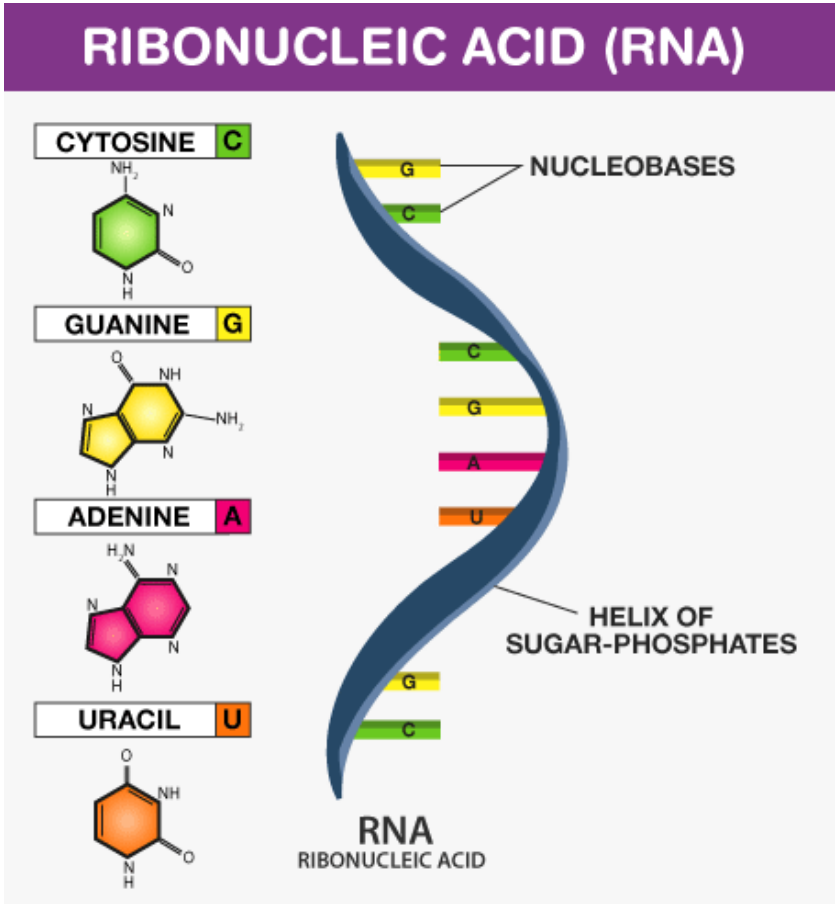
Predviđanje sekundarne strukture ribonukleinskih kiselina

Lazar Vasović, 2006/2021
prof. dr Natalija Polović
Biohemija za informatičare
Seminarski rad, 15. jun 2022.

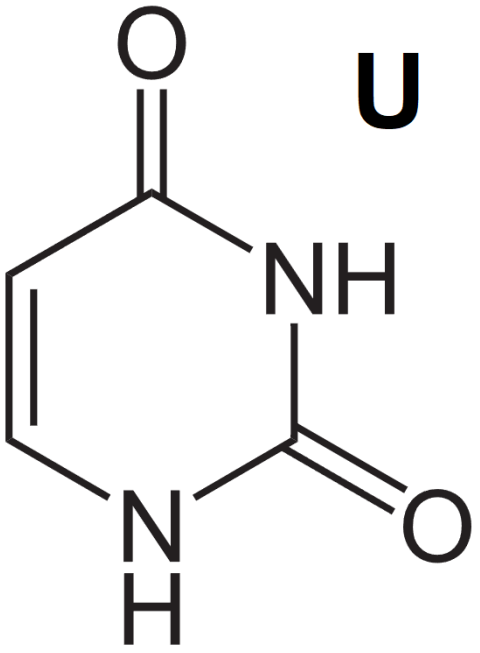
1. RNK I NIVOI STRUKTURE

Primarna struktura

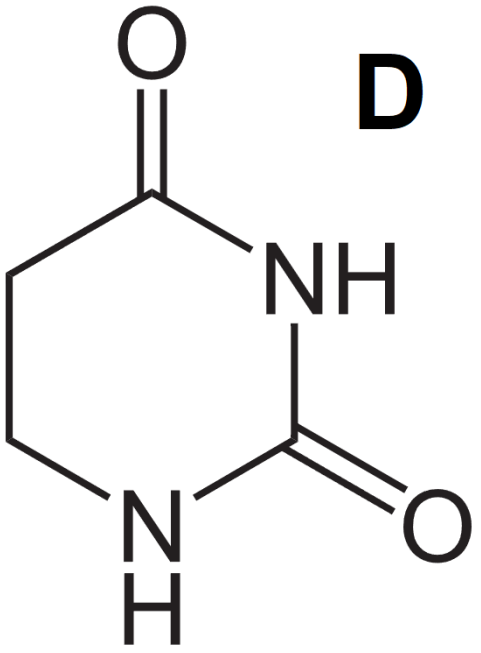
- RNK – lanac (polimer) ribonukleotida
- Ribonukleotid – baza + riboza + fosfat
- Baze – A, G, C, U
- Tip podataka – niska



