

Metódy v bioinformatike, 1-BIN-301/2-AIN-501

Vyučujú:

Broňa Brejová, M-163, bronislava.brejova@fmph.uniba.sk

Tomáš Vinař, M-163, tomas.vinar@fmph.uniba.sk

Jana Černíková, M-25, jana.cernikova@fmph.uniba.sk

Web: <https://fmfi-compbio.github.io/mbi/>

Literatúra:

I-INF-D-23 : Durbin, Eddy, Krogh, Mitchison: Biological sequence analysis.

Cambridge University Press 1998.

I-INF-Z-2 : Zvelebil, Baum: Understanding Bioinformatics. Taylor&Francis 2008.

Skriptá k predmetu, poznámky a videá na webstránke.

Časy a miestnosti

- Prednáška štvrtok 15:40-17:10 F1-109
- Cvičenia informatici štvrtok 14:00-15:30 F1-109
- Cvičenia biológovia štvrtok 17:20-18:50 M-217
(ak prednáška skončí skôr, cvičenia začnú po krátkej prestávke)

“Informatici”: študenti informatiky, bioinformatiky, aplikovanej informatiky, dátovej vedy; zapíšte si 1-BIN-301

“Biológovia”: študenti z PriFUK, študenti biomedicínskej fyziky; zapíšte si 2-AIN-501

Ostatní: porad'te sa, do ktorej skupiny sa zaradiť

Ciele predmetu

- **Všetci:** Prehľad základných metód na výpočtovú analýzu biologických sekvencií a ďalších dát v molekulárnej biológii.
- **Informatici:** Algoritmy a dátové štruktúry, strojové učenie, pravdepodobnosť. Ako prejsť od problému v reálnom svete k matematickej abstrakcii.
- **Biológovia:** Matematické modely tvoriace základ populárnych bioinformatických nástrojov, používanie nástrojov, interpretácia výsledkov.
- **Všetci:** Skúsenosť s interdisciplinárnou spoluprácou.

Známkovanie

3 domáce úlohy 30% (10% každá)

Journal club 10%

Kvízy 10% (každý týždeň 1 bod)

INF: Skúška 50%

BIO: Projekt 50%

Hodnotenie: A: 90+, B: 80+, C: 70+, D: 60+, E: 50+

INF: Zo skúšky treba aspoň polovicu bodov

BIO: Aktívna účasť na cvičeniach

- Dve verzie DÚ: biologická a informatická
- Journal club: čítanie 1 článku v skupine a správa (prípadne nepovinná prezentácia)
- Na skúške povolený ťahák 2 listy A4, môže mať ústnu časť
- Neodpisovať!

Čo nás v tomto predmete čaká

Typická prednáška

- Biologické pozadie problému
- Formulácia ako informatický problém
- Idea algoritmu (riešenia problému)

Typické cvičenia

- Informatici: ďalšie detaily algoritmov, potrebné poznatky z biológie
- Biológovia: aplikácia na konkrétne dáta, význam rôznych parametrov, potrebné poznatky z informatiky

Týždenné kvízy

- Cca 5 krátkych otázok týkajúcich sa prednášky aj cvičení
- Vypĺňajte od štvrtka 19:00 do ďalšej stredy 22:00
- Linku na Moodle s kvízmi nájdete na stránke predmetu
- Cieľ: pripomenúť si aspoň základné pojmy z prednášky a cvičení
- **Prvý kvíz už tento týždeň**

Príklad z nášho výskumu

Kosmáč bielofúzy

(common marmoset, *Callithrix jacchus*, štvrt' kila, 18cm)



Genóm osekvenovaný 2007

(Washington University St. Louis a Baylor College of Medicine, USA)

