Вторичная структура РНК

Практическая биоинформатика 05.11.2024, 12.11.2024

Мичил Трофимов

Genotek

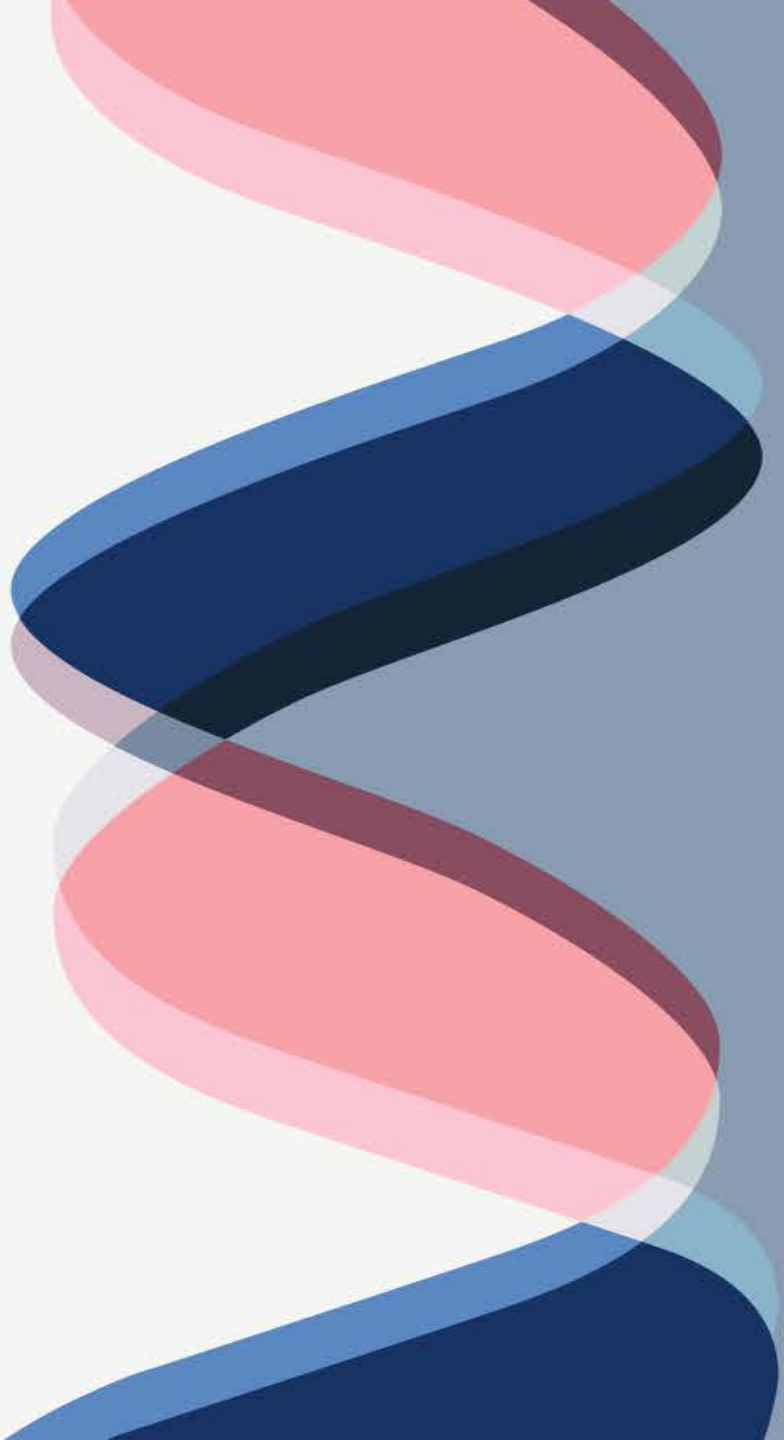
План

Теория:

- Функции РНК и их структуры
- Как предсказывать структуру РНК?
- Базы данных

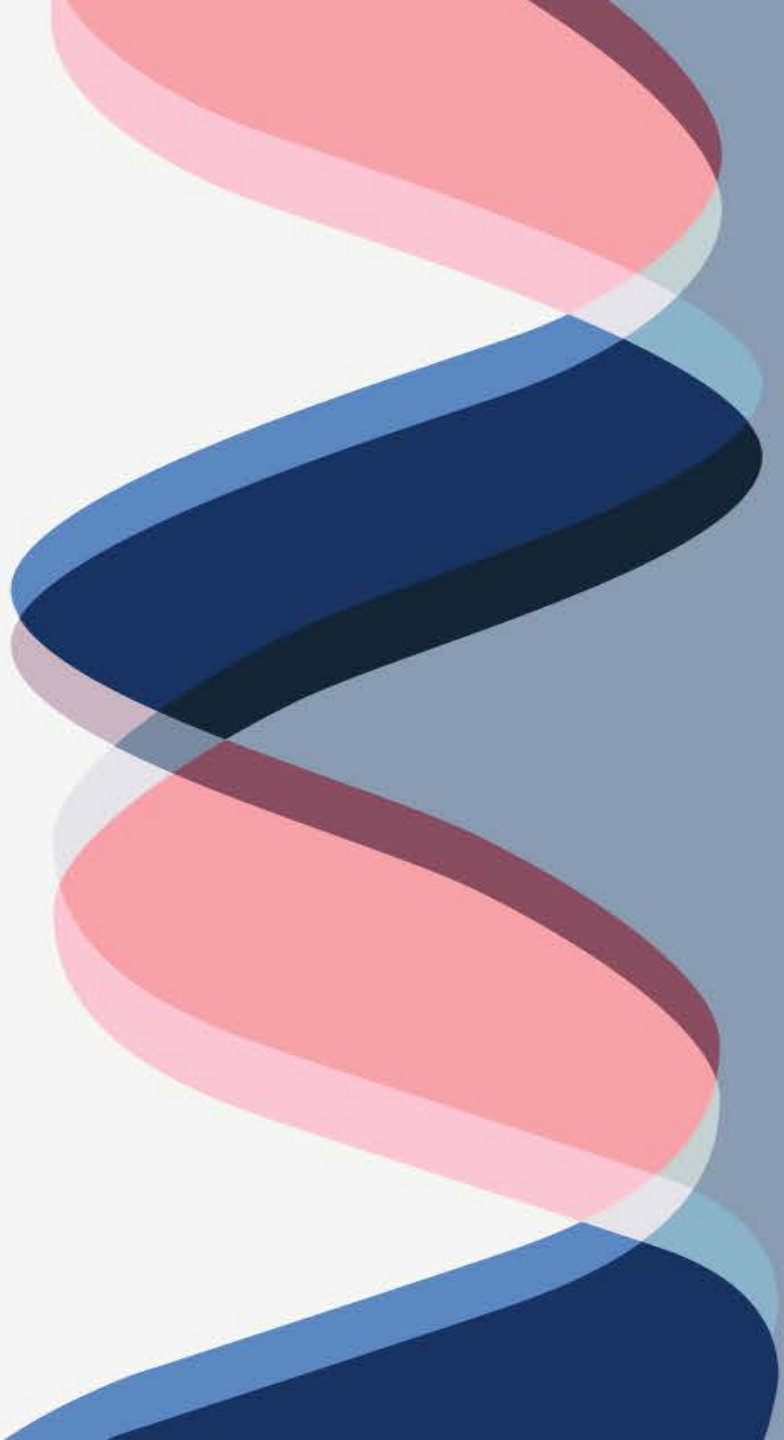
Практика:

- Предсказывание РНК структуры и оценка результата



Что вы будете знать

- Использовать тулы для предсказывания РНК структур
- Оценивать их предсказания

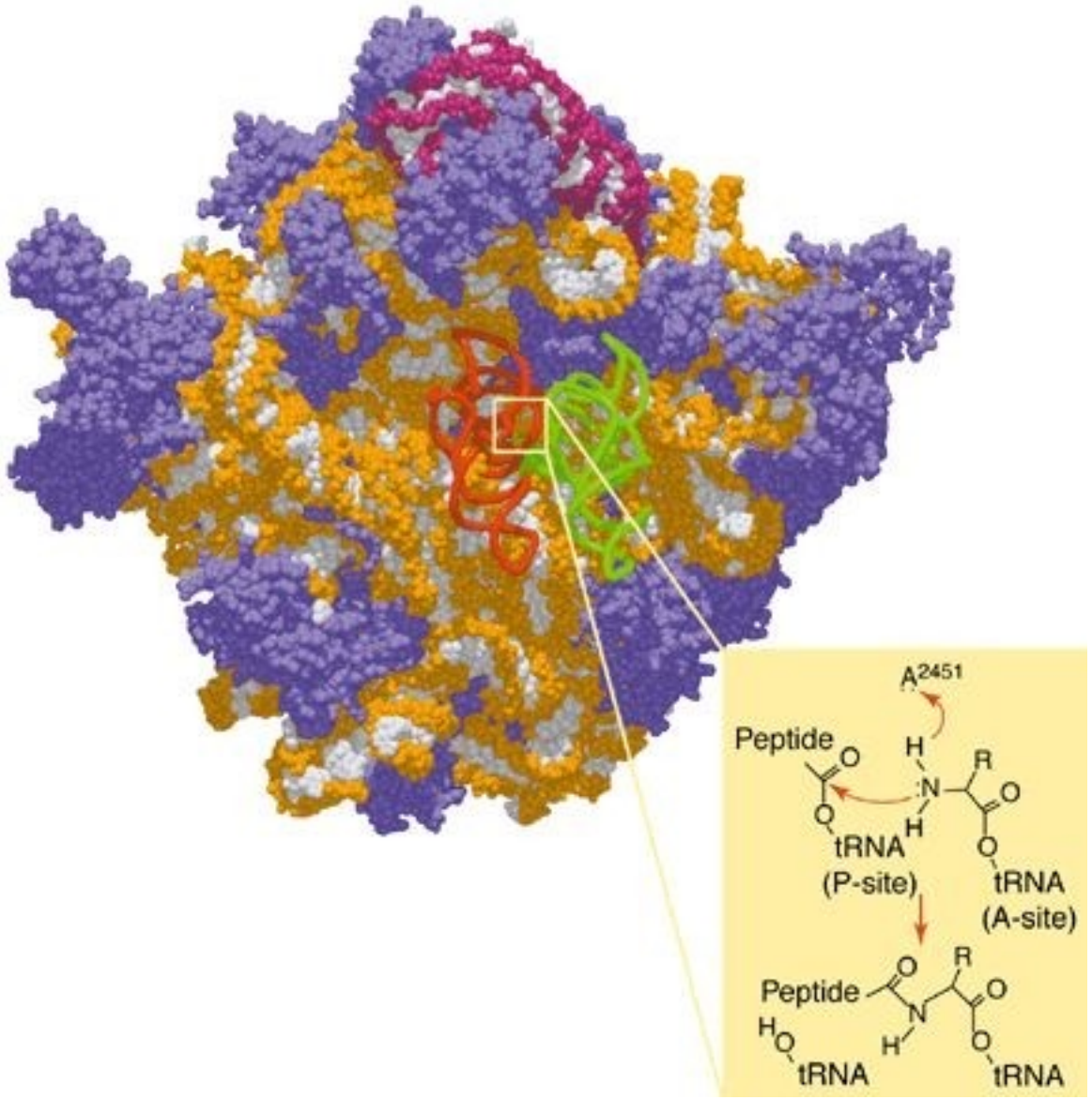


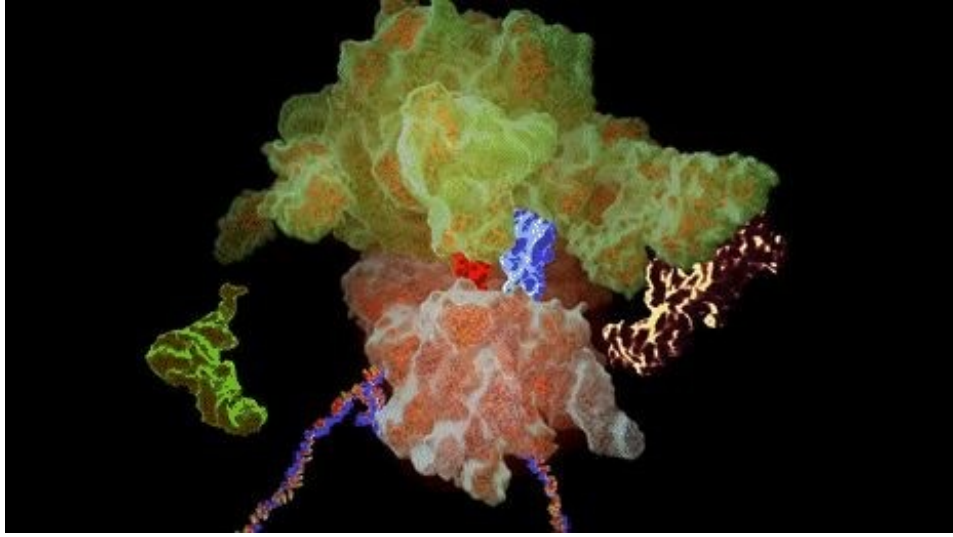
Какие типы РНК вы знаете?

Функциональные РНК

- тРНК
- рРНК
- длинные некодирующие РНК
- Малые ядерные РНК
- Малые ядрышковые РНК

rRNA

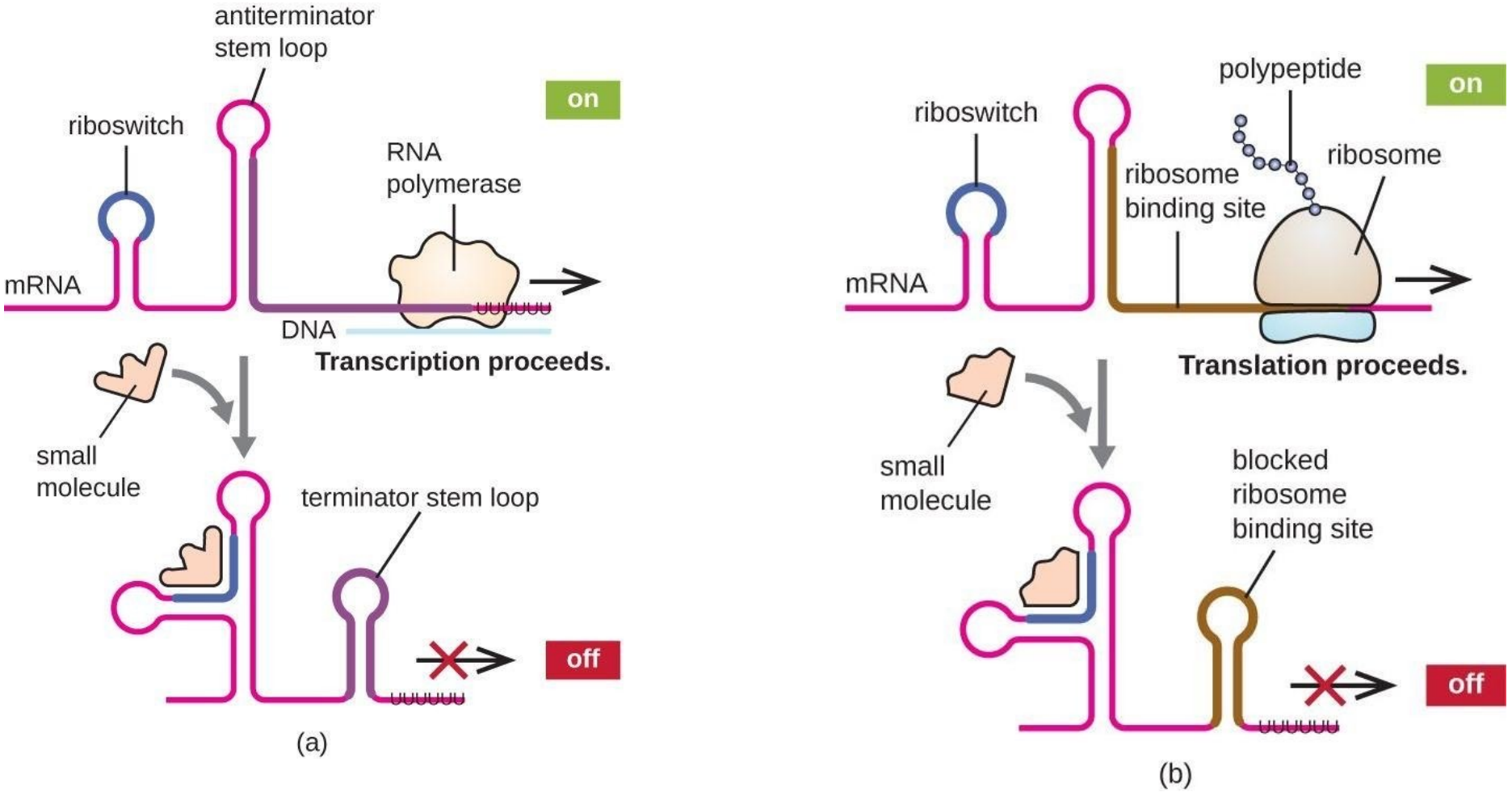




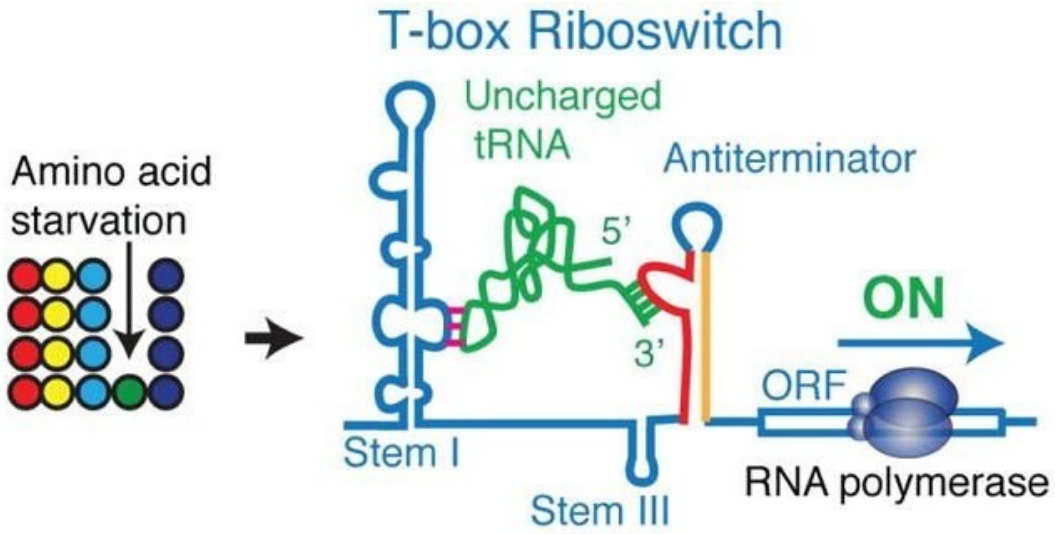
A ribosome is a ribozyme:

- the peptidyl-transfer mechanism is catalyzed by ribosomal RNA as an entropic catalyst that accelerates peptide bond formation primarily by substrate positioning
- the hydrolysis of peptidyl-tRNA is catalyzed also with the help of a release factor