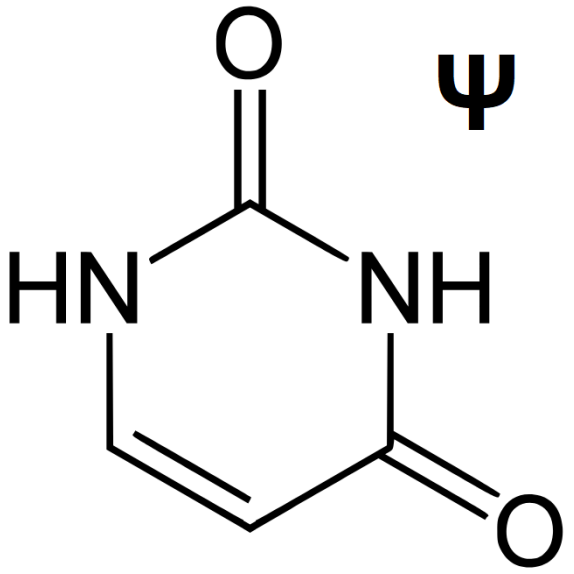
Modifikovane baze



U



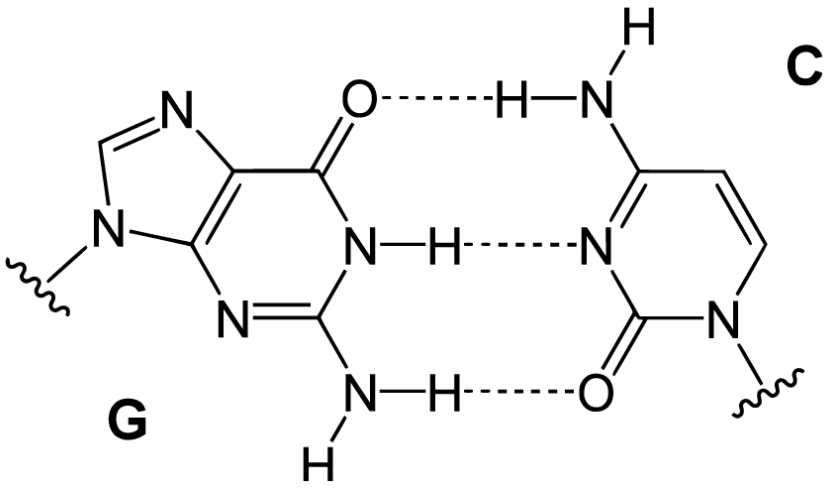
D

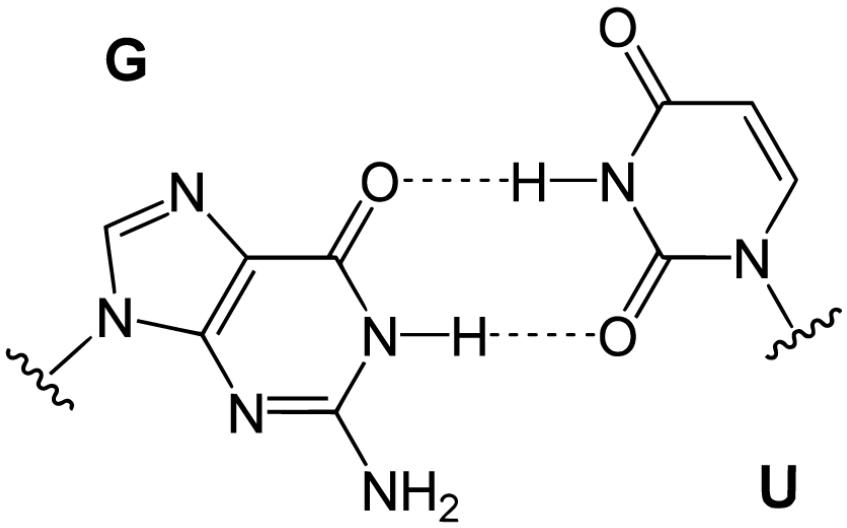


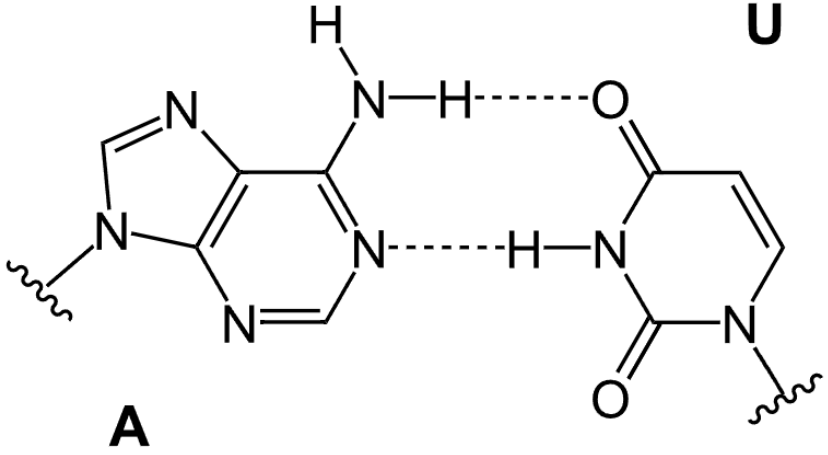
Ψ

- Uracil, zasićeni dihidrouracil, rotirani pseudouracil

Bazni parovi

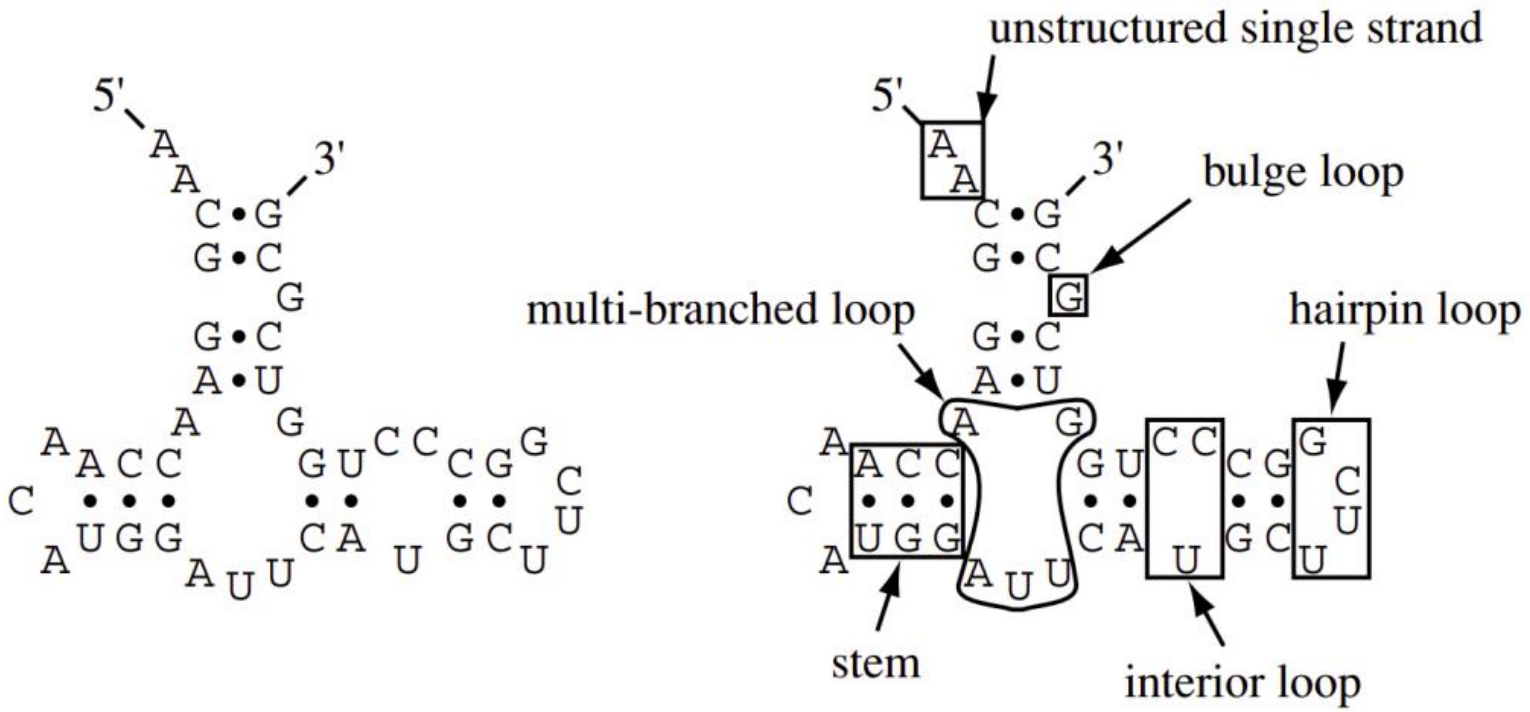






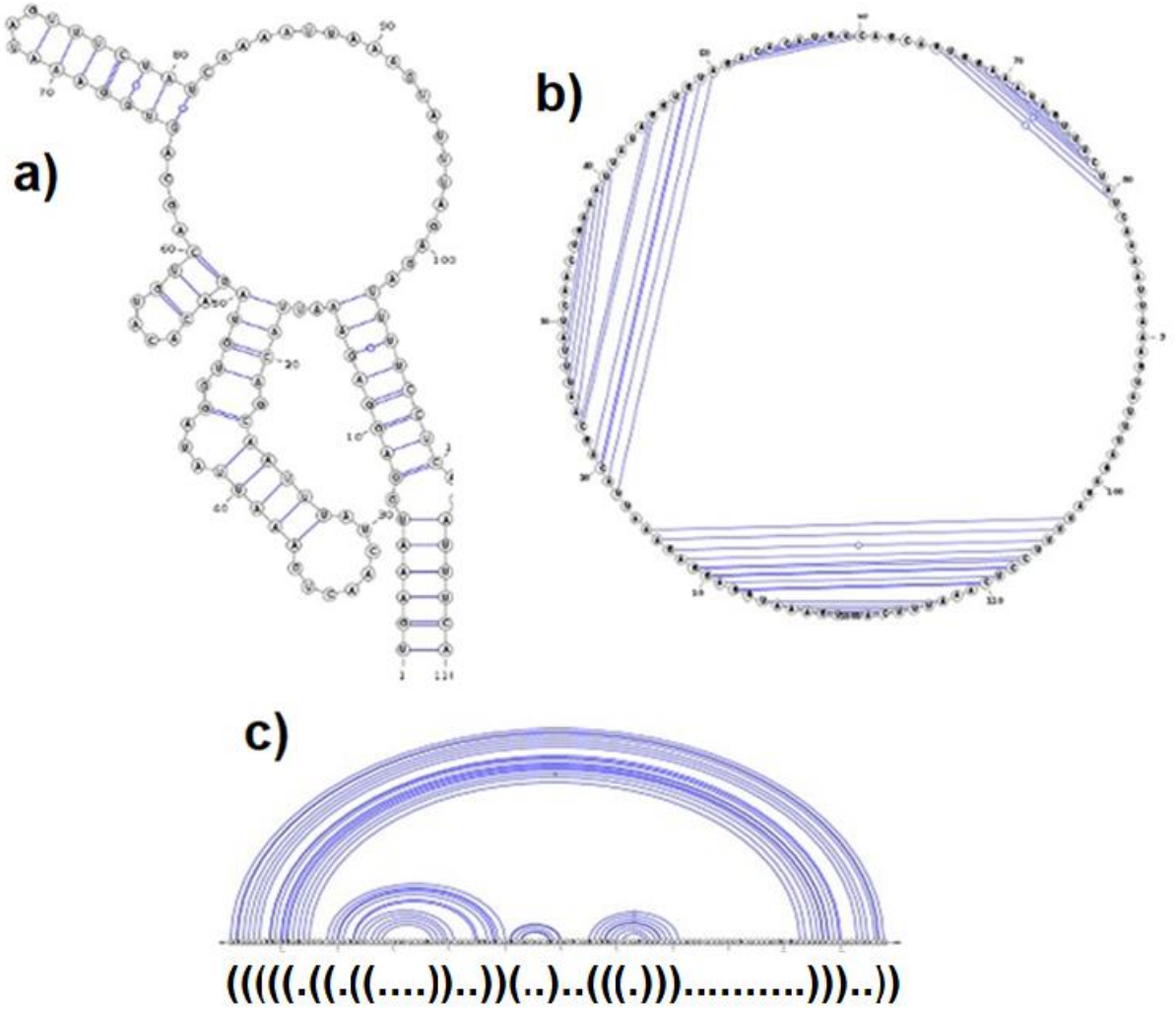
