Oznamy

- Budúci štvrtok 13.12.:
  - posledná prednáška
  - predposledné cvičenia pre informatikov
  - predposledné cvičenia pre biológov
- Štvrtok 20.12.:
  - posledné cvičenia pre informatikov
  - nepovinné prezentácie journal clubu
  - posledné cvičenia pre biológov
  - termín DÚ3
- Piatok 21.12. správy zo journal clubu
- Budúci štvrtok dohodneme:
  - či chcete prezentovať journal club (dohodnite sa v skupinách)
  - termín skúšky (doneste si termíny iných skúšok)

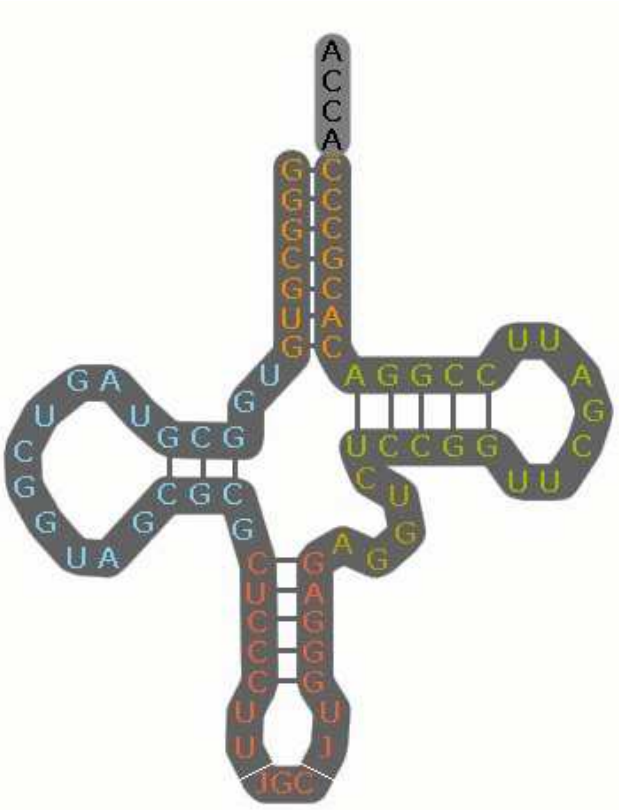
Správa zo journal clubu

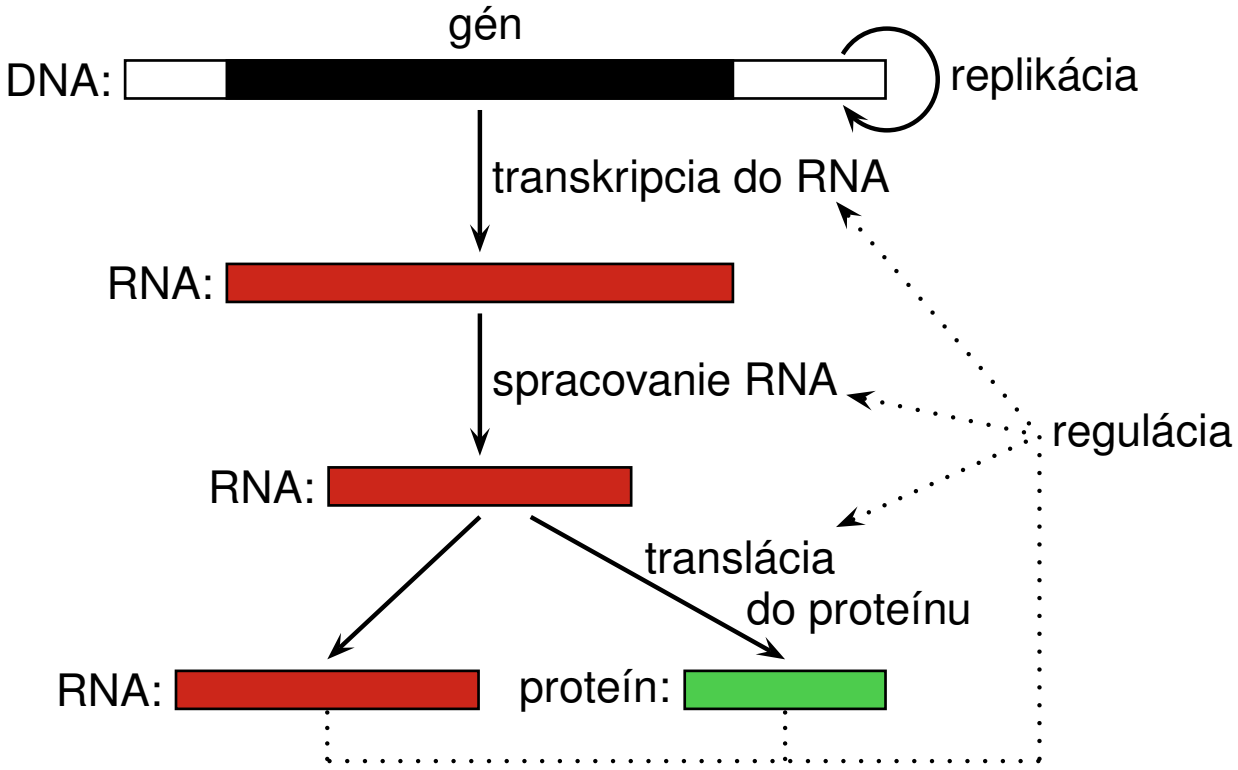