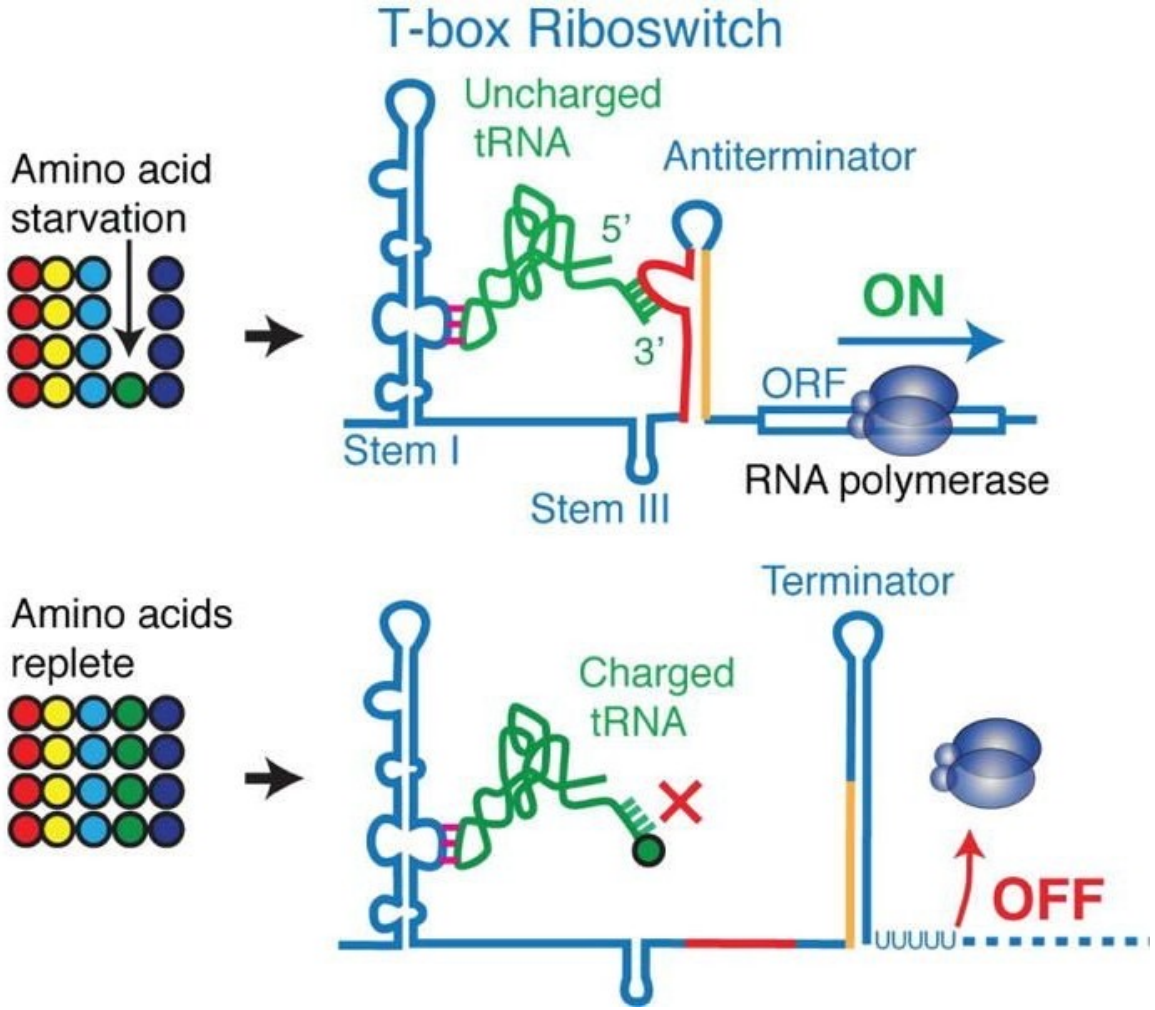
mRNA: Riboswitches



T-boxes regulate genes involved in amino acid metabolism



T-boxes regulate genes involved in amino acid metabolism

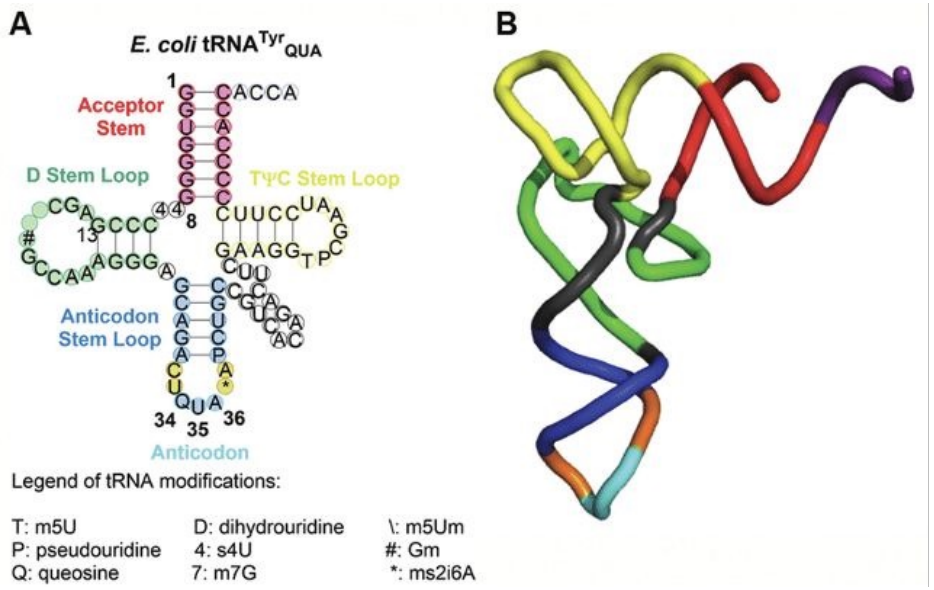


Typical genes:

- aminoacyl-tRNA synthetases
- amino acid transporters
- aminotransferases
- other proteins of the biosynthetic pathways of amino acids

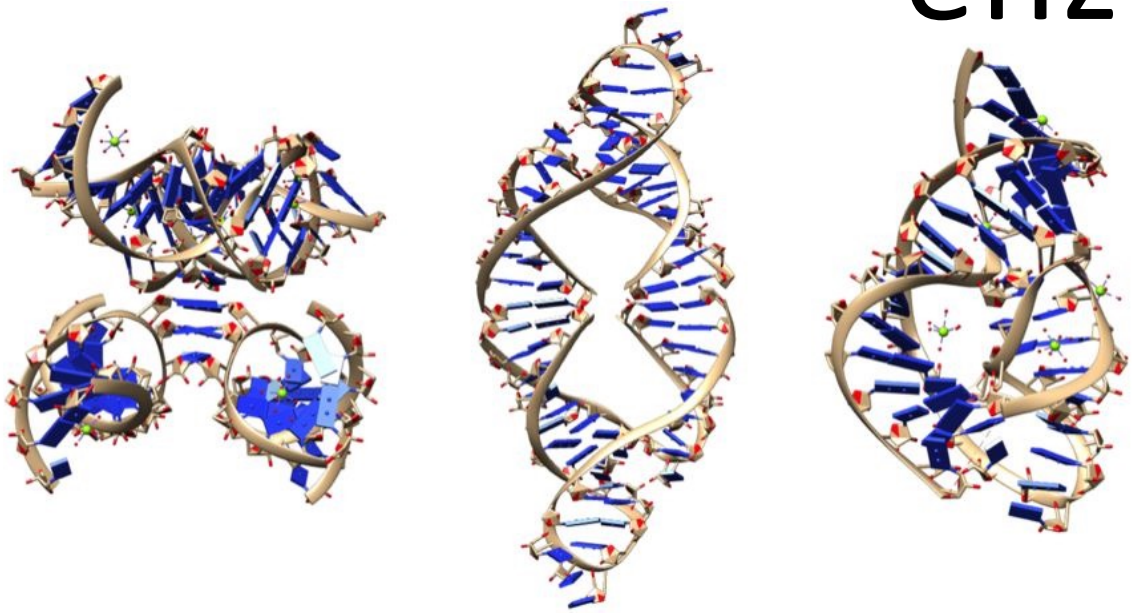
Вторичная структура РНК

Is knowing RNA secondary structure enough for doing biology?



Biddle et al "Modification of orthogonal tRNAs: unexpected consequences for sense codon reassignment," NAR, 10.1093/nar/gkw948.

Ribozymes: what makes them enzymes?

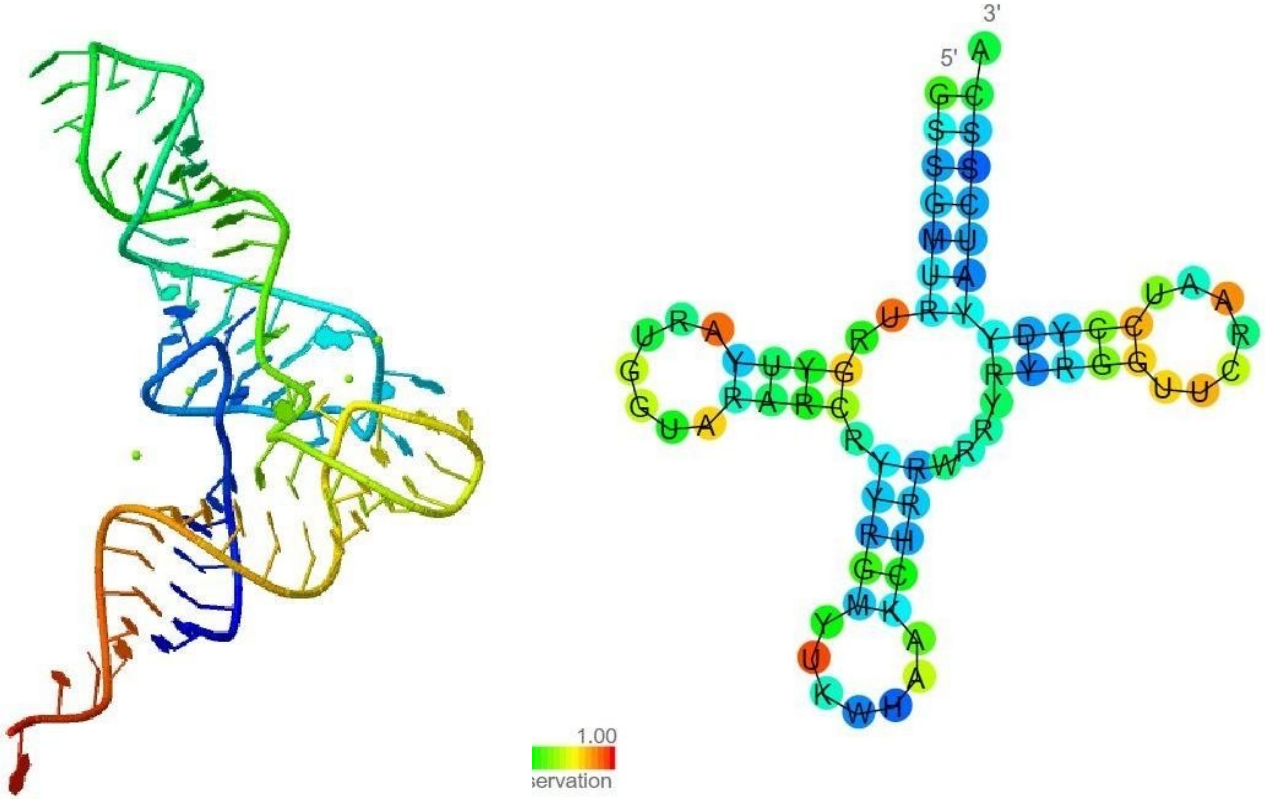


From left to right: leadzyme, hammerhead ribozyme, twister ribozyme

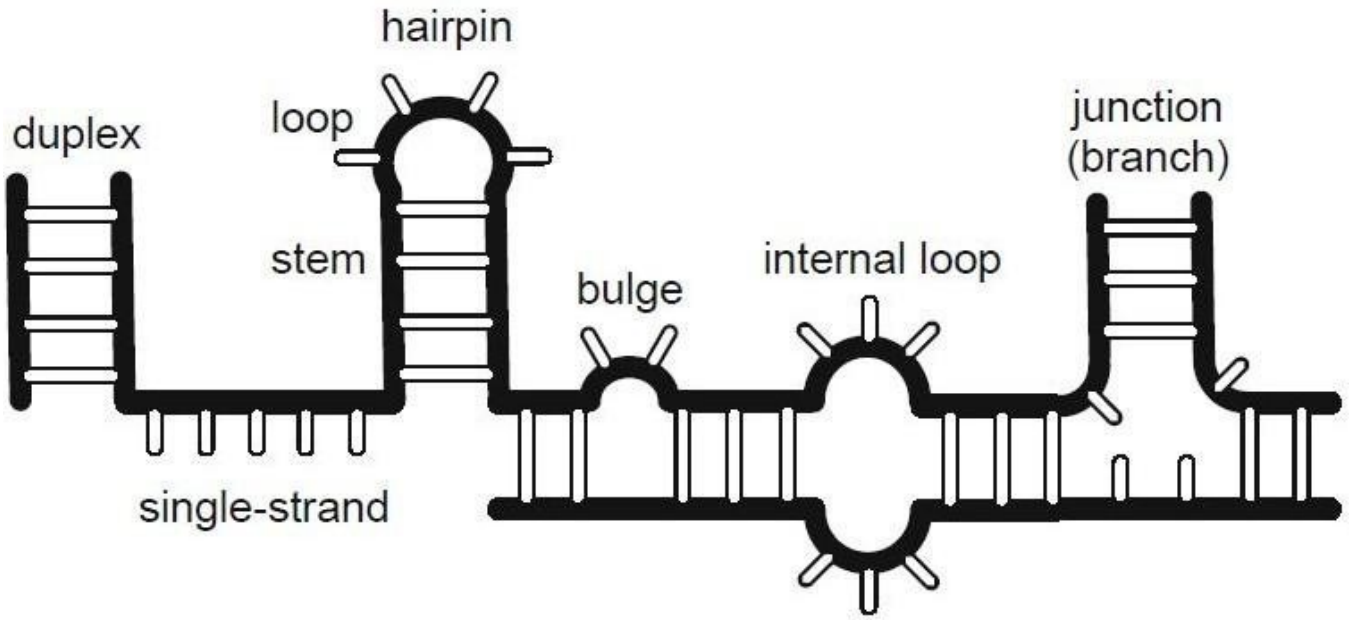
The ability of RNA to form complex secondary/tertiary structures provides a structural basis for catalytic activity like a protein fold does.