Analýza publikovaná v roku 2014

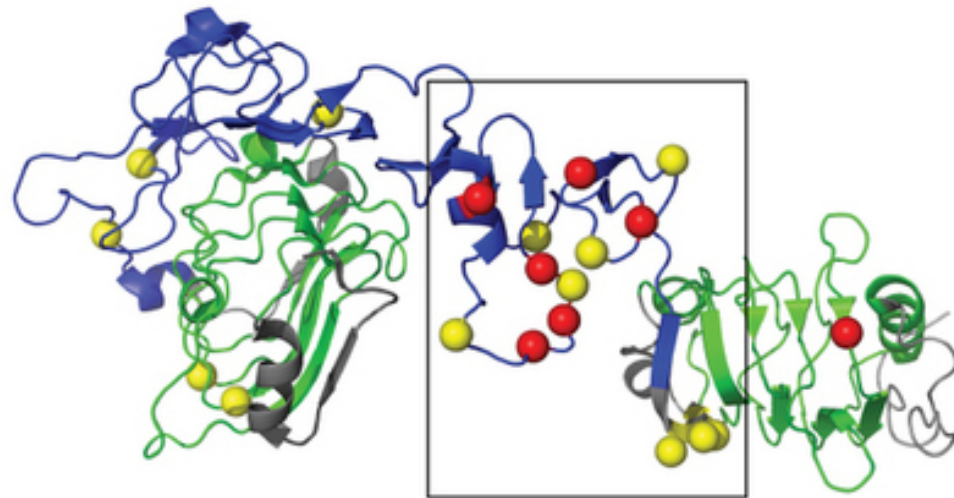
IGF1R: Insulin-like growth factor 1 receptor

Proteín prechádza cez cytoplazmatickú membránu na povrchu bunky

Po naviazaní hormónov IGF1, IGF2 signalizuje dovnútra bunky

Súvisí s rastom a delením bunky, rastom organizmu, rakovinou

human	RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKT
chimp	RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKT
orang	RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKT
macaque	RDFCANILSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKT
marmoset	RQFCASIVSSENENKGFVIHDGECMQDCPSGFIRDTHSMQCIPCKGPCPKVC-D-EQMAK
mouse	RDFCANIPNAESSDSDGFVIHDDECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKT
rat	RDFCANIPNAESSDSDGFVIHDGECMQECPSGFIRNSTQSMYCIPCEGPCPKVCGDEEKKT
dog	RDFCANIPSAESSDSEGFVIHDGECMQECPSGFIRNGSQSMYCIPCEGPCPKVC-EEEKKT



Aké bioinformatické nástroje boli potrebné k tomuto výsledku?

1. Zostavenie genómu
2. Nájdenie zarovnaní s inými genómami
3. Hľadanie génov kódujúcich proteíny
4. Hľadanie génov s pozitívnym výberom
5. Určovanie štruktúry a funkcie proteínov