- Vlastnými slovami hlavné metódy a výsledky článku
- Pochopiteľná pre študentov tohto predmetu (inf aj bio)
- Netreba pokryť všetko a naopak, môžete využiť aj iné zdroje
- Skúste vložiť vlastný pohľad na tému
- Rozsah cca 1-2 strany na osobu, jeden ucelený text
- V správe vymenujte členov skupiny, ktorí sa podieľali na jej spísaní, dostanú rovnako bodov
- Pdf odovzdať cez Moodle (stačí 1 za skupinu)

RNA

Broňa Brejová

6.12.2018




gén
DNA:
replikácia
transkripcia do RNA
RNA:
spracovanie RNA
regulácia
RNA:
translácia
do proteínu
RNA:
proteín:

Vlastnosti RNA

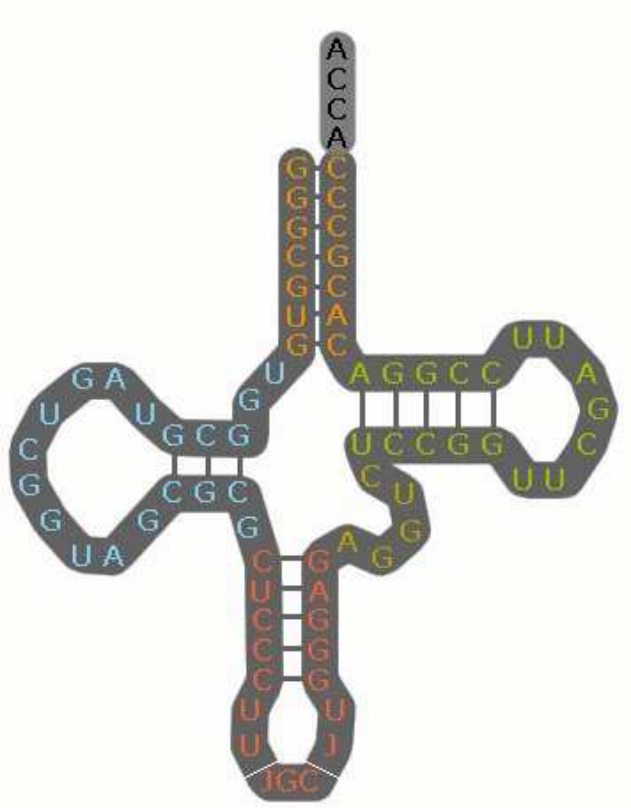
Ako sa líši od DNA?