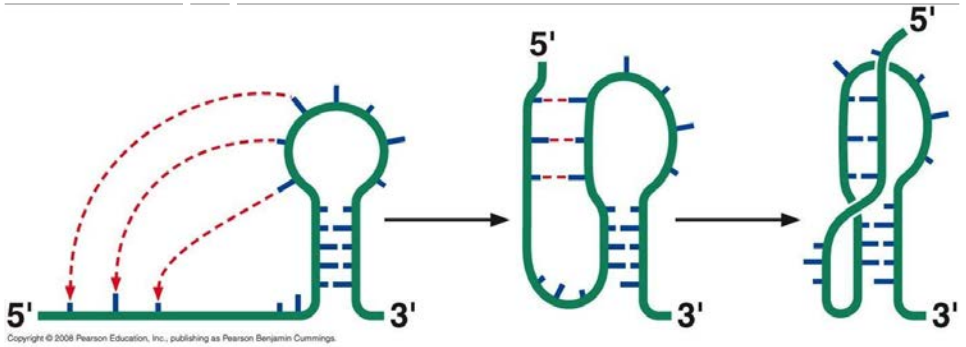
NB! Ribozymes often need a metal ion for the catalytic activity

Secondary structure is just one level …but we can work (only) with it



Naming conventions



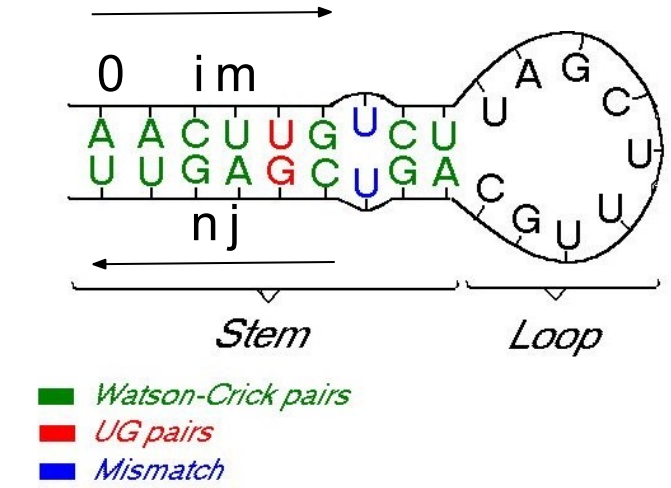


Common assumptions

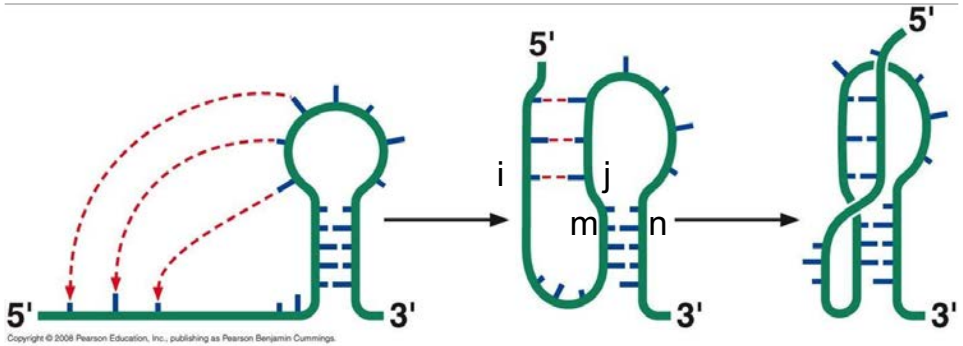
- No knots: no pairs (i, j) and (m, n) such that
  i < m < j < n
  - No "close" base pairs:
    (j - i) always > t for some t > 0
    often t = 4
  - Complementary base pairs:

    C≡G , A=U, the wobble pair G~U

A hairpin without pseudoknots

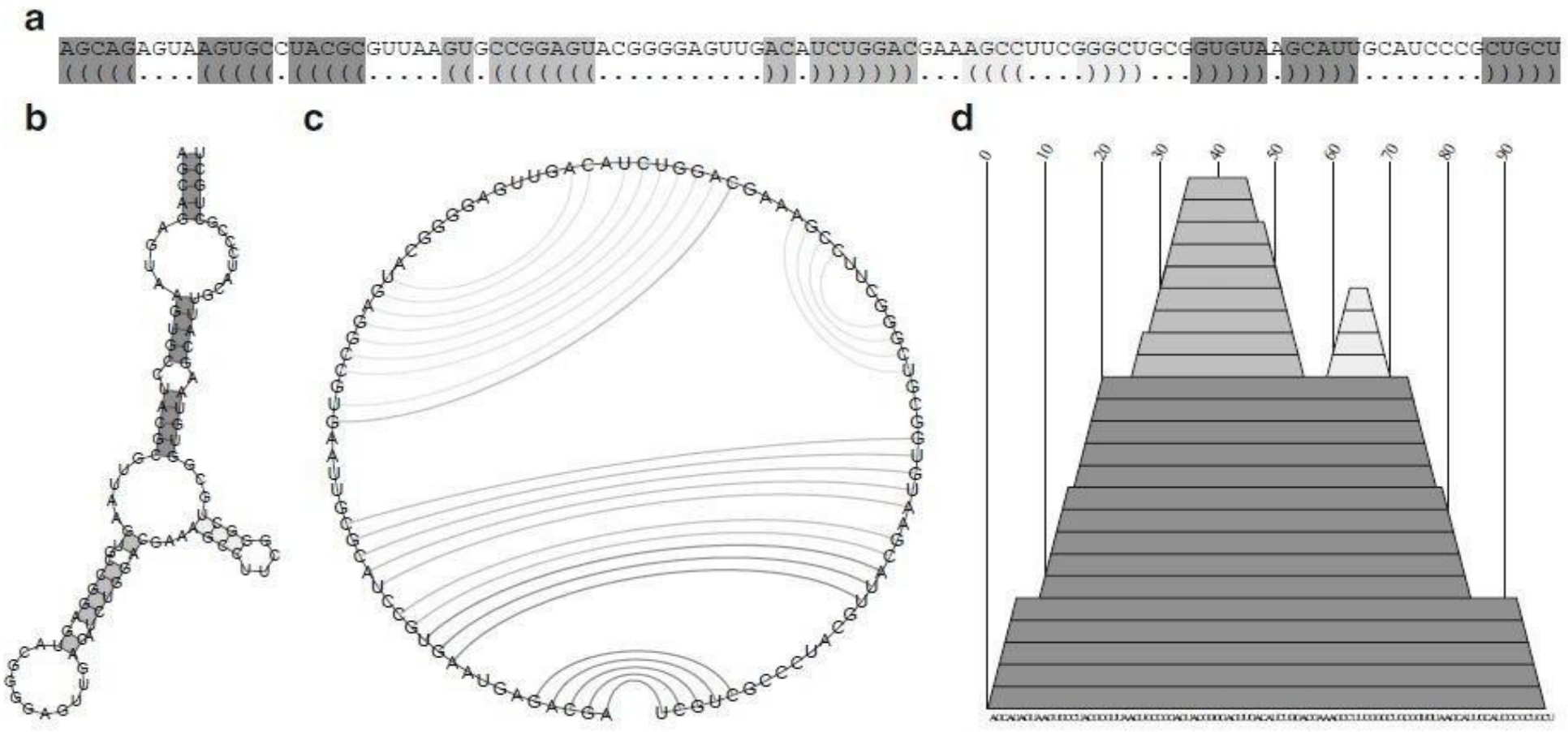


A pseudoknot



http://www.bioinf.man.ac.uk/resources/phase/manual/node72.html

How to visualize

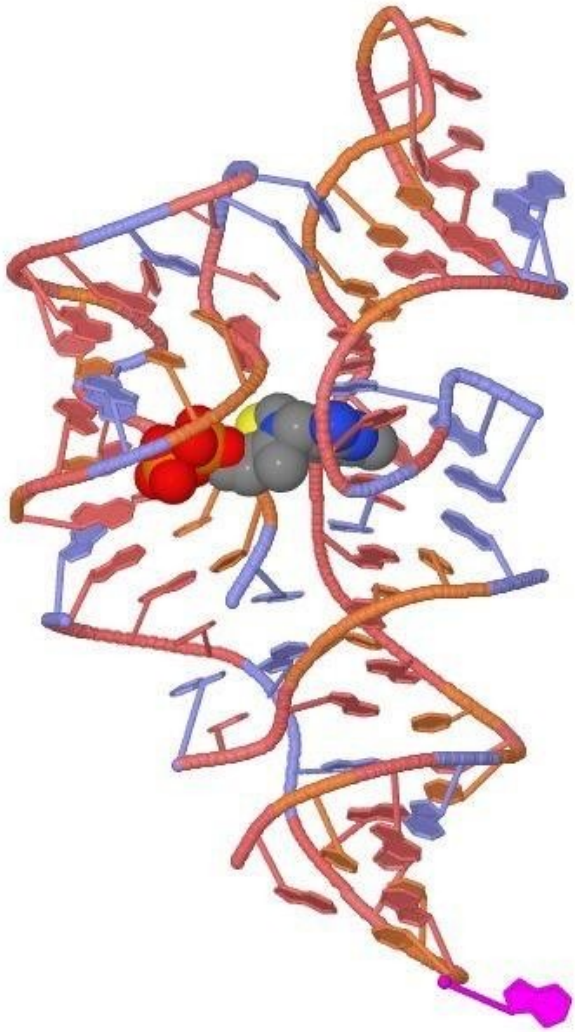


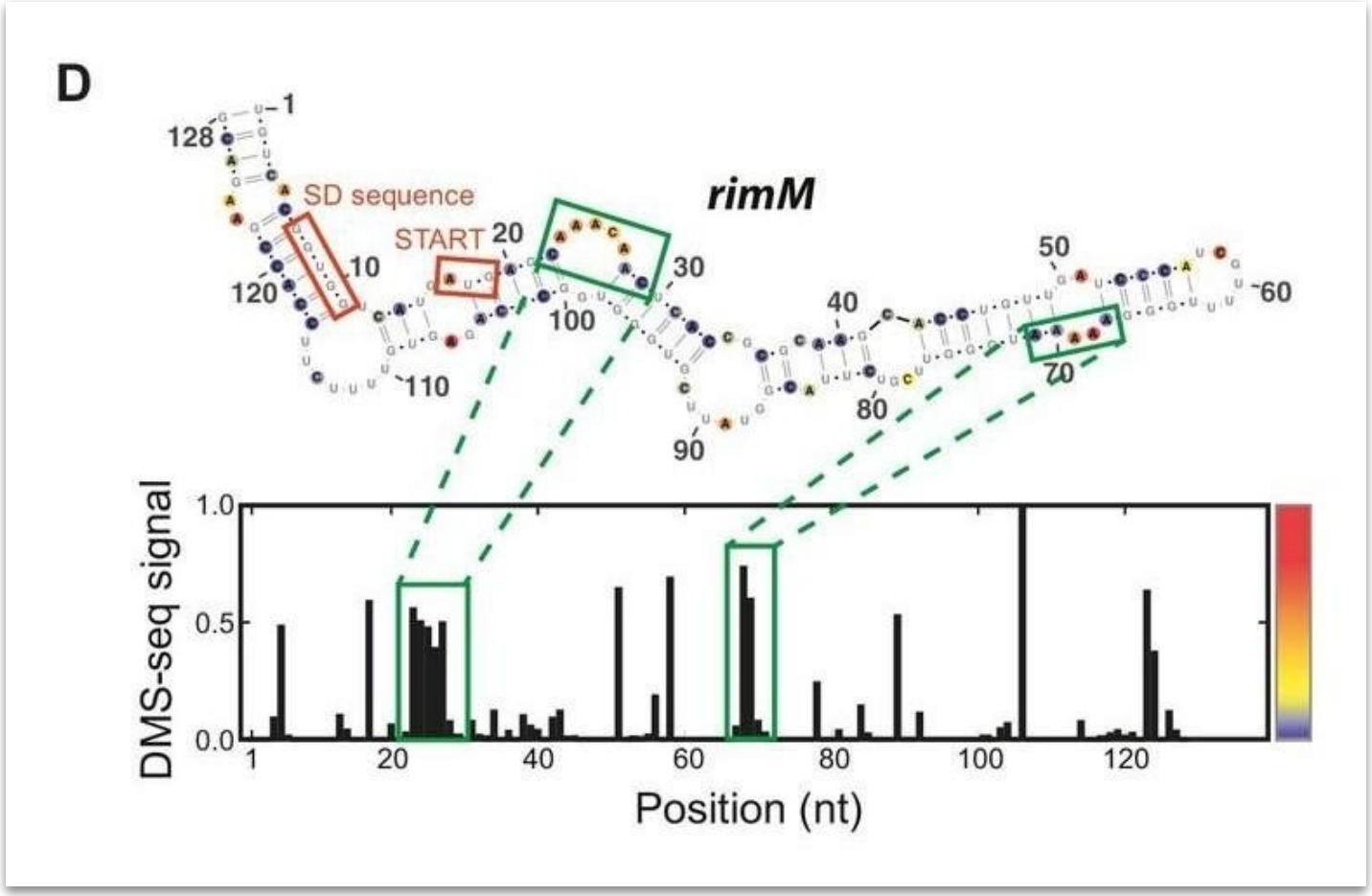
(a) dot bracket notation
(b)a secondary structure diagram
(c) a planar circle plot
(d) a mountain plot