- Uglavnom W–C
- Nekanonski parovi

Sekundarna struktura



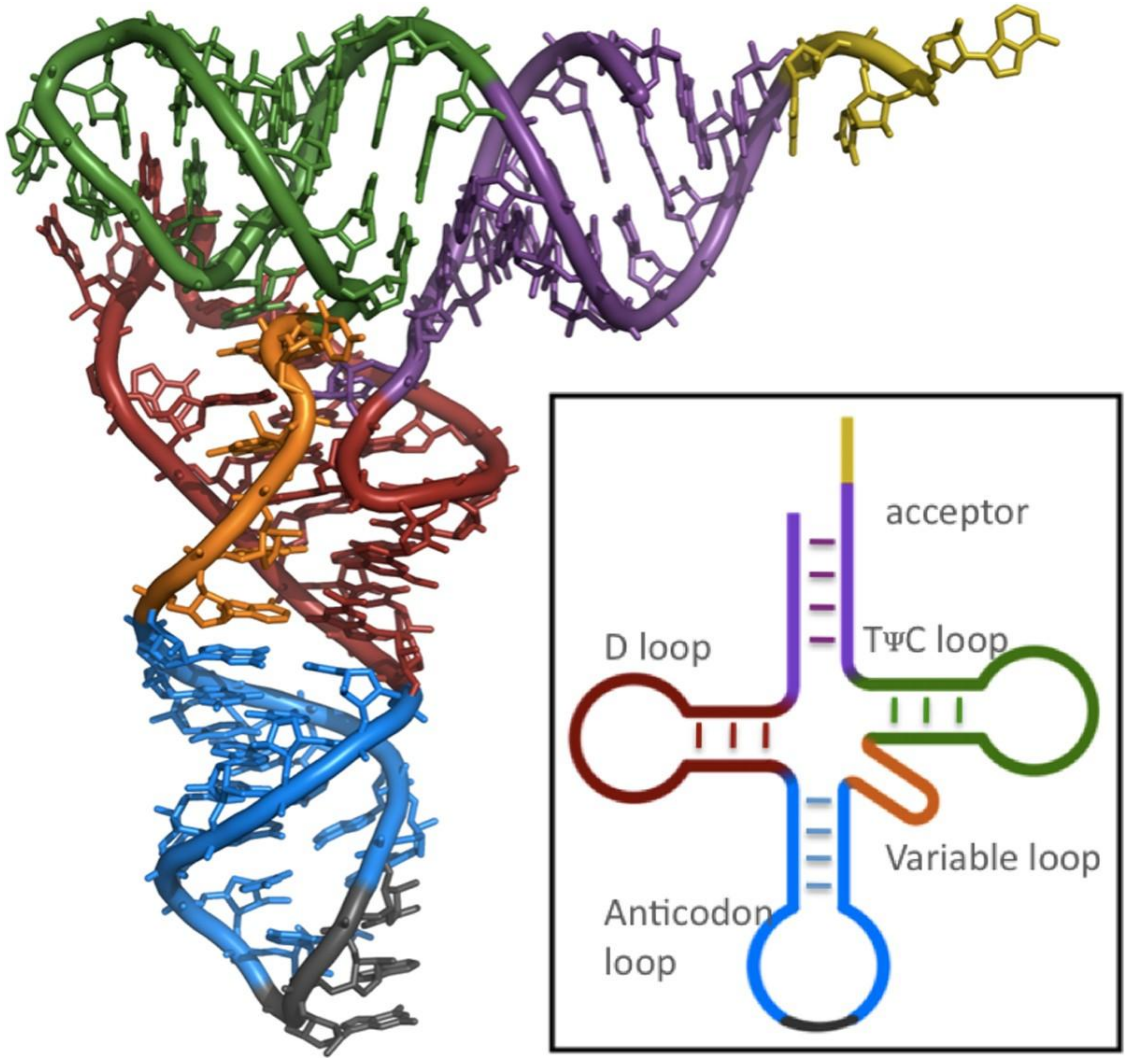
- Intramolekularno uvijanje i formiranje segmenata uparenih ribonukleotida – po pravilu A-heliks

Dodatni načini prikaza (lukovi)

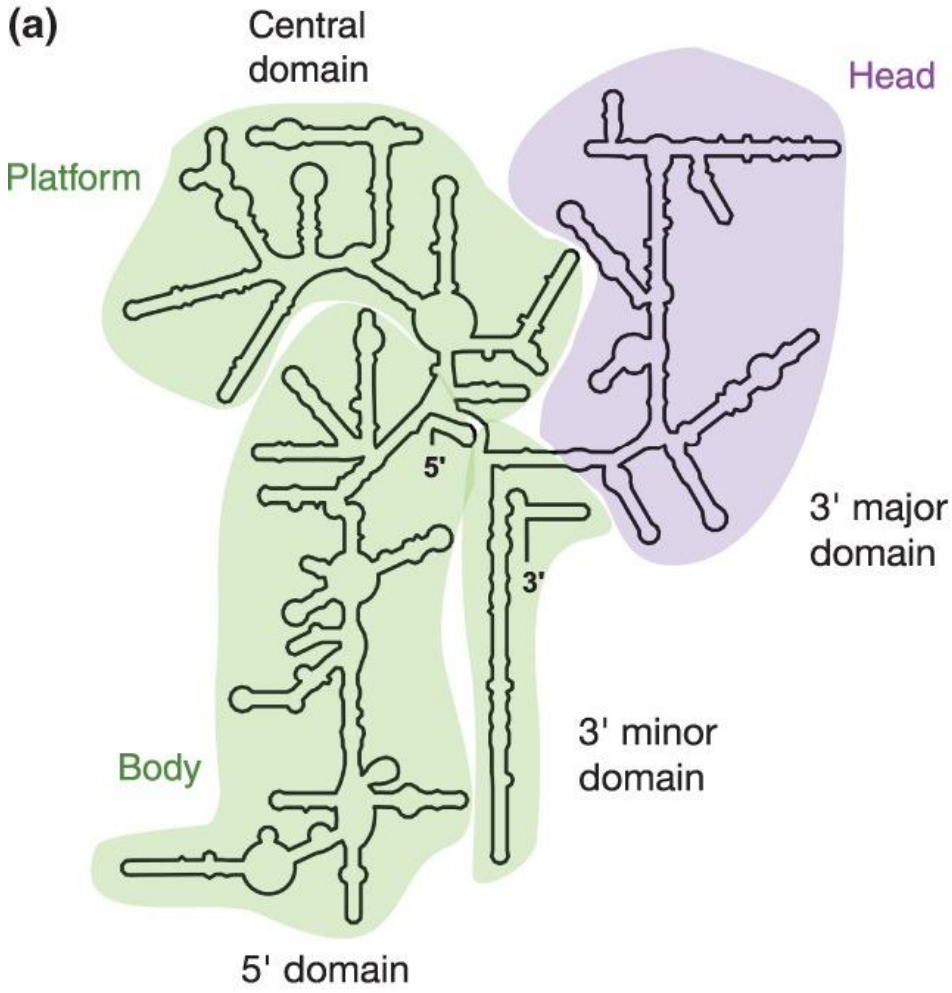


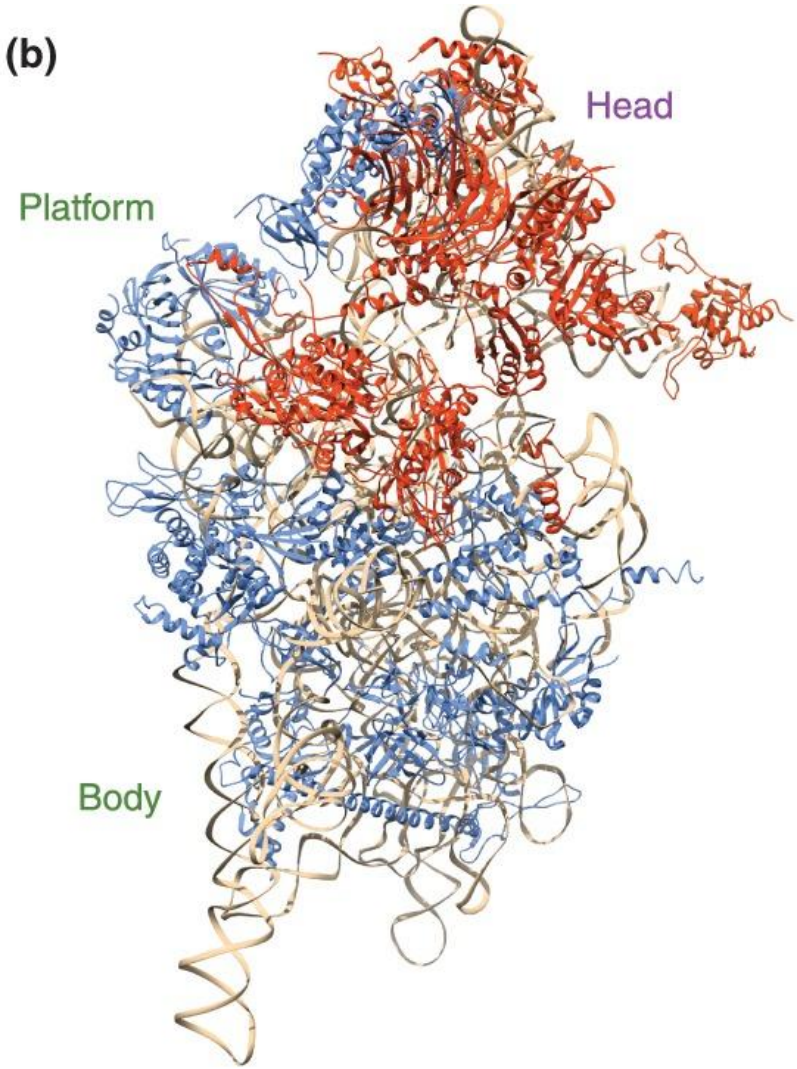
2. PRIMERI RAZLIČITIH RNK

Transportna RNK (detelina)