- obsahuje ribózu namiesto deoxyribózy
- obsahuje uracil namiesto tymínu (bázy A,C,G,U)
- jednovláknové reťazce, zvyčajne kratšie
- zložitá sekundárna štruktúra: spárované komplementárne úseky
- okrem párov A-U, C-G aj nekanonické páry (napr. G-U)

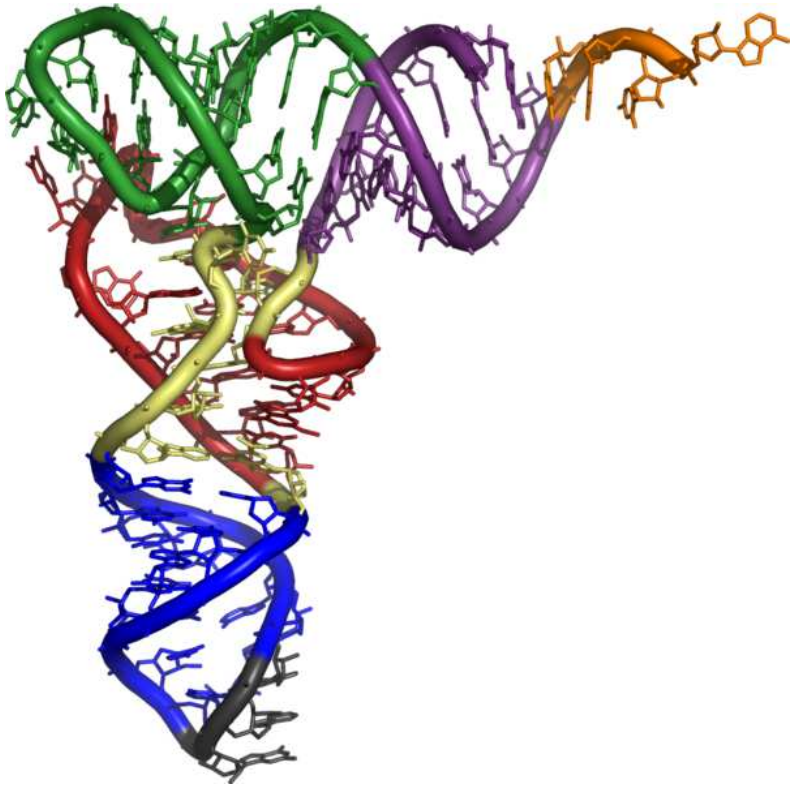
Štruktúra RNA

Príklad: transferová RNA (transfer RNA)