Gorodkin at al, 2014

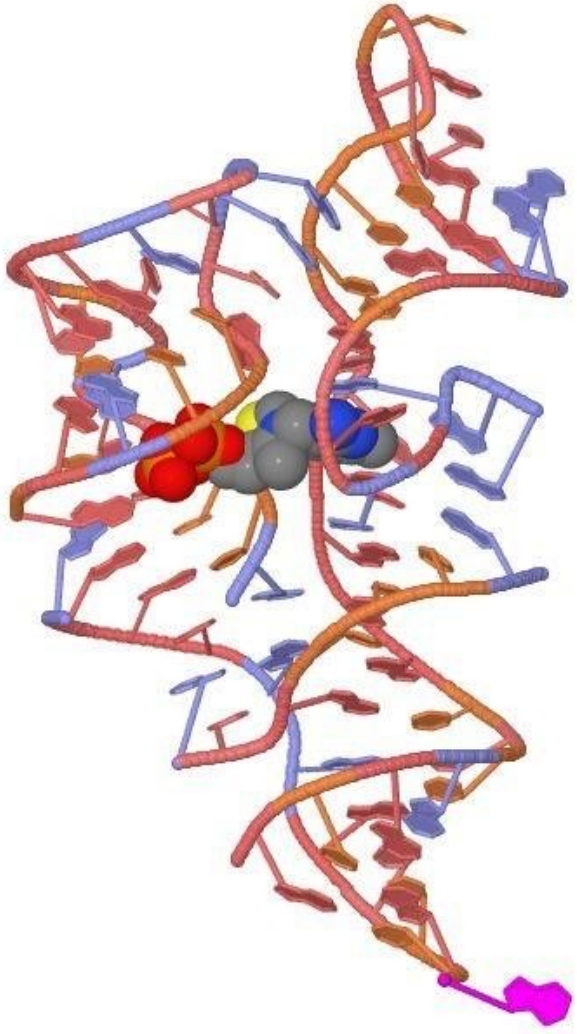
Зачем нам предсказывать вторичную структуру РНК?

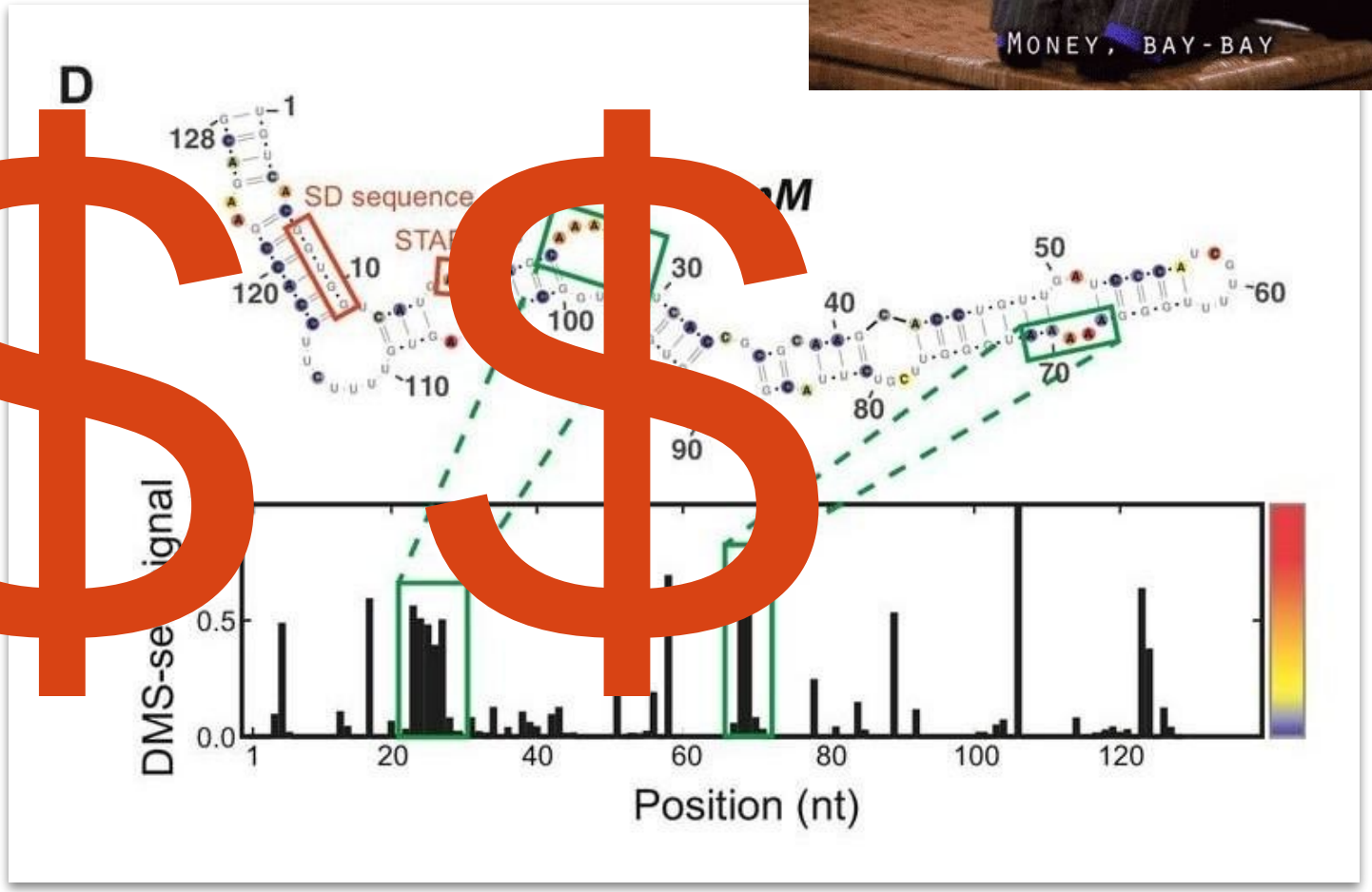
Experimental approaches





Experimental data

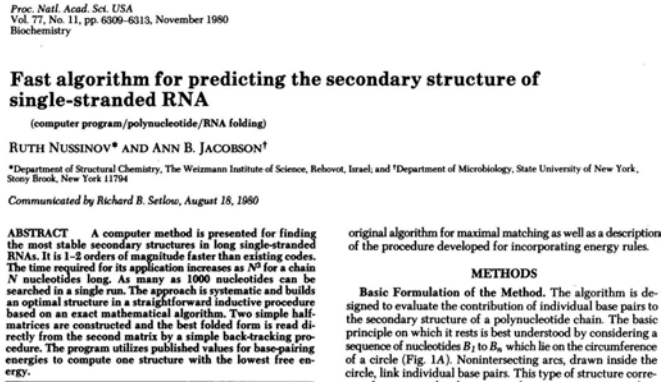




Как предсказывать структуру РНК?

Вычислительные подходы

Ruth Nussinov.
Developed dynamic programming approach for RNA SS prediction.



Proc. Natl. Acad. Sci. USA
Vol. 77, No. 11, pp. 6309–6313, November 1980
Biochemistry

Fast algorithm for predicting the secondary structure of single-stranded RNA

(computer program/polynucleotide/RNA folding)

RUTH NUSSINOV* AND ANN B. JACOBSON†

*Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot, Israel; and †Department of Microbiology, State University of New York, Stony Brook, New York 11794

Communicated by Richard B. Setlow, August 18, 1980

ABSTRACT A computer method is presented for finding the most stable secondary structures in long single-stranded RNAs. It is 1–2 orders of magnitude faster than existing codes. The time required for its application increases as N^3 for a chain N nucleotides long. As many as 1000 nucleotides can be searched in a single run. The approach is systematic and builds an optimal structure in a straightforward inductive procedure based on an exact mathematical algorithm. Two simple half-matrices are constructed and the best folded form is read directly from the second matrix by a simple back-tracking procedure. The program utilizes published values for base-pairing energies to compute one structure with the lowest free energy.

original algorithm for maximal matching as well as a description of the procedure developed for incorporating energy rules.

METHODS

Basic Formulation of the Method. The algorithm is designed to evaluate the contribution of individual base pairs to the secondary structure of a polynucleotide chain. The basic principle on which it rests is best understood by considering a sequence of nucleotides B_1 to B_n which lie on the circumference of a circle (Fig. 1*A*). Nonintersecting arcs, drawn inside the circle, link individual base pairs. This type of structure corre-

If you have only one sequence:

- Minimal free energy approach
- Machine (Deep) learning methods
- Comparison with covariation models

If you have several related sequences:

- Align then fold
- Align and fold
- Building/comparison with covariation models

How to choose an approach?