1. Zostavenie genómu

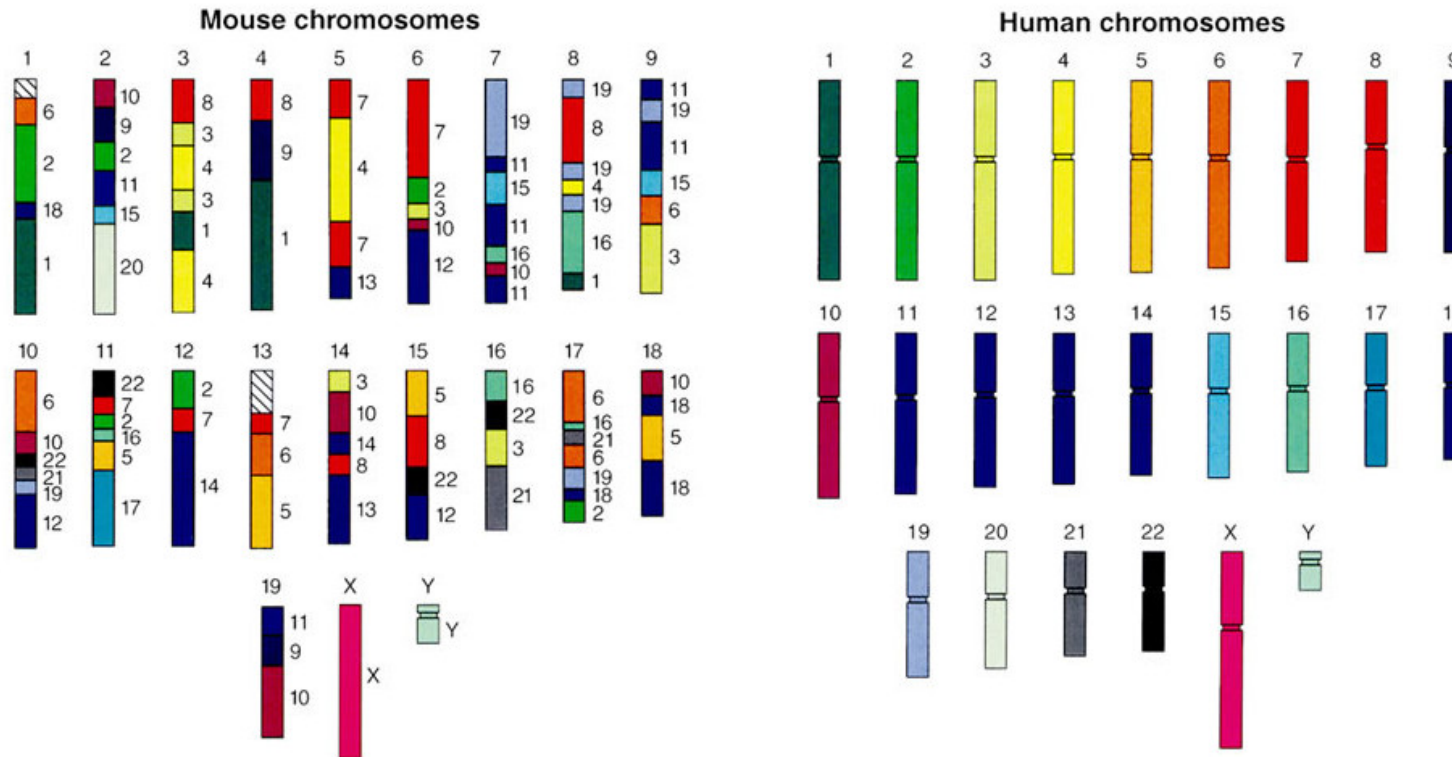
- Pri sekvenovaní DNA vieme čítať len krátke kúsky (napr. dĺžky 1000)
- Každé miesto v genóme prečítame viackrát (u kosmáča priemer $6\times$)



- Čítania “zliepame” na základe prekryvov
- Veľmi veľa dát \Rightarrow potreba veľmi efektívnych programov

2. Nájdenie zarovnaní s inými genómami

Ku každému miestu v genóme kosmáča chceme nájsť zodpovedajúce časti iných genómov (napr. človek, šimpanz, myš, ...)