Sekundárna štruktúra
(secondary structure):
páry nukleotidov

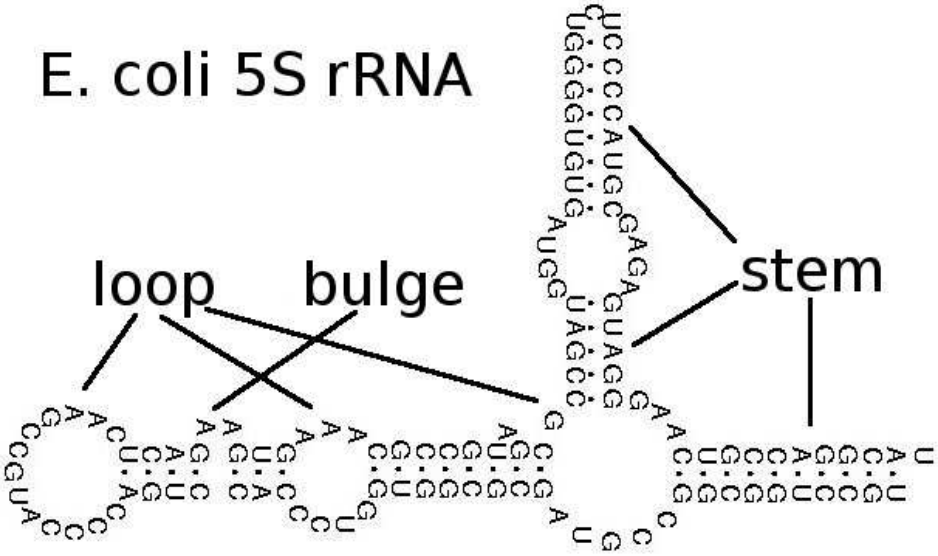


Terciárna štruktúra
(tertiary structure):
3D súradnice



Sekundárna štruktúra RNA

Prvky sekundárnej štruktúry



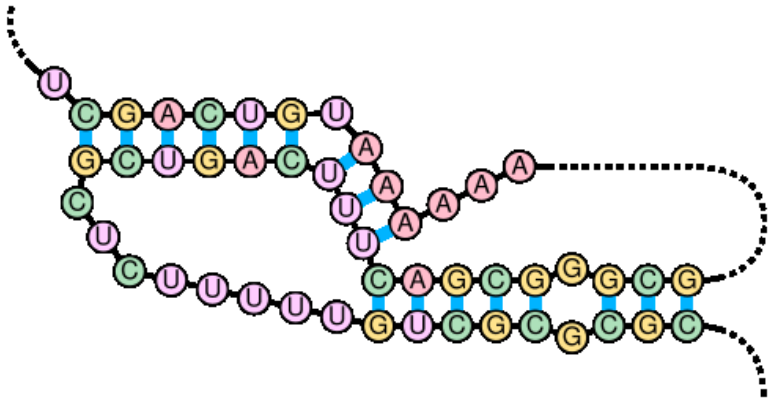
V tomto prípade spárované bázy tvoria **dobre uzátvorkovaný výraz:**

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

t.j. ak máme páry medzi pozíciami i a j a i' a j' a $i < i'$, tak buď $i < i' < j' < j$ alebo $i < j < i' < j'$.

Sekundárna štruktúra RNA

Pseudouzol: výnimka z dobrého uzátvorkovania



Mnohé algoritmy na prácu so sekundárnou štruktúrou ignorujú pseudouzly.

Zhruba 1.4% RNA nukleotidových párov v pseudouzloch.

Problém: určovanie štruktúry RNA

Vstup: RNA sekvencia

Cieľ: nájsť spárované bázy

Veľmi zjednodušená formulácia: nájdi dobre uzátvorkované spárovanie s najväčším počtom komplementárnych párov A-U, C-G.

Príklad: `((.((((()))(((.))))))))`

`GAACACAUGUAAAAUUUGUC`

Nussinovovej algoritmus

Dynamické programovanie:

Majme RNA $X_1, \ldots, X_n$.

Nech $A[i, j]$ je maximálny počet párov v podreťazci $X_i, X_{i+1}, \ldots, X_j$.

Rekurencia:

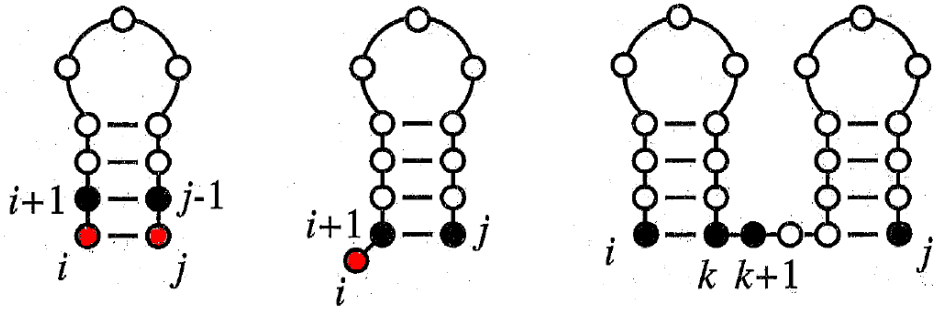
Podreťazce dĺžky 1: žiadne páry $A[i, i] = 0$

Dlhšie podreťazce: 3 prípady

– X_i a X_j sú pár: $A[i, j] = A[i + 1, j - 1] + 1$

– X_i je nespárované: $A[i, j] = A[i + 1, j]$

– X_i je pár s X_k pre $i < k < j$: $A[i, j] = A[i, k] + A[k + 1, j]$

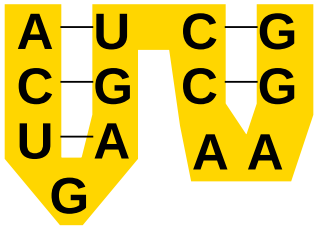


Rekurencia: $A[i,j] = \max \begin{cases} A[i+1, j-1] + c(X_i, X_j), \\ A[i+1, j], \\ \max_{k=i+1 \ldots j-1} \{A[i,k] + A[k+1, j]\} \end{cases}$