	Completely new class of structures (*De novo*)	**The new member of a known class**
We want to predict a structure for one particular object	MFE-based approach	Scan with known covariation models (etc.)
We want to describe a new class of structures	Build covariation model	

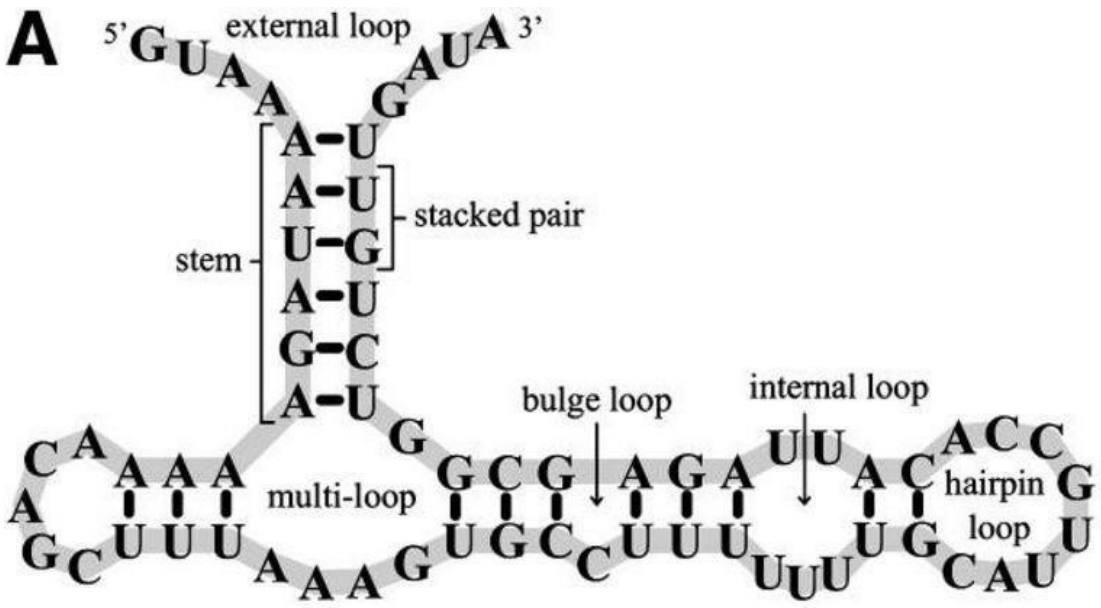
Minimum Free Energy approach

Assumption: The energy of each base pair is independent of all of the other pairs and the loop structure.

Consequence: Total free energy is the sum of all of the base pair free energies.

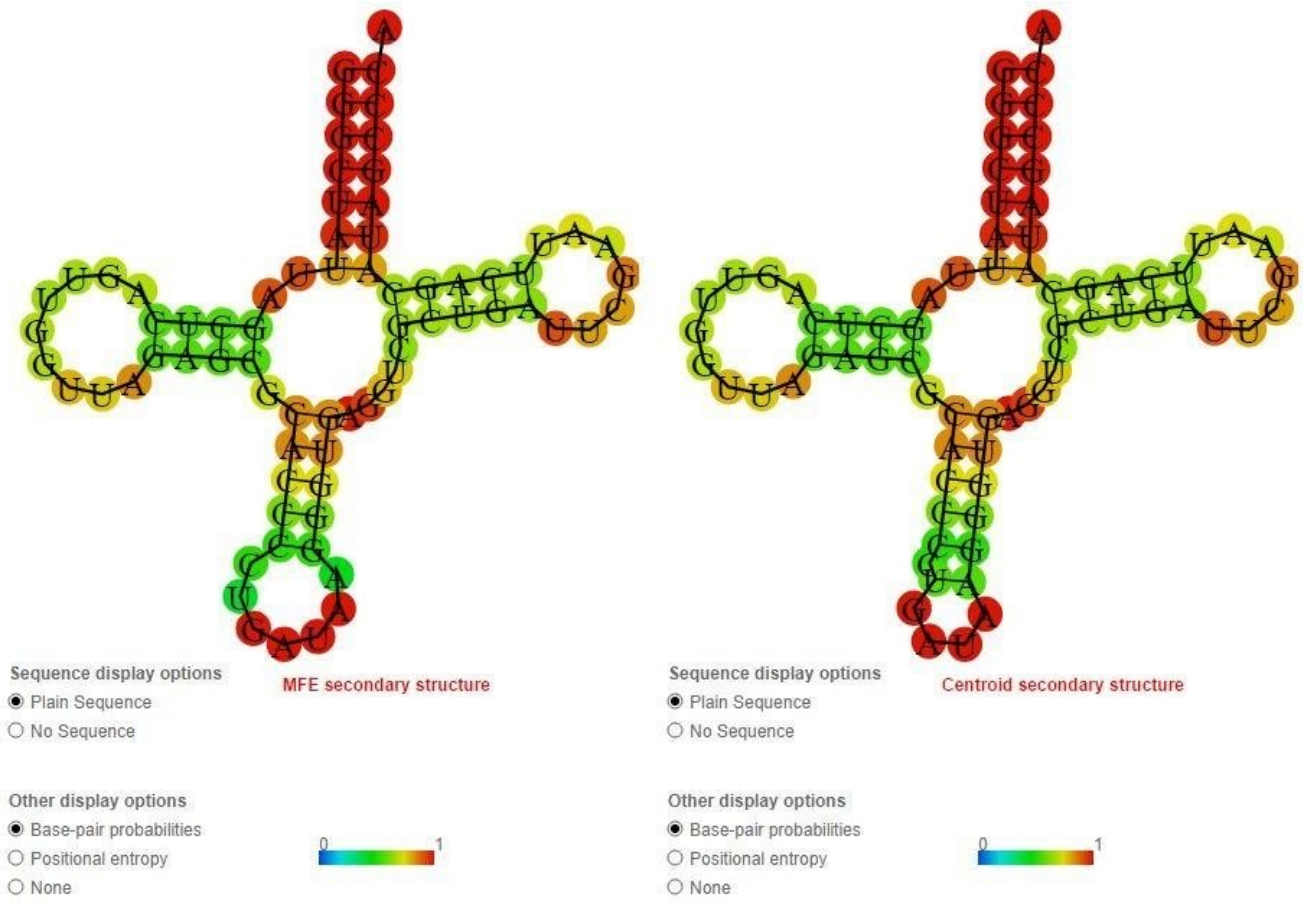
The size of free energy:

- is related to the type of base pairing
- affected by adjacent base pairs
- vary for the different structures (hair-loop, inner-loop, etc.)



RNAfold

http://nibiru.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

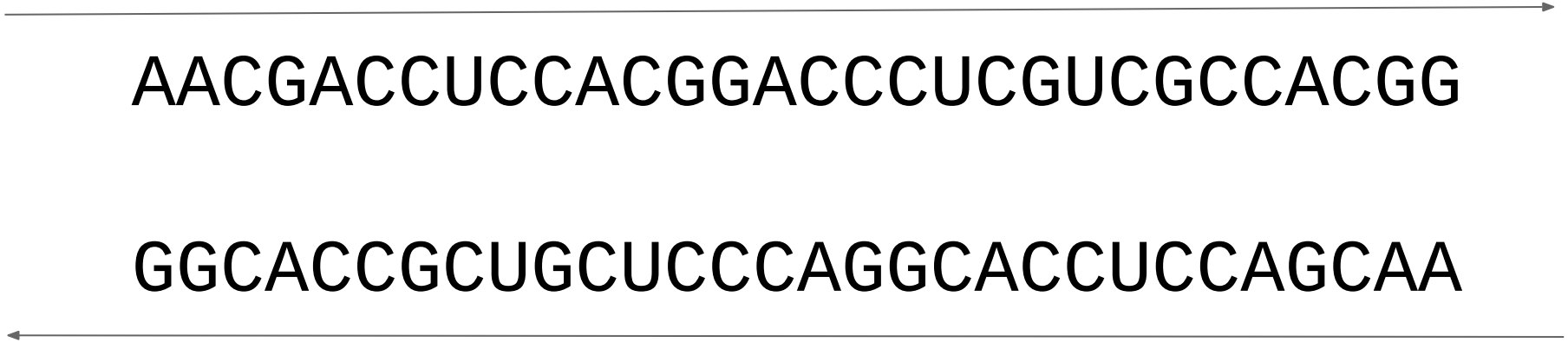


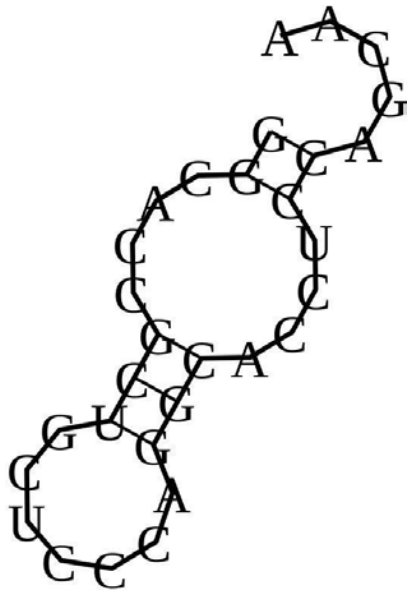
Task 1. Prediction with individual sequences.

Let's take 2 sequences from your dataset (any of them) and fold them individually with RNAfold:

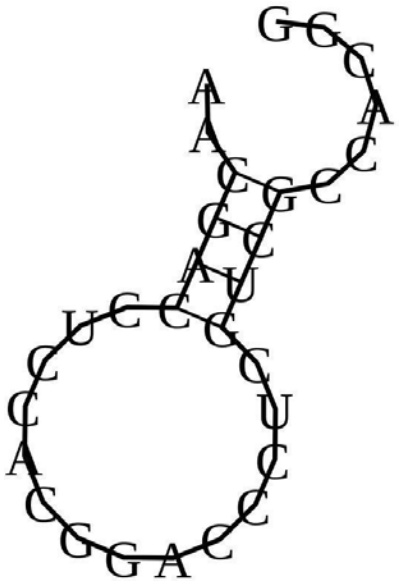
1) Copy 1 random sequence from your data (with ID).
2) Go to http://nibiru.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
3) Paste the sequence there
4) Press Process
5) After the computation is done, find secondary structure of MFE
6) Find a link to downloadable Vienna* record and click it. A window with the secondary structure will open.
7) Copy this secondary structure and save it (using any text editor). Name it using this sequence ID.
8) Go to the pictures of secondary structures. Find the options for downloading them and save both pictures in any convenient format (you will include them to the report).
9) Repeat all procedure with another randomly picked sequence.

Will an inverted sequence form the same structure with the same energy?





-3.10
kcal/mol



-2.30
kcal/mol

Concerns

- Suboptimal structures may also fold *in vivo*.
- Hard to distinguish folding of a single, functional RNA sequence from the folding of a random sequence with the same nucleotide content.

Let’s use evolutionary information

- **Molecules with similar functions and different nucleotide sequences will form similar structures.**

- **Positions that co-vary with one another are possible pairing partners.**

Task 2. Prediction from an alignment. (align&fold)

Create a structural alignment using any 20 (to make it faster) of the sequences from your dataset using LocARNA:

1) Go to https://rna.informatik.uni-freiburg.de/LocARNA/Input.jsp
2) Upload the file with your data (or part of it)
3) Wait — locaRNA is slow
4) After the job is done, download 3 files: Stockholm, FASTA and structure.

LocaRNA is slow. We can have a 10 minute break.

Let's compare RNAfold and locaRNA

Now for each sequence you've selected at step 1you have two predicted secondary structures: one from RNAfold, another one from the alignment of all sequences (consensus structure).

You can compare these two secondary structure predictions for any of your sequences from step 1using a tool R-Chie https://e-rna.org/r-chie/

Task 3. Compare RNA structures produced by two methods.