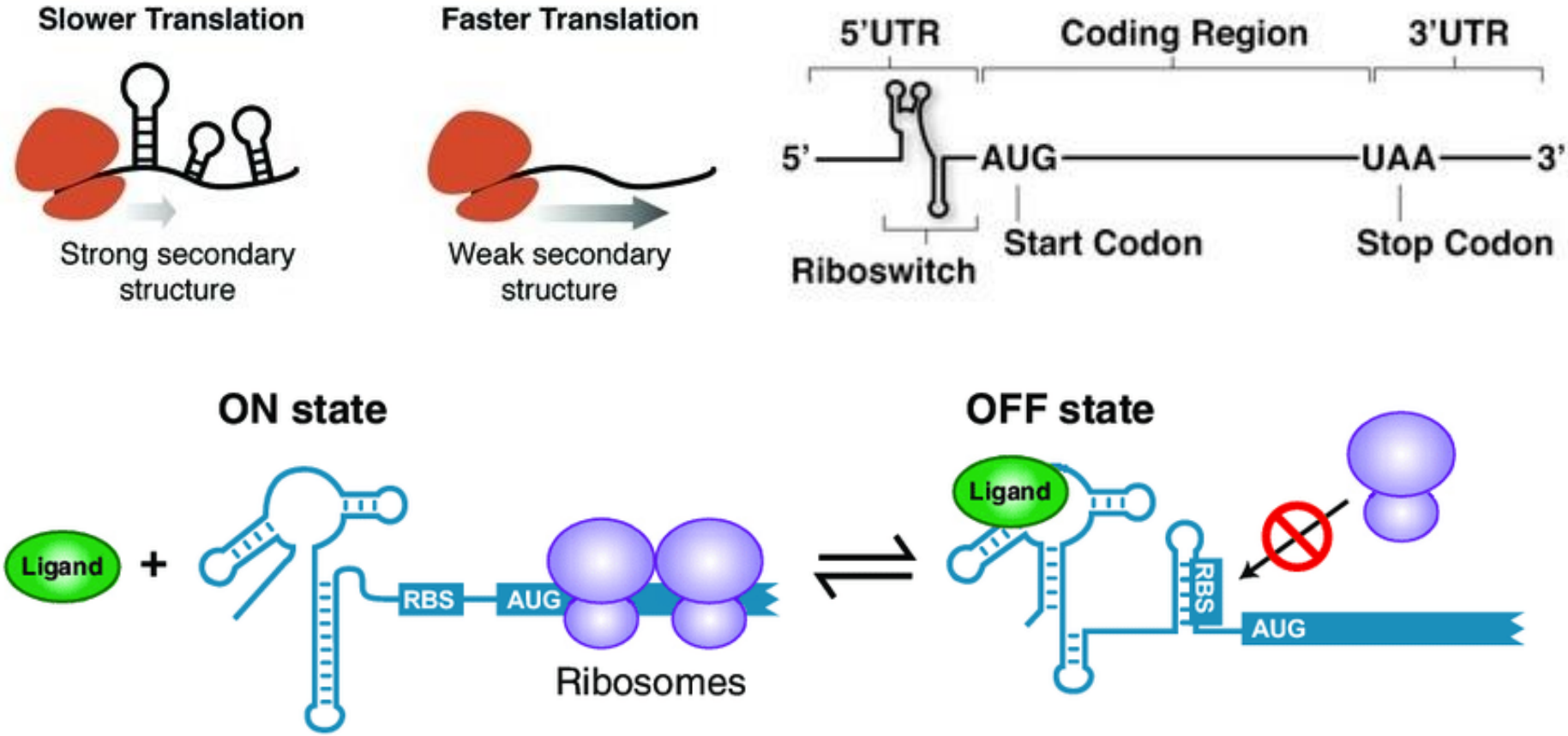


Ribozomska RNK (18S)



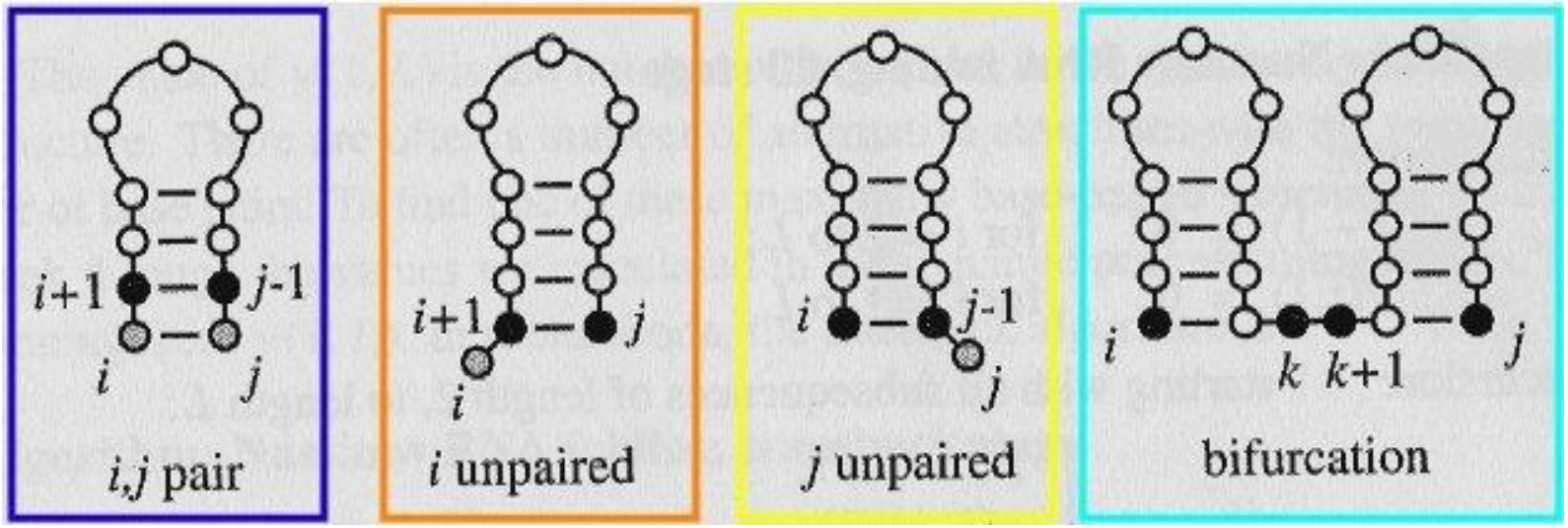


Informaciona RNK (riboprekidač)



3. OBRADA POJEDINAČNIH SEKVENCI

Maksimizacija uparivanja

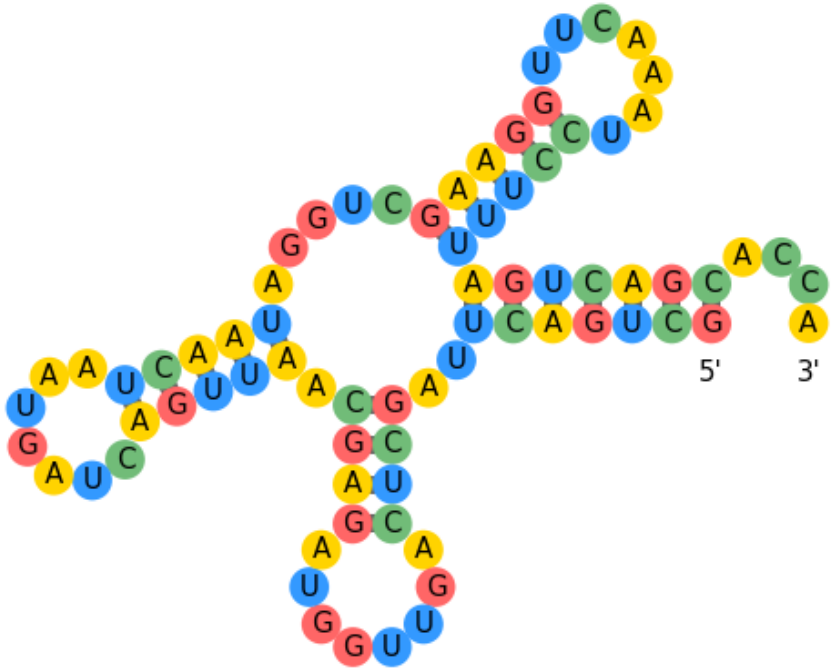


- Struktura sa najvećim brojem baznih parova, eventualno uz varijabilnu cenu uparivanja
- Važne strukturne karakteristike nisu uzete u obzir – preferencije ka određenim dužinama petlji ili kombinacijama susednih parova u zavojnici