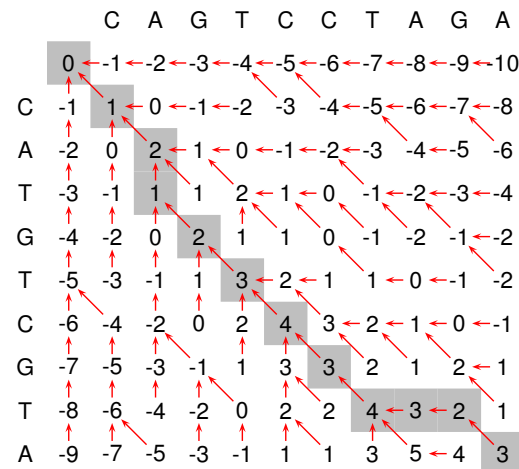
2. Nájdenie zarovnaní s inými genómami

- Hľadáme podobnosti medzi DNA sekvenciami

```

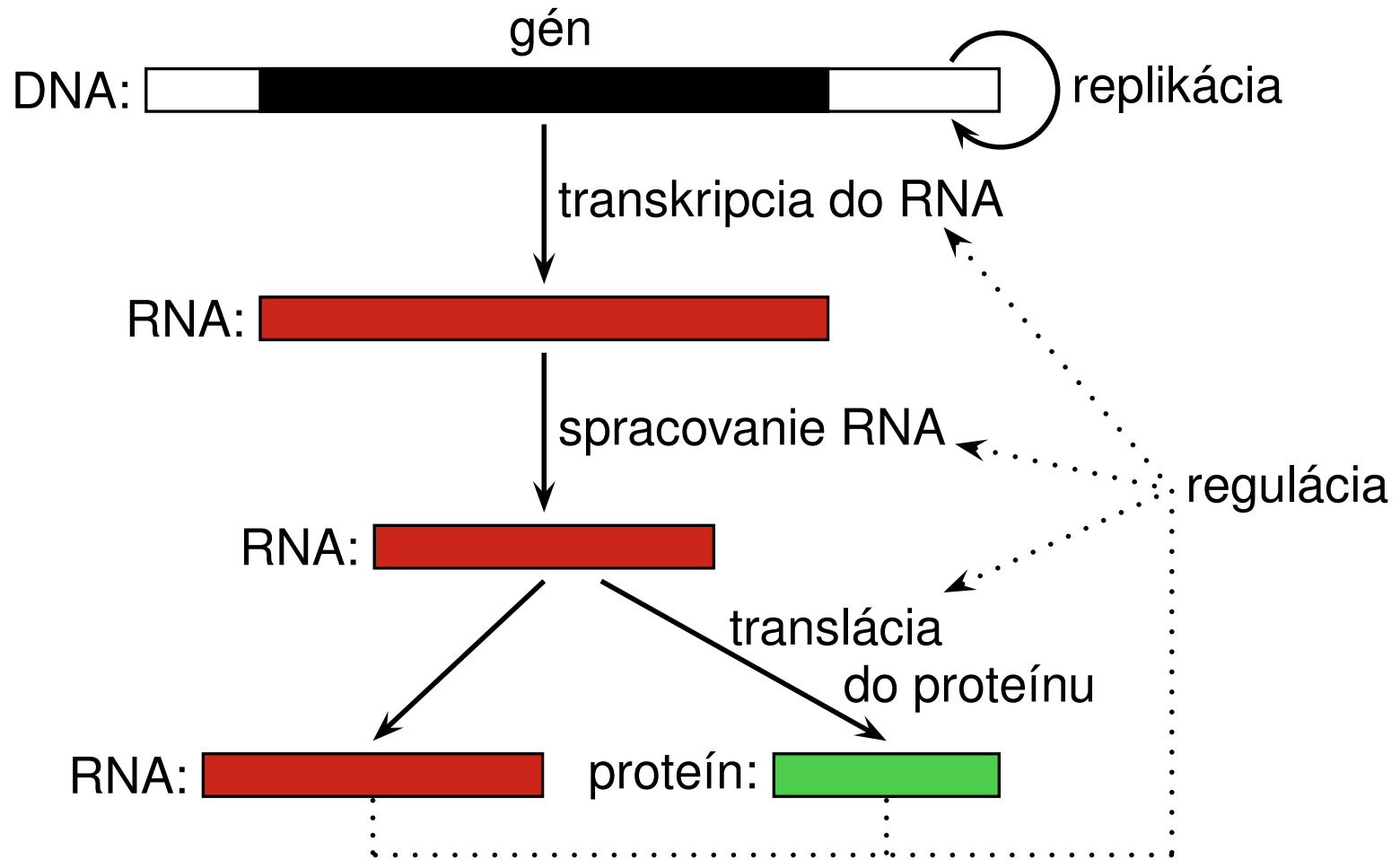
Human  AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGA
Rhesus  AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGA
Mouse   GGTGGCTGCCGGGCTG---GGTGGCTGAGGCCTTGTTGGTGGGGT
Dog     AGTGGCTGCCCGGCTG---GGTGGCTGAGGCCTTATTTGCAGGGA
Chicken AGTGGCTGCCAGTCTGCGCCGTGGCCGACGTCTTGCTCGGGGGAA
    
```

- Základ je technika **dynamického programovania**, ktorá veľký problém rozkladá na veľa malých podproblémov



- Tabuľka je veľmi veľká, v praxi treba pridať veľa vylepšení

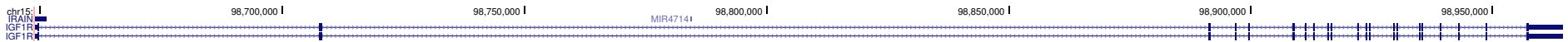
3. Hľadanie génov kódujúcich proteíny



Ktoré časti osekvenovaného genómu kódujú proteíny?