1) You have to match the dot-bracket formula from RNAfold with LocaRNA consensus structure by length. Do it using this colab code (the instructions are inside): https://colab.research.google.com/drive/1dQZQgV2Y63VPjoqrqPC1B8Y5xG_HfzPv?usp=sharing

2) Then go to https://e-rna.org/r-chie/ and click Create a Plot
3) Upload the corrected RNAfold structure to Secondary Structure Input
4) Click Add another structure to plot and upload consensus structure from locaRNA
5) Click Add nucleotide sequence(s) to plot and upload the alignment from locaRNA in FASTA format
6) Press Plot button
7) Save the produced figure in any convenient format

RNA sequence is often not conserved…

RNA
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | The RNA Society

RNA. 2009 Dec;15(12):2075–2082. doi: 10.1261/rna.1556009

The tedious task of finding homologous noncoding RNA genes

Peter Menzel [1,2], Jan Gorodkin [1], Peter F Stadler [2,3,4,5,6]

Author information Copyright and License information

PMCID: PMC2779685 PMID: 19861422

Abstract

…but RNA structure is still maintained

RNA
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | The RNA Society

RNA. 2009 Dec;15(12):2075–2082. doi: 10.1261/rna.1556009

The tedious task of finding homologous noncoding RNA genes

Peter Menzel [1,2], Jan Gorodkin [1], Peter F Stadler [2,3,4,5,6]

Author information · Copyright and License information

PMCID: PMC2779685 PMID: 19861422

Abstract

Comparative Study > Cell. 2000 Mar 3;100(5):503-14. doi: 10.1016/s0092-8674(00)80687-x.

Secondary structure of vertebrate telomerase RNA

J L Chen [1], M A Blasco, C W Greider

Affiliations + expand

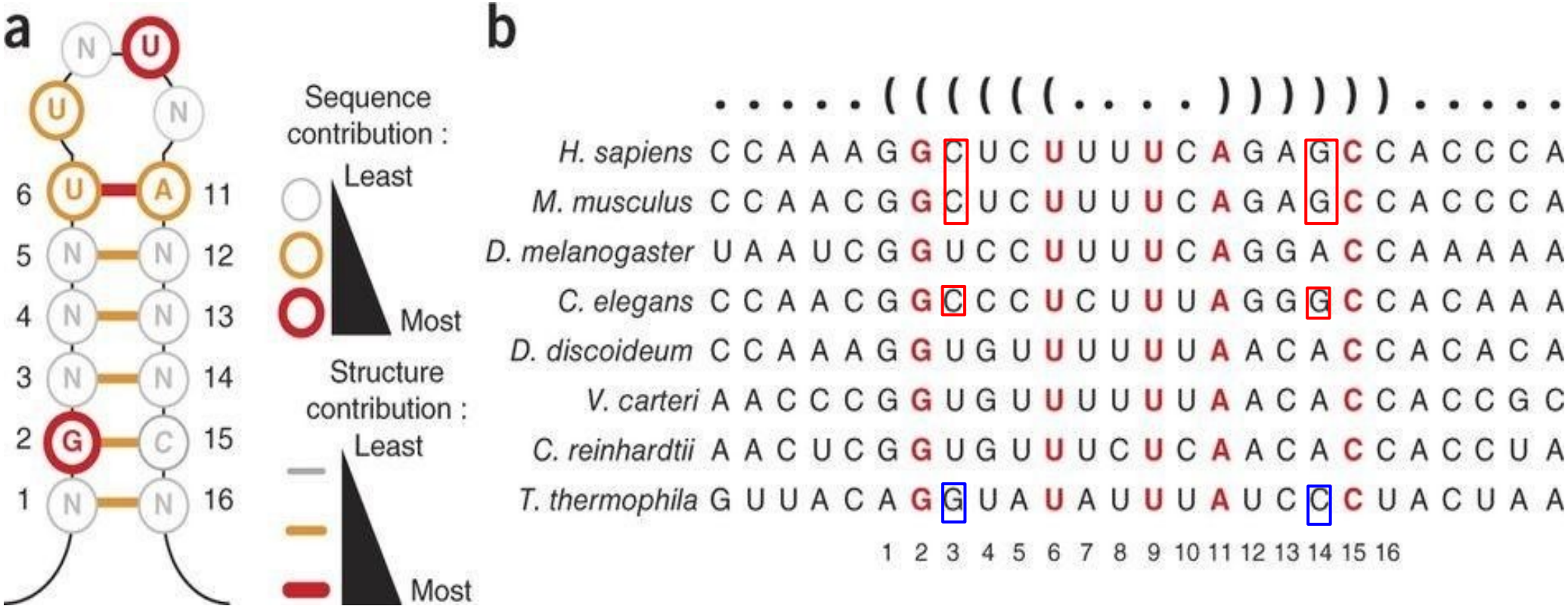
PMID: 10721988 DOI: 10.1016/s0092-8674(00)80687-x

Free article

Abstract

Telomerase is a ribonucleoprotein enzyme that maintains telomere length by adding telomeric sequence repeats onto chromosome ends. The essential RNA component of telomerase provides the template for telomeric repeat synthesis. To determine the secondary structure of vertebrate telomerase RNA, 32 new telomerase RNA genes were cloned and sequenced from a variety of vertebrate species including 18 mammals, 2 birds, 1 reptile, 7 amphibians, and 4 fishes. Using phylogenetic comparative analysis, we propose a secondary structure that contains four structural domains conserved in all vertebrates. Ten helical regions of the RNA are universally conserved while other regions vary significantly in length and sequence between different classes of vertebrates. The proposed vertebrate telomerase RNA structure displays a strikingly similar topology to the previously determined ciliate telomerase RNA structure, implying an evolutionary conservation of the global architecture of telomerase RNA.

Covariation of important pairs maintain the structure



We can build covariation models from the data

LocaRNA alignment

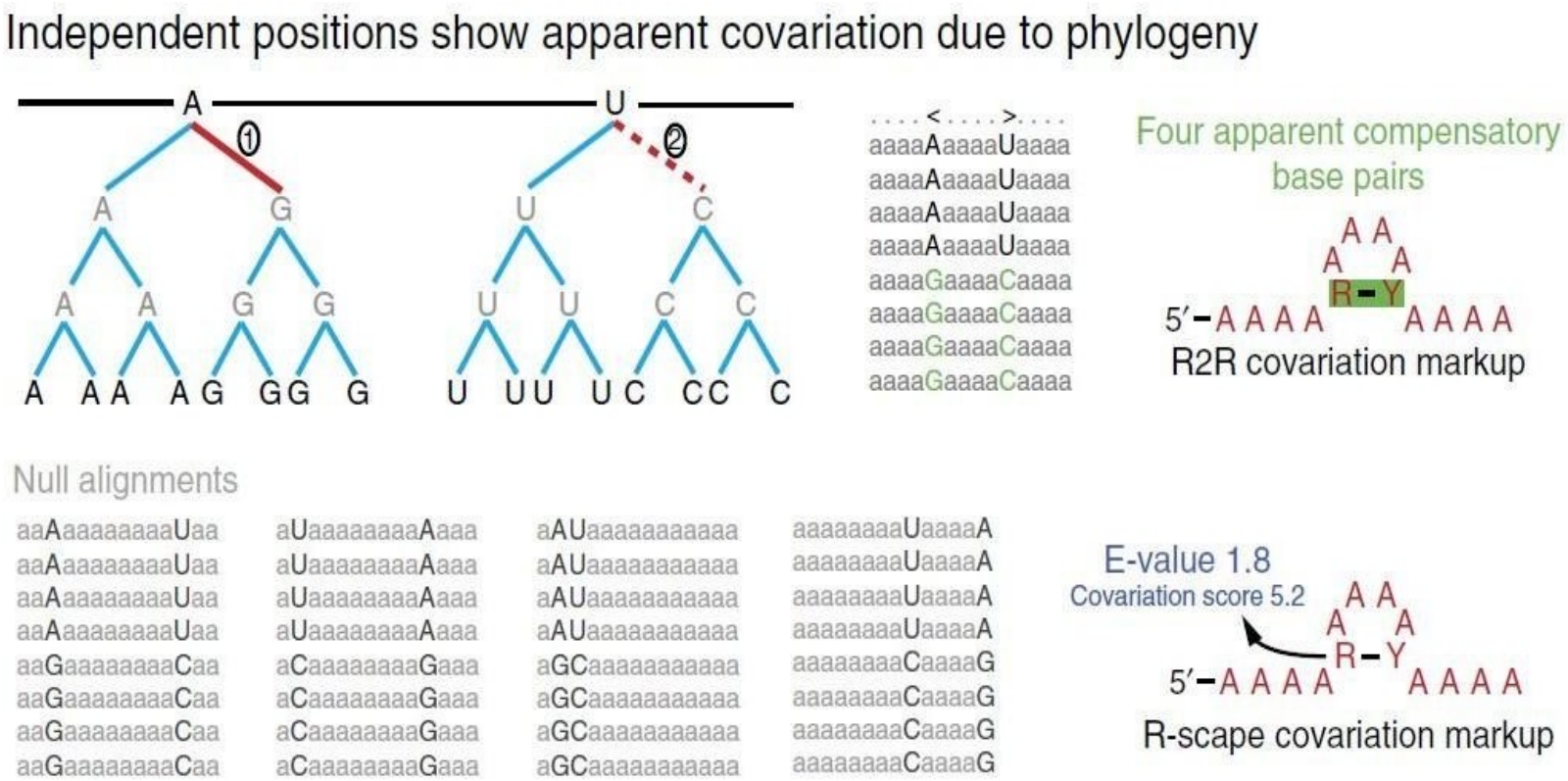
⬇

Infernal tool

http://eddylab.org/infernal/

Infernal can build a covariance model (we will not do it today; just FYI)

Bad news: covariation may not indicate structure conservation



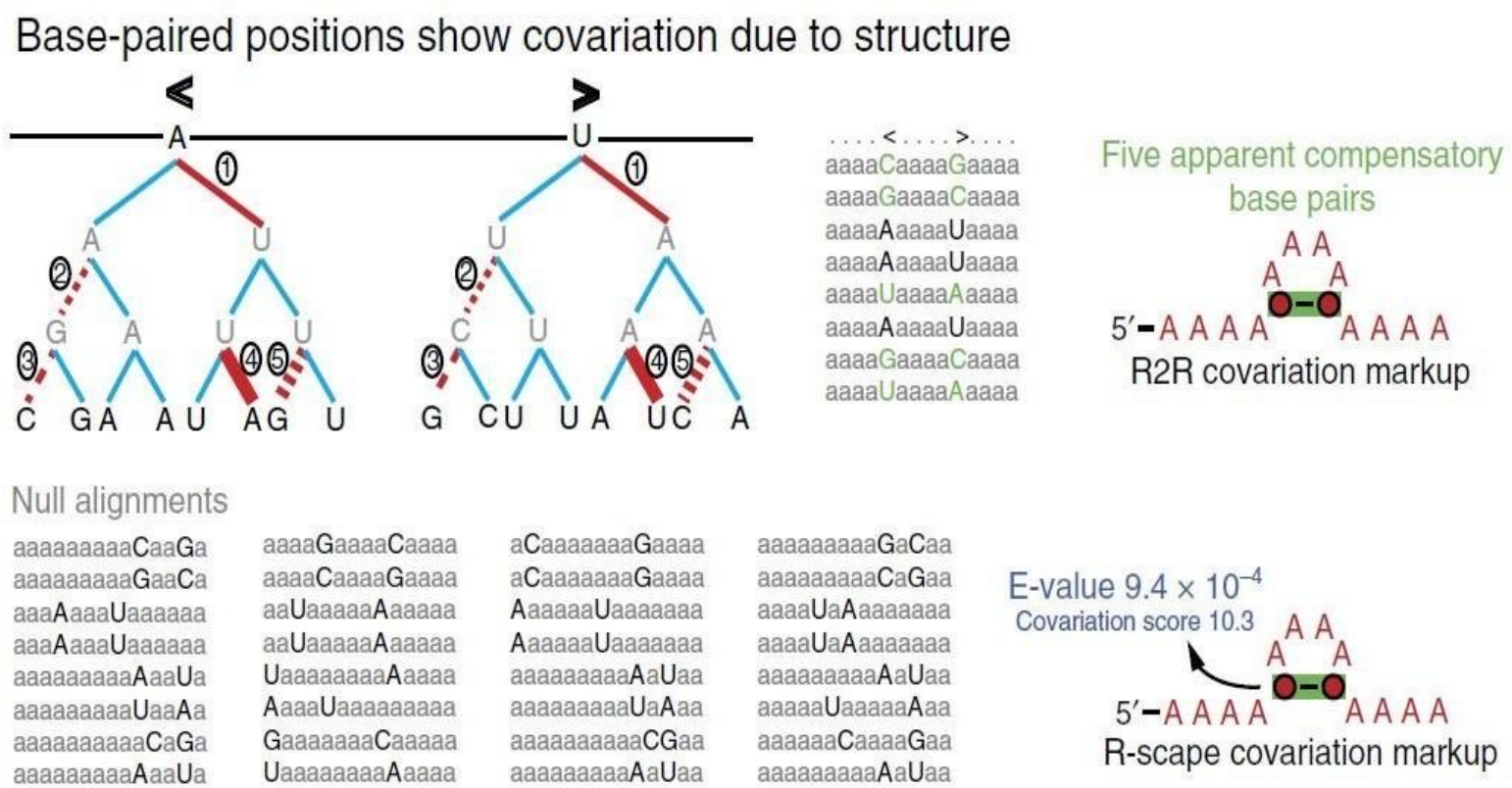
An E-value E(x) is the number of column pairs expected to give a covariation score of at least x when they are evolving independently, under no RNA structure constraint.