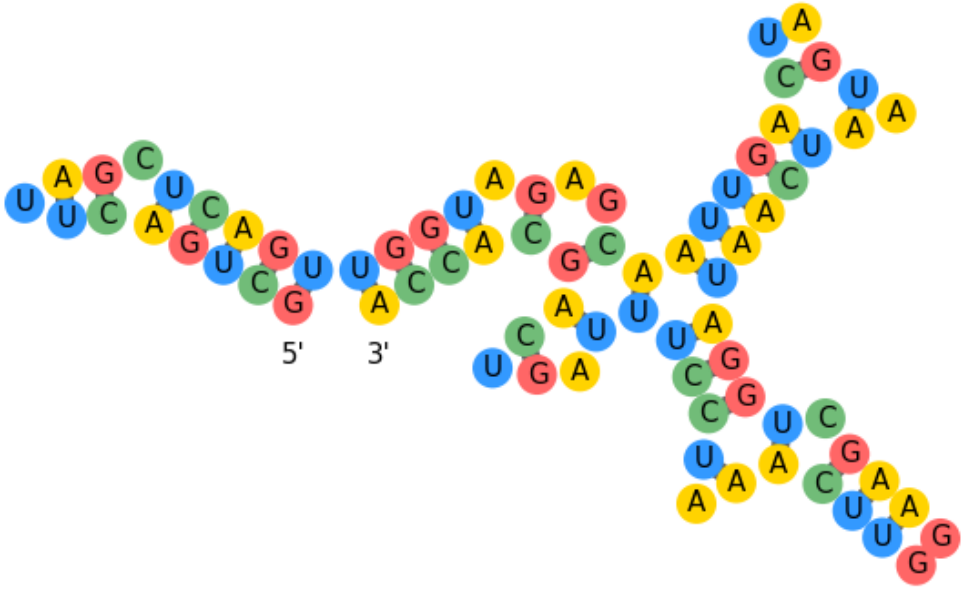
Maksimizacija nema smisla

>tdbR00000433|Mycoplasma_capricolum|2095|Thr|AGU
GCUGACUUAGCUCAGUUGGUAGAGCAAUUGACUAGUAAUCAAUAGGUCGAAGGUUCAAAUCCUUUAGUCAGCACCA

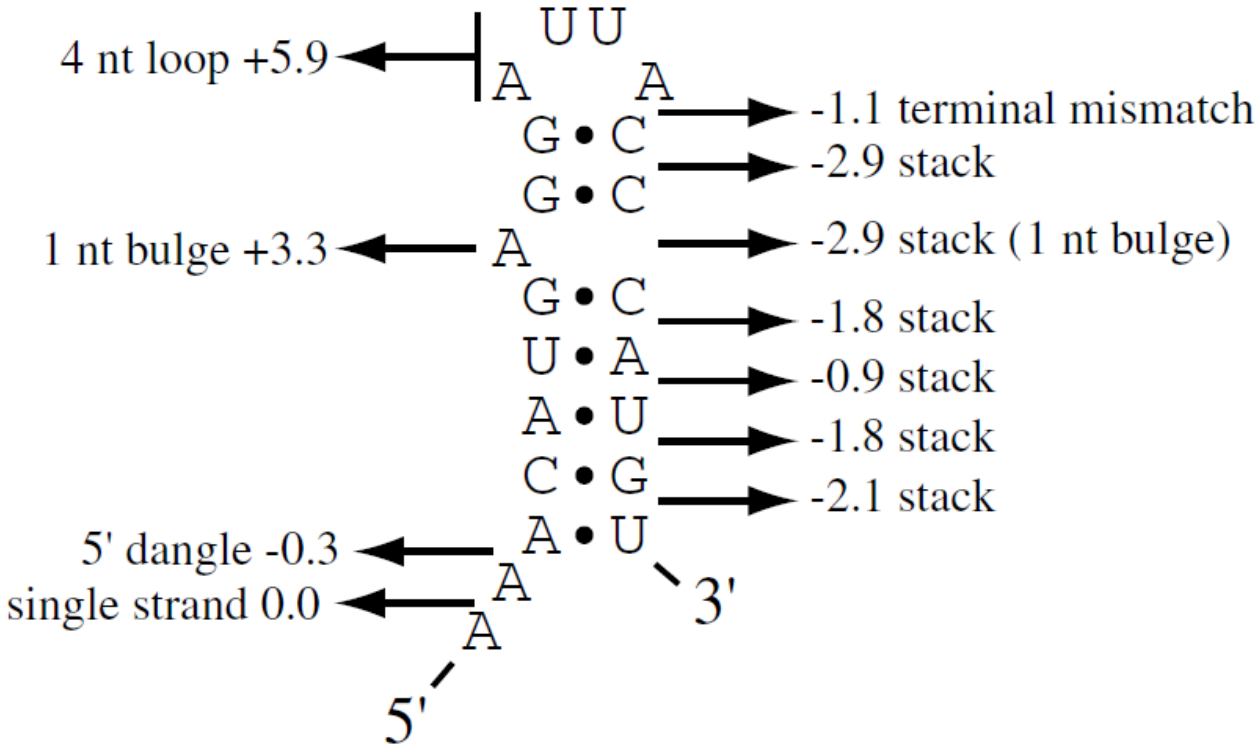
Stvarna struktura
(((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))....



Nussinov struktura
(((((((.)).)))))(((((.(..(((((((..)(.)))))))(((((.(((..))))(.)))))(.(.)))))))))



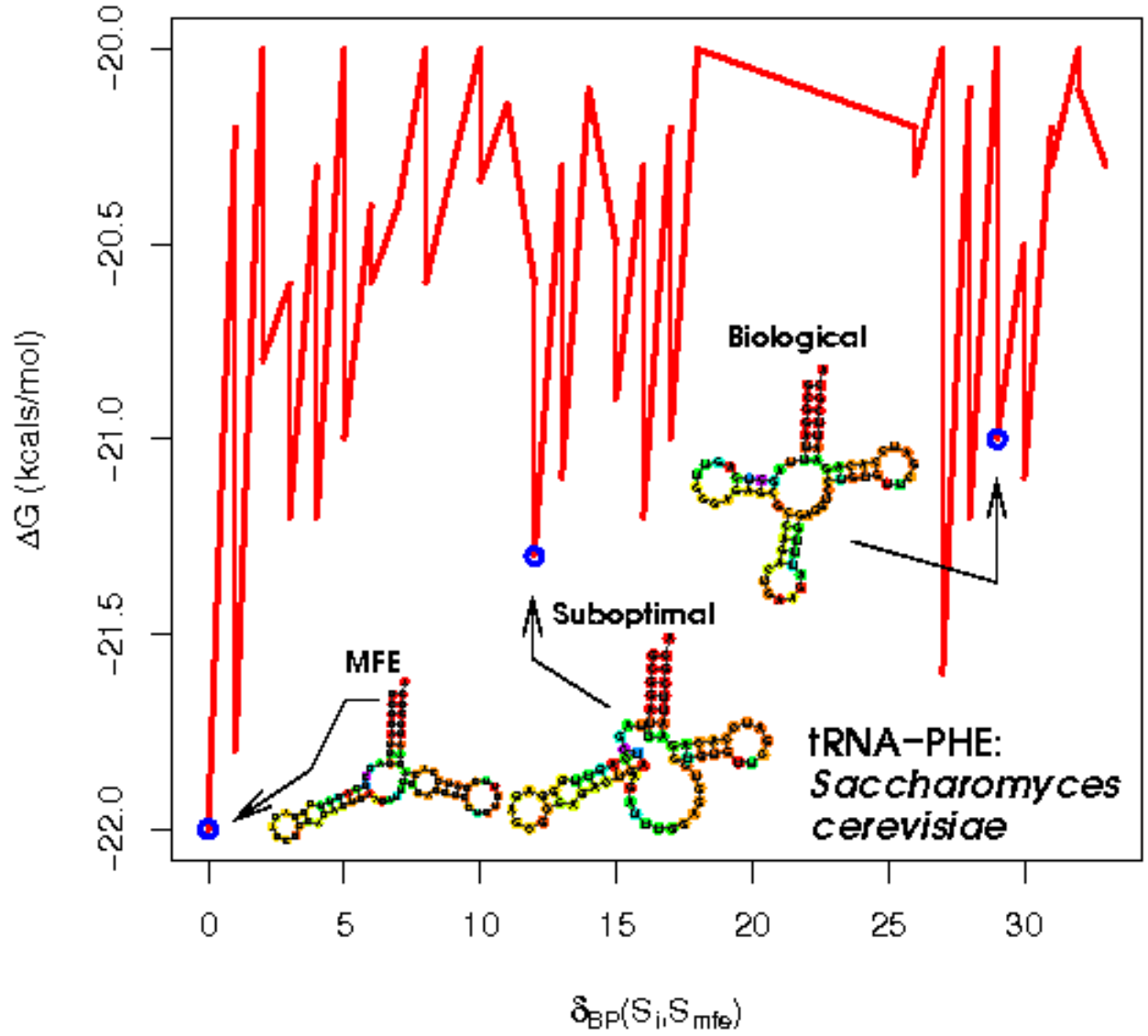
Zukerov termodinamički model



overall ΔG = -4.6 kcal/mol

- Uvijanje diktiraju biofizički procesi – slepljivanje baza (jake steking interakcije), min. slob. energija (MFE)