	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

$$c(X_i, X_j) = \begin{cases} 1 & \text{ak } X_i\text{-}X_j \text{ môže byť pár} \\ 0 & \text{inak} \end{cases}$$

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



Zložitosť:

$O(n^3)$ čas

$O(n^2)$ pamäť

Štruktúra s minimálnou voľnou energiou (MFE folding)

Realistickejšia formulácia problému určovania sek. štruktúry RNA.

Predpoklad: molekula v rovnovážnom stave
s minimálnou Gibbsovou voľnou energiou (Gibbs free energy).
Energie pre niektoré sekvencie experimentálne zmerané.

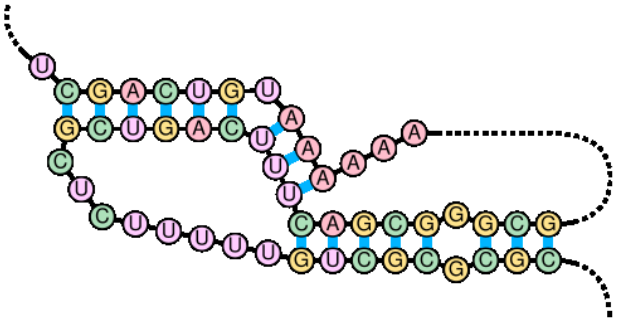
Nearest neighbor model: sada parametrov, energie pre dvojice susedných párov v helixoch, dĺžky slučiek atď.
Odvodené z nameraných dát.

Príklad:

```
                        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   .
```

Štruktúra s minimálnou energiu sa dá nájsť podobným (ale zložitejším) dyn. programovaním [Zuker and Stiegler, 1981].

Algoritmy dovoľujúce pseudouzly



Vo všeobecnosti NP-ťažký problém [Lyngso and Pedersen, 2000].

Pomalé dyn. programovanie $O(n^4)$ – $O(n^6)$ nájde niektoré typy pseudouzlov [Rivas and Eddy, 1999].

Tiež môžeme použiť heuristiky [Ren et al., 2005] (opakované vytváranie silných helixov).

Pravdepodobnostné modely na predikciu štruktúry

Chceme: model, ktorý generuje dvojice sekvencia a sek. štrukúra

Použitie: pre danú sekvenciu nájsť najpravdepodobnejšiu štruktúru

HMM nevhodné: závislosti medzi vzdialenými spárovanými bázami.

Stochastická bezkontextová gramatika,
stochastic context free grammar (SCFG):

Rozšírenie bezkontextových gramatík

Pravidlám pridáme pravdepodobnosti

Stochastické bezkontextové gramatiky (SCFG)

neterminály (veľké písmená) podobné na stavy v HMM,
terminály (malé písmená) reprezentujú nukleotidy.
Pravidlá prepisujú neterminál na reťazec terminálov a neterminálov.
Každé pravidlo má pravdepodobnosť.

Príklad: jeden neterminál, 14 pravidiel (ϵ =prázdny reťazec)

$$S \to \overbrace{aSu}^{0.1} \mid \overbrace{uSa}^{0.1} \mid \overbrace{cSg}^{0.1} \mid \overbrace{gSc}^{0.1} \mid$$
$$\overbrace{aS}^{0.05} \mid \overbrace{cS}^{0.05} \mid \overbrace{gS}^{0.05} \mid \overbrace{uS}^{0.05} \mid \overbrace{Sa}^{0.05} \mid \overbrace{Sc}^{0.05} \mid \overbrace{Sg}^{0.05} \mid \overbrace{Su}^{0.05} \mid \overbrace{SS}^{0.1} \mid \overbrace{\epsilon}^{0.1}$$

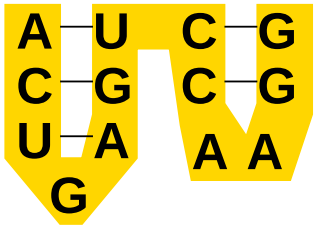
V každom kroku zvoľ jeden (napr. najľavejší) neterminál,
prepíš ho náhodne zvoleným pravidlom:

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

Stochastické bezkontextové gramatiky

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

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



Bázy vygenerované v jednom kroku sú spárované.