Toy alignment (top left) with two independent substitutions (marked 1,2) on the same branch, resulting in an apparent pairwise covariation annotated by R2R (top right). R-scape simulated null alignments (bottom left) retain this confounding covariation signal, and it is judged insignificant (bottom right).

Rivas et al., Nature Methods 2016

Good news: we know how to deal with it

http://eddylab.org/R-scape/



An E-value E(x) is the number of column pairs expected to give a covariation score of at least x when they are evolving independently, under no RNA structure constraint.

Toy alignment with five compensatory base pair substitutions (marked 1–5) showing a covariation pattern that is destroyed in the R-scape simulated null alignments, and thus judged significant.

Rivas et al., Nature Methods 2016

Task 4. Evaluation of structure

http://eddylab.org/R-scape/

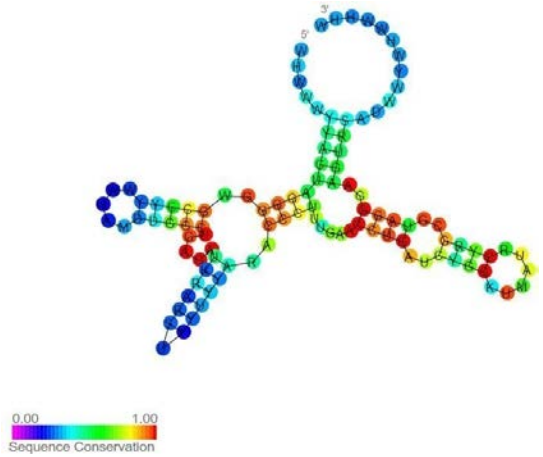
Use R-scape to evaluate the covariation in the alignment produced by locaRNA

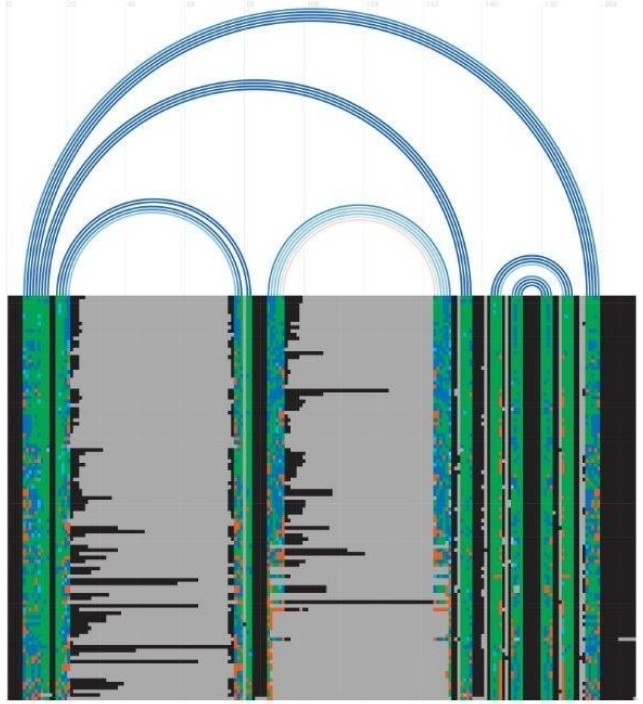
1) **Go to R-scape**
2) **Upload your alignment in .stk format**
3) **Submit the job**

Rfam database is built using Infernal

http://eddylab.org/R-scape/

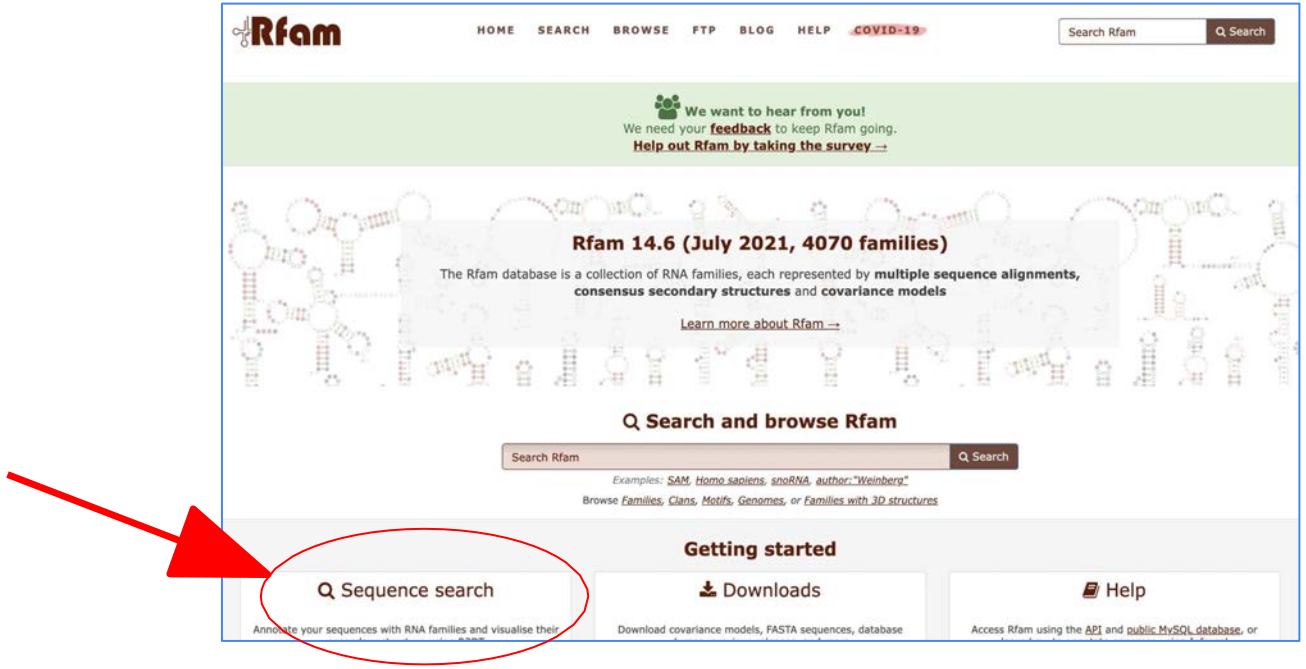
http://rfam.xfam.org





Task 5. Search for the known structure by sequence.

Take the sequence with the best MFE structure and perform sequence search in Rfam.

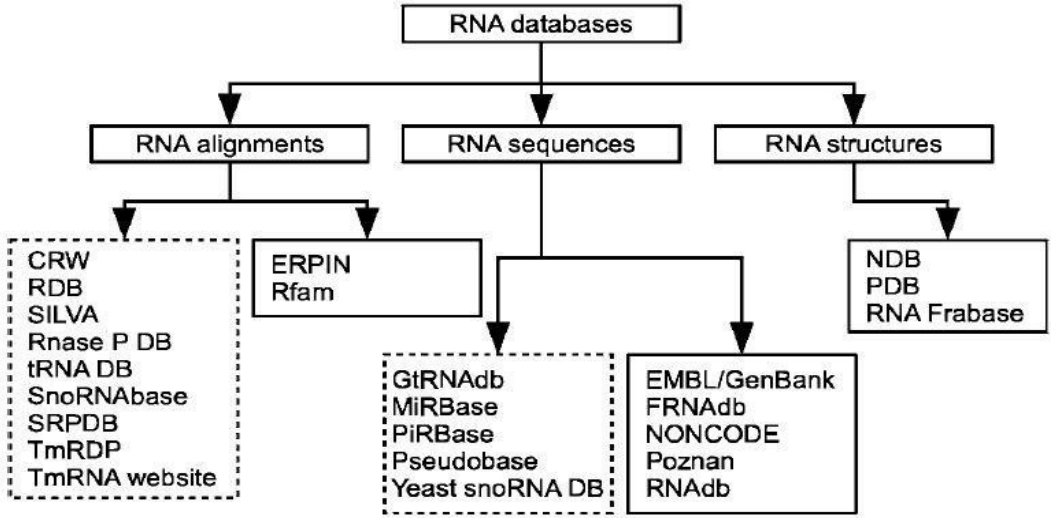


Task 6. Command line practice

1) Take Stockholm file (the alignment produced by locaRNA)
2) In command line prompt, type a command that will show you all lines starting with the same ID as you used in Task 1 (any of the two).
3) Make a screenshot and include in your report.
4) Then, type a command that count number of lines NOT containing this ID.
5) Make a screenshot and include it in your report.
6) Bonus task (0.5 compensatory points): write a command that can make a FASTA record from the output from (2).

Data sources

- http://rfam.xfam.org
- RNAcentral - a non-coding RNA sequences database aggregating data from 40 databases: https://rnacentral.org/



- Nucleic Acid Database - an analog of PDB: http://ndbserver.rutgers.edu/
- DSSR - a tool for annotation of RNA tertiary structure: http://wdssr.x3dna.org/

Report structure

Task 1. Include 2 sequences you picked, their predicted secondary structures in dot-bracket format, and pictures of their MFE and centroid structures. Compare the pictures (MFE and centroid for each sequence, and between them). Can these sequences have a conserved secondary structure? Support your opinion using the comparison of the pictures. (2 points)

Task 2. Analyse the alignment in locaRNA visually. Do you observe a good sequence conservation? Can this dataset contain sequences with a conserved secondary structure? You may add a screenshot of your alignment. (2 points)

Task 3. Include the figure produced by R-Chie. Do structures predicted by different methods coincide? Describe the main differences/similarities. Do these observations support the hypothesis that our data contain RNAs with conserved secondary structure? (2 points)

Task 4. Include the table and the picture of a secondary structure produced by R-Scape. Analyse them. Does R-Scape support the hypothesis that our dataset contain sequences with a conserved secondary structure? Support your opinion using the analysis of R-Scape output. (2 points)

Task 5. Include Rfam ID of the 1st found record and a screenshot of its secondary structure. Is it similar to your predictions? (1 point)

Task 6. Complete the task and include the required screenshots. (1 point)