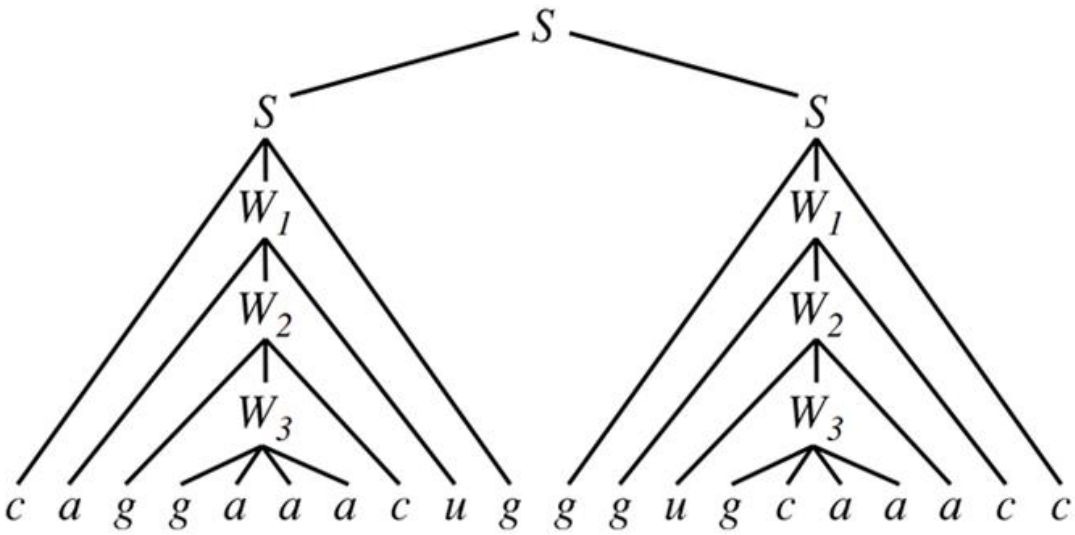
Zuker je sofisticiran, ali nije savršen



Stohastičke ili probabilističke kontekstno slobodne gramatike

- Statistički model sekvenci sa ugnežđenim zavisnostima – uklapa se u strukturu RNK
- Generativni model – sekvenca se izvodi na osnovu pravila

$$
\begin{aligned}
S &\rightarrow SS \\
S &\rightarrow aW_1u \mid cW_1g \mid gW_1c \mid uW_1a \\
W_1 &\rightarrow aW_2u \mid cW_2g \mid gW_2c \mid uW_2a \\
W_2 &\rightarrow aW_3u \mid cW_3g \mid gW_3c \mid uW_3a \\
W_3 &\rightarrow gaaa \mid gcaa
\end{aligned}
$$



Mogućnosti statističkih modela

- Obučavanje – nadgledano (na osnovu obeleženih struktura, prosto prebrojavanje) ili nenadgledano (algoritam iznutra-spolja)
- Predviđanje strukture – određivanje izvođenja najveće verovatnoće (Viterbi, tj. CYK)
- Evaluacija – određivanje verovatnoće niske (RNK) u jeziku koji generiše data gramatika (algoritmi iznutra/inside i spolja/outside)

Primena statističkih modela

- Moćni – mogu da simuliraju i maksimizaciju baznih parova, ali i termodinamičke modele
- Jednostavne gramatike mogu biti vrlo uspešne

$$S \rightarrow LS \text{ (nizanje elemenata)} \mid L \text{ (poslednji element)}$$

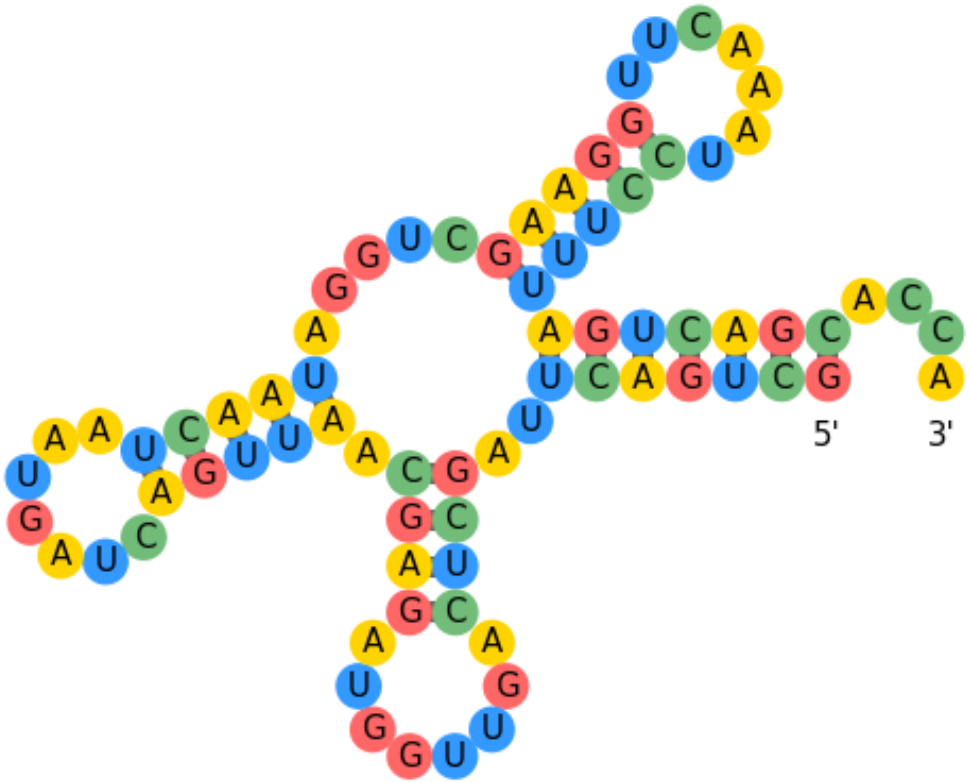
$$L \rightarrow s \text{ (neuparena baza)} \mid dFd \text{ (početak zavojnice)}$$

$$F \rightarrow dFd \text{ (nastavak zavojnice)} \mid LS \text{ (unutrašnjost zavojnice)}$$

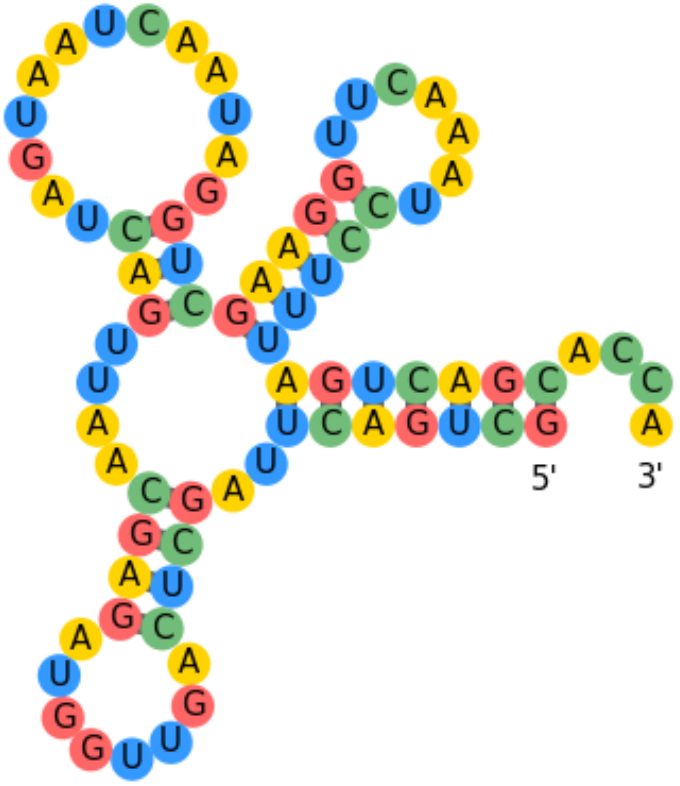
Primer omašene, ali bliske strukture

>tdbR00000433|Mycoplasma_capricolum|2095|Thr|AGU
GCUGACUUAGCUCAGUUGGUAGAGCAAUUGACUAGUAAUCAAUAGGUCGAAGGUUCAAAUCCUUUAGUCAGCACCA