Úloha: Nájdi najpravdepodobnejšie odvodenie danej RNA
$\Rightarrow$ určuje sekundárnu štruktúru

Riešenie: Dynamické programovanie (CYK algoritmus), $O(n^3)$

Trénovanie parametrov: zo známych RNA štruktúr

Gramatiky vs. minimalizácia energie

Výhody gramatík:

- možno automaticky trénovať, netreba náročné experimenty.
- rozšíriteľné na modely viacerých sekvencií.

Nevýhody gramatík:

- jednoduché gramatiky nevystihujú komplexitu problému
- nižšia presnosť ako minimalizácia energie.

Conditional log-linear models: podobné na SCFG, ale:

- maximalizujeme podmienenú pravdepodobnosť správnej odpovede (diskriminatívne trénovanie: nemodelujeme rozdelenie pravdepodobností samotných RNA sekvencií, iba štruktúr)
- dosahujú dokonca lepšiu presnosť ako minimalizácia energie.

Evolúcia RNA sekvencií

Často vidíme koreláciu medzi mutáciami v spárovaných bázach.
Napr. C sa zmení na A, spárované G sa súčasne zmení na U

Príklad: niekoľko sekvencií z D ramena tRNA

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

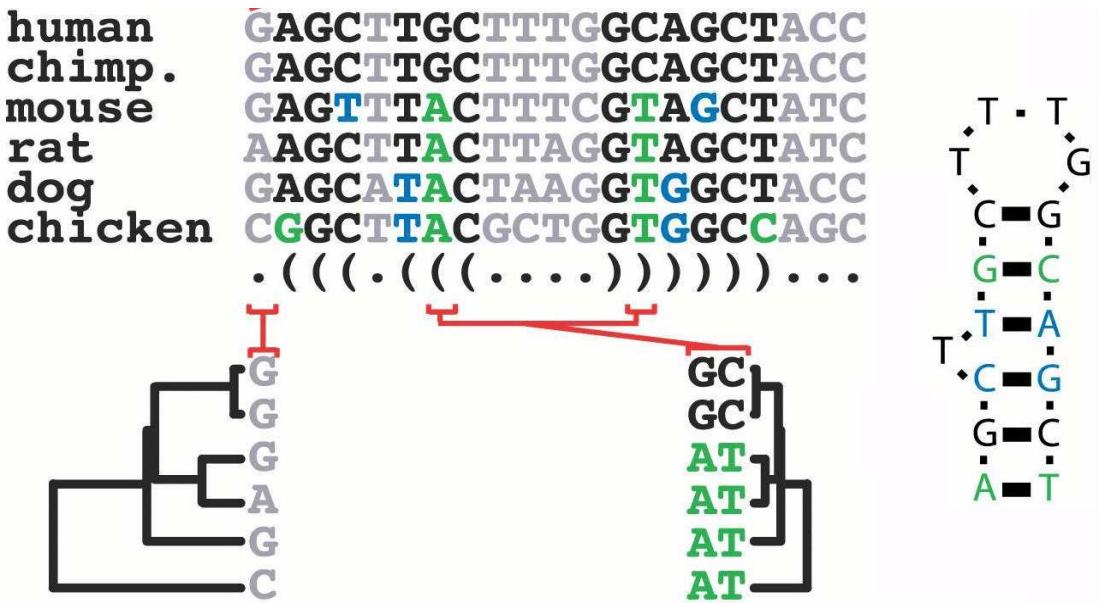
Úloha: daných je niekoľko (zarovnaných) sekvencií RNA
nájdite ich spoločnú RNA štruktúru