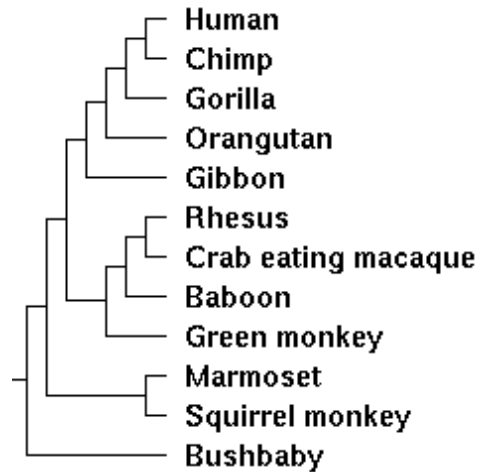
3. Hľadanie génov kódujúcich proteíny

- Hľadanie ihly v kope sena: iba cca 1% ľudského genómu kóduje proteíny
- Kód pre jeden proteín rozbitý do veľa krátkych exónov
- Napr. IGF1R zaberá 315 569nt, z toho kóduje 4101nt v 21 exónoch



- Zoberieme známe gény, spravíme rôzne štatistiky
potom hľadáme iné oblasti s podobným štatistickým profilom

4. Hľadanie génov s pozitívnym výberom



- Štúdium evolučných procesov
- V DNA vznikajú mutácie, tie však podliehajú prirodzenému výberu
- Väčšina náhodných zmien v proteíne je škodlivých, preto sa proteíny menia pomerne pomaly

4. Hľadanie génov s pozitívnym výberom

- Niekedy sa však proteín mení rýchlejšie, nakoľko náhodné zmeny sú užitočné (pozitívny výber)
- Veľké množstvo zmien v proteíne môže znamenať zmeny vo funkcii

human	R	D	F	C	A	N	I	L	S	A	E	S	S	D	S	E	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	G	S	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	-	E	E	E	K	K	T	K
chimp	R	D	F	C	A	N	I	L	S	A	E	S	S	D	S	E	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	G	S	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	-	E	E	E	K	K	T	K
orang	R	D	F	C	A	N	I	L	S	A	E	S	S	D	S	E	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	G	S	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	-	E	E	E	K	K	T	K
macaque	R	D	F	C	A	N	I	L	S	A	E	S	S	D	S	E	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	G	S	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	-	E	E	E	K	K	T	K
marmoset	R	Q	F	C	A	S	I	V	S	S	E	N	S	E	N	N	K	F	V	I	H	D	G	E	C	M	Q	D	C	P	S	G	F	I	R	D	T	T	H	S	M	Q	C	I	P	C	K	G	P	C	P	K	V	C	-	D	-	E	Q	M	A	K
mouse	R	D	F	C	A	N	I	P	N	A	E	S	S	D	S	D	G	F	V	I	H	D	D	E	C	M	Q	E	C	P	S	G	F	I	R	N	S	T	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	G	D	E	E	K	K	T	K
rat	R	D	F	C	A	N	I	P	N	A	E	S	S	D	S	D	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	S	T	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	G	D	E	E	K	K	T	K
dog	R	D	F	C	A	N	I	P	S	A	E	S	S	D	S	E	G	F	V	I	H	D	G	E	C	M	Q	E	C	P	S	G	F	I	R	N	G	S	Q	S	M	Y	C	I	P	C	E	G	P	C	P	K	V	C	-	E	E	E	K	K	T	K

5. Určovanie štruktúry a funkcie proteínov

- Spravili sme kroky 1-4 a dostali sme zoznam 37 génov pod vplyvom pozitívneho výberu v kosmáči
- Čo tie gény robia, ktoré by mohli súvisieť s veľkosťou?
- Aký má daný proteín tvar, kde sú pozície, ktoré sa v evolúcii zmenili?
- Štruktúra (tvar) proteínov sa dá určovať experimentálne je to drahé, namiesto toho predikcia 3D štruktúry