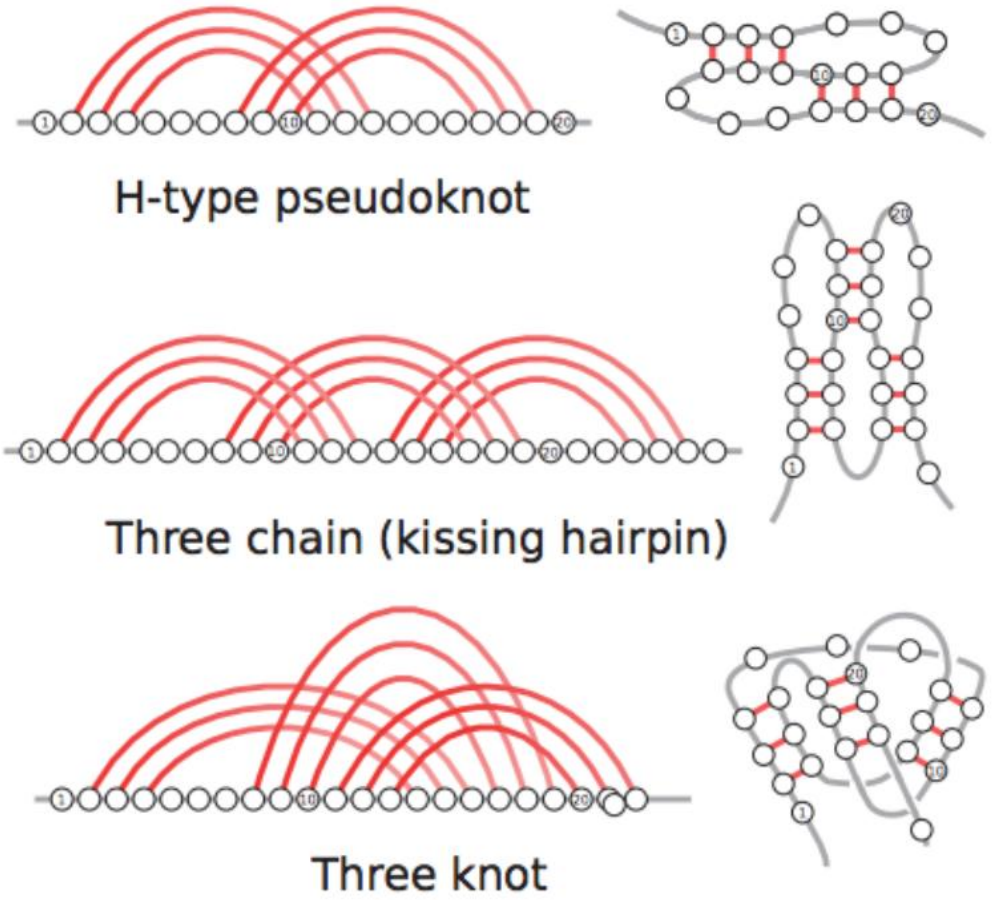
Stvarna struktura
(((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))....



KH-99 struktura
(((((((..((((........)))).....(((.............)))(((((.......)))))))))))....

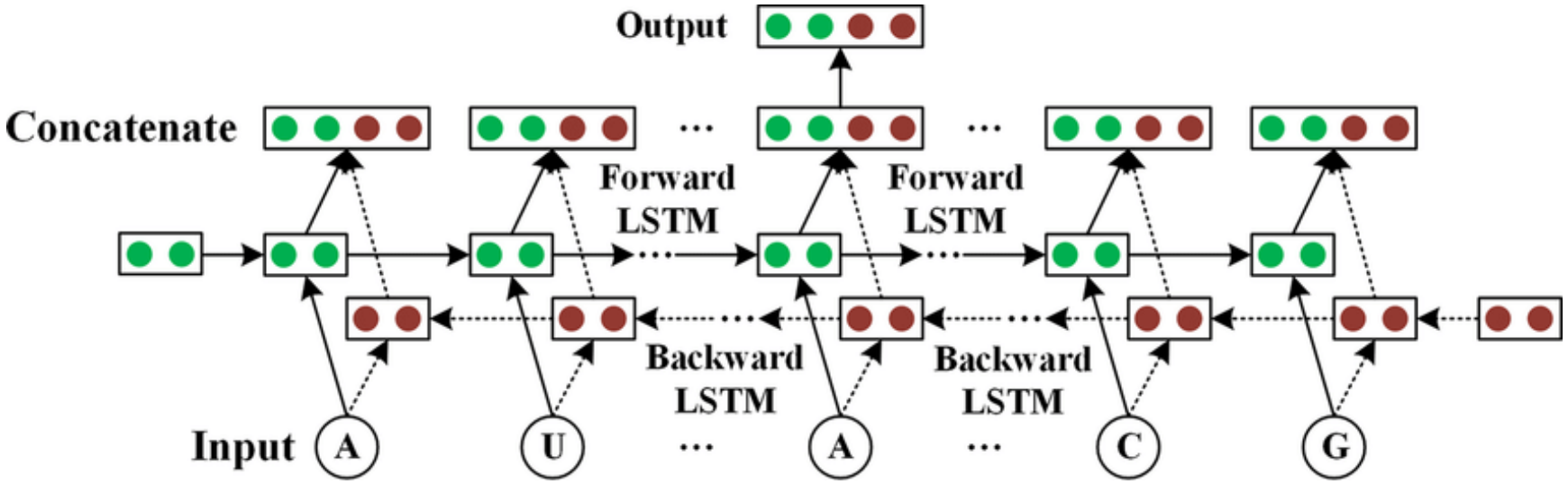


Ugnežđenost uparivanja



- Pseudočvorovi – neugnežđene interakcije
- Neki tipovi su mogu predvideti algoritmima visoke složenosti
- Neki samo homologijom

Duboko učenje



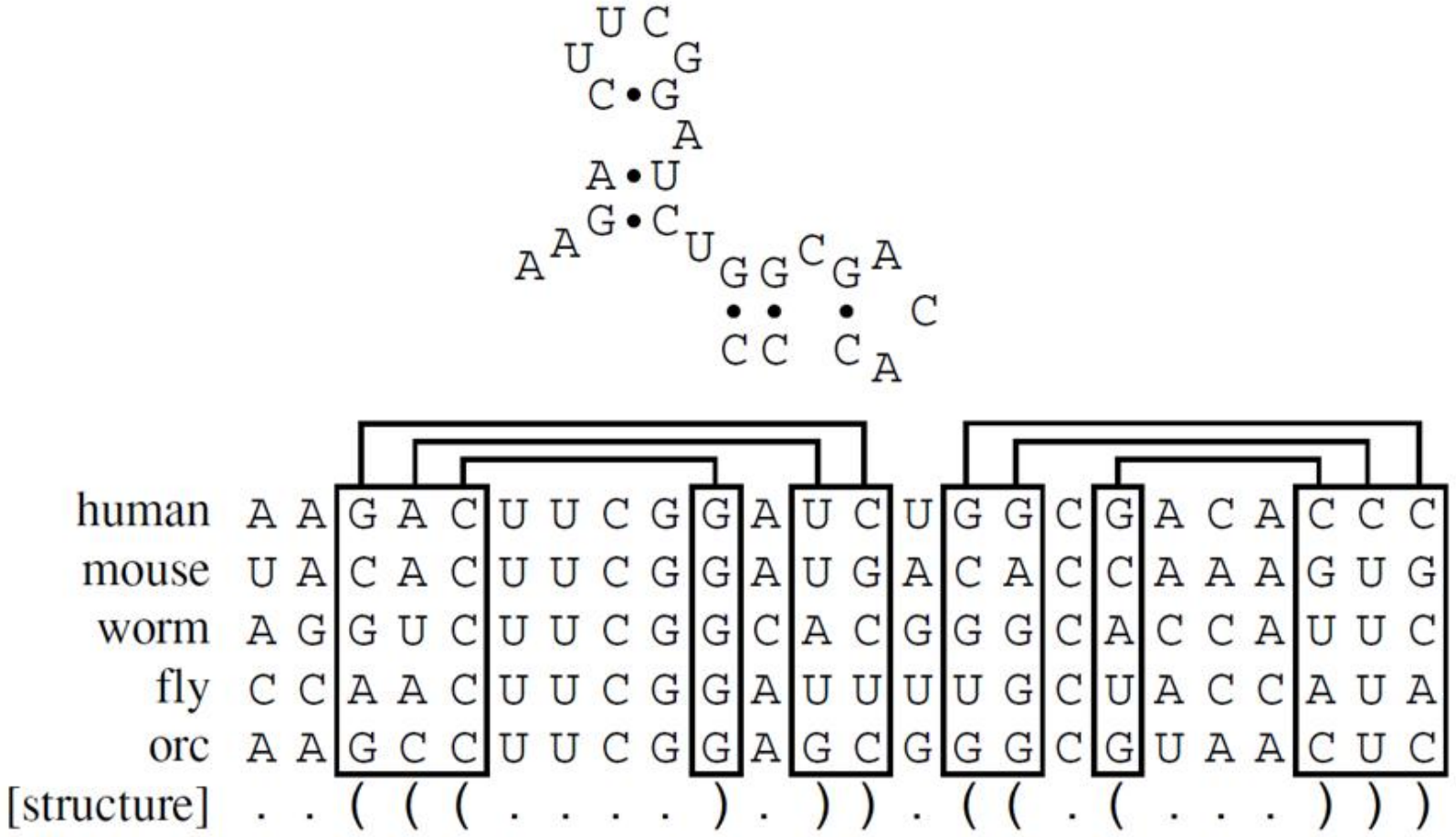
- Savremeni pristupi zasnovani na neuronskim mrežama i različitim reprezentacijama sekvence
- Još nisu nadmašile ostale modele u mnogim aspektima, ali jesu npr. kod pseudočvorova