(korelácie medzi spárovanými bázami potvrdzujú správnosť štruktúry)

Hľadanie spoločnej štruktúry pre viacero sekvencií

Phylo-SCFG:

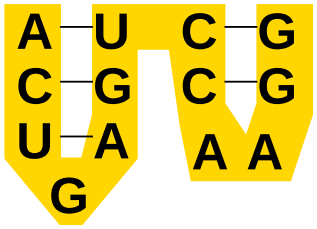
- namiesto jednotlivých báz emituje stĺpce zarovnania podľa fylogenetického stromu.
- nespárované bázy emituje bežnou substitučnou maticou,
- spárované bázy substitučnou maticou dvojíc (16×16).



Problém: hľadanie známych typov RNA génov v genóme

- Obdoba databázy Pfam $\Rightarrow$ databáza Rfam
  2473 rodín podobných RNA génov
- Pre každú rodinu zarovnanie a pravdepodobnostný model
- Proteínové rodiny reprezentujeme profilovými HMM
- Pre RNA kovariančné modely (covariance model, CM):
  špeciálny typ SCFG

Kovariančný model



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

- S =start, E_i =end
  P_i =pár, L_i =nespárovaná báza vľavo, R_i =nespárovaná báza vpravo
  ďalšie neterminály modelujú indely.
- terminály (bázy) sa emitujú s pravdepodobnosťami podľa príslušného stĺpca zarovnania
  Napr. $P_1 \to \overbrace{aP_2u}^{0.2} \mid \overbrace{uP_2a}^{0.2} \mid \overbrace{cP_2g}^{0.4} \mid \overbrace{cP_2u}^{0.1}$
- veľkosť gramatiky úmerná dĺžke modelovanej RNA rodiny

Kovariančný model

Použitie:

hľadaj výskyty génu v DNA (lokálne zarovnanie),
nájdi štruktúru nového génu z tej istej rodiny (globálne zarovnanie).

Dynamické programovanie: čas $O(MND^2)$,
$M =$ počet neterminálov v gramatike, úmerný dĺžke zarovnania,
$N =$ dĺžka DNA sekvencie,
$D =$ max. dĺžka RNA génu v DNA (úmerná M).

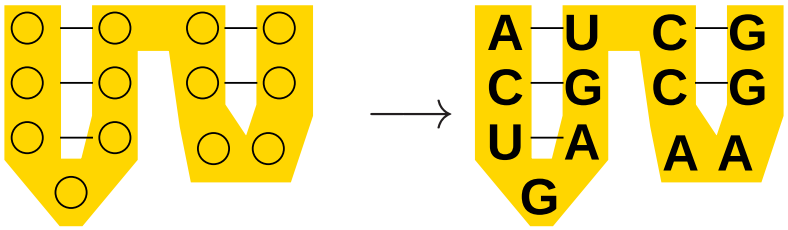
Zrýchlenie:

nájdi sľubné úseky podobné na sekvencie v RNA rodine
(iba na základe podobnosti sekvencií)
aplikuj CM iba na sľubné úseky

Problém: dizajn RNA

Daná RNA sekundárna štruktúra (párovanie).
Nájdi sekvenciu, pre ktorú je táto štruktúra optimálna.

Nie je známy efektívny algoritmus, heuristiky často nájdu sekvenciu pomerne rýchlo.



Použitie: skúmanie možných RNA štruktúr, vývoj liekov (ribozymes, riboswitches), RNA pre laboratórne techniky, RNA nanoštruktúry

Zhrnutie

- Určovanie sekundárnej štruktúry RNA:
  minimalizácia energie, pravdepodobnostné SCFG
- Lepšie výsledky, keď použijeme zarovnanie viacerých sekvencií (PhyloSCFG)
- Známe rodiny reprezentujeme pomocou kovariančných modelov v nových sekvenciách hľadáme výskyty rodín z databázy Rfam
- Väčšina problémov sa dá riešiť dynamickým programovaním, ktoré je pomerne pomalé a ignoruje pseudouzly.
- Ďalšie zaujímavé problémy: napr. dizajn RNA štruktúr