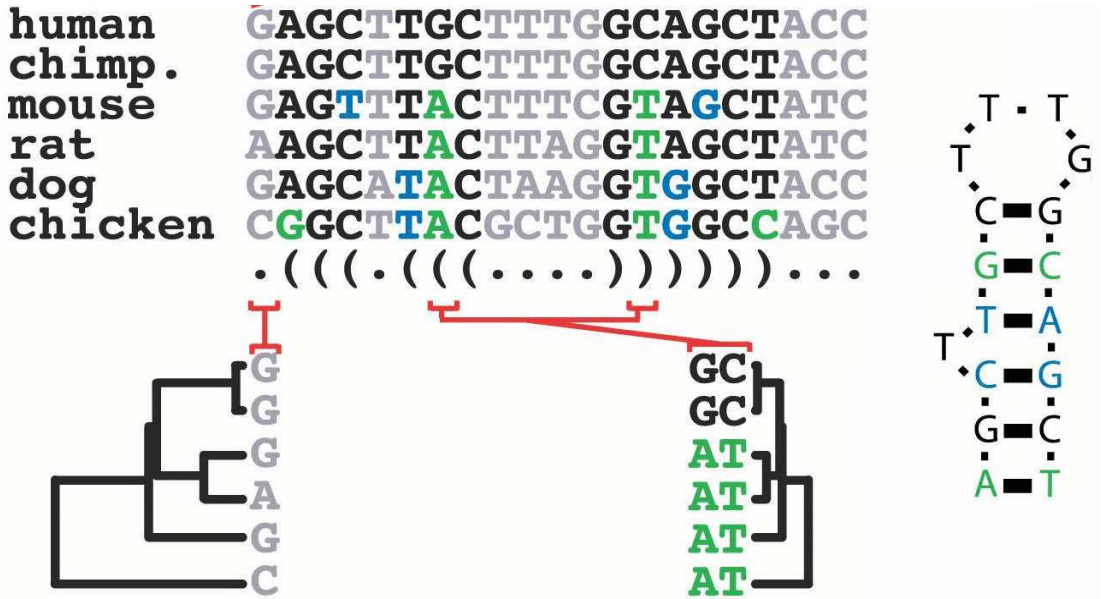
Problem: given a multiple alignment of RNA sequences
find a common RNA structure

(common structure will exhibit correlations between paired bases)

Common RNA structure for several RNA sequences

Phylo-SCFG:

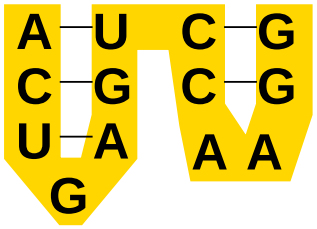
- terminals will be **whole alignment columns**
  use phylogenetic tree structure
- unpaired bases emitted using a regular substition matrix
- paired bases emitted using a 16×16 substition matrix (all pairs)



Problem: Finding known types of RNA genes in genomes

- Rfam database contains structures for > 4000 RNA families represented using probabilistic models
- Similar idea to profile HMMs used for representation of protein families (Pfam database)
- Special type of SCFGs called **covariance models**

Covariance models (CMs)



$$
\begin{array}{lll}
S \to B_1 & P_1 \to aP_2u & P_4 \to cP_5g \\
B_1 \to P_1P_4 & P_2 \to cP_3g & P_5 \to gL_2c \\
 & P_3 \to uL_1a & L_2 \to aL_3 \\
 & L_1 \to gE_1 & L_3 \to aE_2 \\
 & E_1 \to \epsilon & E_2 \to \epsilon
\end{array}
$$

- S =start, E_i =end, P_i =pair,
  L_i =unpaired base on the left, R_i =unpaired base on the right
  other non-terminals to represent indels
- terminals (bases) emitted with probabilities **specific to each alignment column**
  e.g. $P_1 \to \overbrace{aP_2u}^{0.2} \mid \overbrace{uP_2a}^{0.2} \mid \overbrace{cP_2g}^{0.4} \mid \overbrace{cP_2u}^{0.1}$

Covariance models (CMs)

Uses:

finding occurrences of a gene in DNA (local alignment),
finding structure of a new gene from a known family (global alignment).

Dynamic programming: time $O(MND^2)$

$M =$ the number of non-terminals, proportional to the alignment length

$N =$ the length of DNA ,

$D =$ max. length of an RNA gene (related to M).

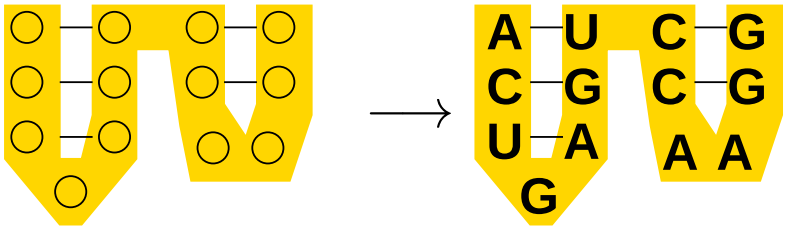
Heuristic speedup: find potential sites with sequences similar to known family representatives, apply CM only there

Problem: RNA secondary structure design

Given RNA secondary structure (pairing)

Find a sequence for which this is the optimal structure.

No known efficient algorithm, but fast heuristics work well



Use: research on possible RNA structures, drug design (ribozymes, riboswitches), RNA for laboratory techniques, RNA nanostructures

Summary

- RNA secondary structure prediction:
  energy minimization, probabilistic SCFGs
- Can achieve better results if we use a multiple alignment of several RNA sequences with a common structure (PhyloSCFG)
- Known RNA families can be represented by covariance models, these can be used to locate occurrences in novel sequences
- Rfam database
- Most problems can be solved by dynamic programming
  – somewhat slow
  – ignores pseudoknots
- Other interesting problems: RNA design