4. OBRADA FAMILIJA RNK

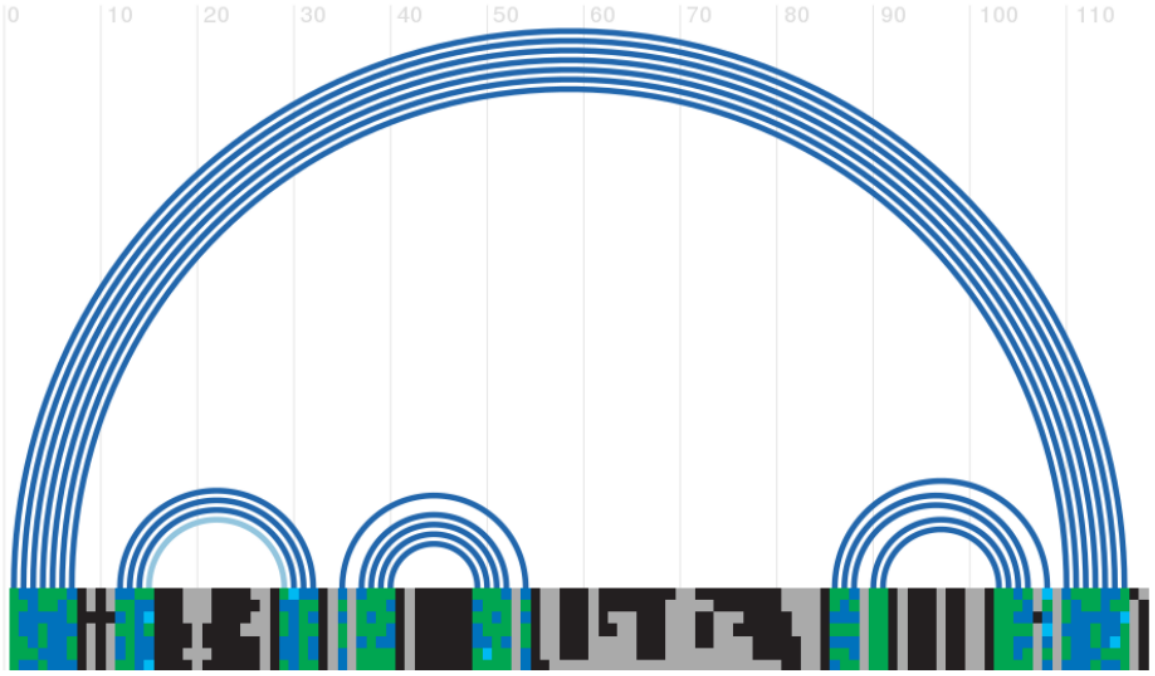
Familije RNK

- Sekvence istog porekla mogu imati različitu primarnu strukturu, ali identičnu sekundarnu
- Sekundarna struktura se menja sporije od primarne (mutacije se često kompenzuju), pa je ona glavna za modelovanje familija RNK, odnosno za modelovanje homologije RNK
- I za ovaj zadatak se mogu iskoristiti stohastičke kontekstno slobodne gramatike (SCFG, PCFG)

Modeli kovarijacije (CM)



Kovarijacioni model tRNK



	Acc-stem		D-stem	D-loop	D-stem		Ac-stem	Ac-loop	Ac-stem	V-region	T-stem	T-loop	T-stem	Acc-stem		CCA
-1	1	8	10	14	22	26	27	32	39	44	49	53	61	66	73	74
-	GGGCCCG	UA	GCUC	AGCCAGGAC-A	GAGC	G	CCGGC	CUUCUAA	GCCGG	UG----------CUG	CCGGG	UUCAAAU	CCCGG	CGGGCCC	G	CCA
-	GCCGCGG	UA	GUAU	AGCCUGGACUA	GUAU	G	CGGGC	CUGUCAA	GCCCG	UG----------A-C	CCGGG	UUCAAAU	CCCGG	CCGCGGC	G	CCA
-	GCCGGGG	UG	GCCG	AGC--GGUCUA	AGGC	G	GCGGG	CUGCAGA	CCCGU	UA-------G----------UUC	CCGGG	UUCGAAU	CCCGG	CCCCGGC	U	CCA
-	GGGCCCG	UA	GCUC	AGCCUGGU--A	GAGC	G	GCGGG	CUCUUAA	CCCGC	GAGG-----GAGG------AAGUC	CCGGG	UUCAAAU	CCCGG	CGGGCCC	G	CCA
-	GGGCCCG	UA	GCUC	AGCCCGGC--A	GAGC	G	GCGGG	CUUUUAC	CCCGC	GG-------AAG---------GUC	CCGGG	UUCAAAU	CCCGG	CGGGCCC	G	CCA
-	GGGCCCG	UA	GCUC	AGCCAGGU--A	GAGC	G	CCCGG	CUCAUAA	CCGGG	UG----------GUC	GGGGG	UUCAAAU	CCCCC	CGGGCCC	A	CCA

- Kovarijacioni modeli imaju vrlo preciznu konsenzus strukturu
- CM tRNK – tačno četiri petlje, kao i zavojnice

Gramatika familije tRNK

$$G \rightarrow s \text{ (glava je tačno jedan nukleotid)}$$

$$R \rightarrow Gcca \text{ (glava i } CCA \text{ rep)} \mid G \text{ (samo glava)}$$

$$S \rightarrow GZ^S R \text{ (glava, zavojnica i rep)} \mid Z^S R \text{ (zavojnica i rep)}$$

$$Z^S \rightarrow dZ^S d \text{ (uparivanje)} \mid GZ^S G \text{ (promašaj)} \mid GGZ^U Z^U LZ^U \text{ (unutrašnjost)}$$

$$Z^U \rightarrow dZ^U d \text{ (uparivanje)} \mid GZ^U G \text{ (promašaj)} \mid L \text{ (unutrašnjost)}$$

$$L \rightarrow GL \text{ (nizanje baza)} \mid GG \text{ (dve baze)}$$

- I sama gramatika eksploatiše dobro očuvani skelet strukture familije, npr. transportnih RNK u primeru
- Sa visokom tačnošću predviđa strukture u modelovanoj familiji

Pregled alata za predviđanje

- Freiburg RNA Tools – maksimizacija, statistički MEA model (najveća očekivana preciznost)
- RNAfold (Vienna RNA) – termodinamički (MFE) model, model centroida (“najbolji” predstavnik)
- CONTRAfold (Stanford) – poboljšane uslovne probabilističke kontekstno slobodne gramatike
- Mnogi drugi – UNAFold/mFold (MFE), SPOT-RNA (transformer), Ufold (enkoder-dekoder)...

Zaključak

- RNK se prepisuje kao jedan lanac, ali se zatim intramolekularno uvija u više nivoe strukture
- Transportna RNK ima specifičan oblik deteline s tri lista, informaciona strukturom kontroliše prevođenje, dok je ribozomalna vrlo složena
- Sekundarna struktura RNK može se predvideti maksimizacijom uparivanja (neuspešno), termodinamičkim i statističkim modelima, metodama homologije, dubokim učenjem...
- Problemi – modifikacije, pseudočvorovi, manjak baza podataka (zadatak za budućnost)...

Literatura

- Natalija Polović (2021) *Osnove biohemije*. Hemijski fakultet, Univerzitet u Beogradu.
- R. Durbin, S. Eddy, A. Krogh, G. Mitchison (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
- Robin D Dowell, Sean R Eddy (2004) *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction*. BMC Bioinformatics 5(1):71.
- Zhao Q, Zhao Z, Fan X, Yuan Z, Mao Q, et al. (2021) *Review of machine learning methods for RNA secondary structure prediction*. PLOS Computational Biology 17(8):e1009291.
- Laiyi F, Yingxin C, Jie W, Qinke P, Qing N, Xiaohui X (2022) *UFold: fast and accurate RNA secondary structure prediction with deep learning*. Nucleic Acid Research 50(3):e14.