

Organizačné poznámky

- DÚ1 bude zverejnená dnes, odovzdávanie do stredy 13.11. 22:00
- Journal club: Rozdelenie do skupín / inštrukcie na konci prednášky
- Najbližší journal club termín 22.11.

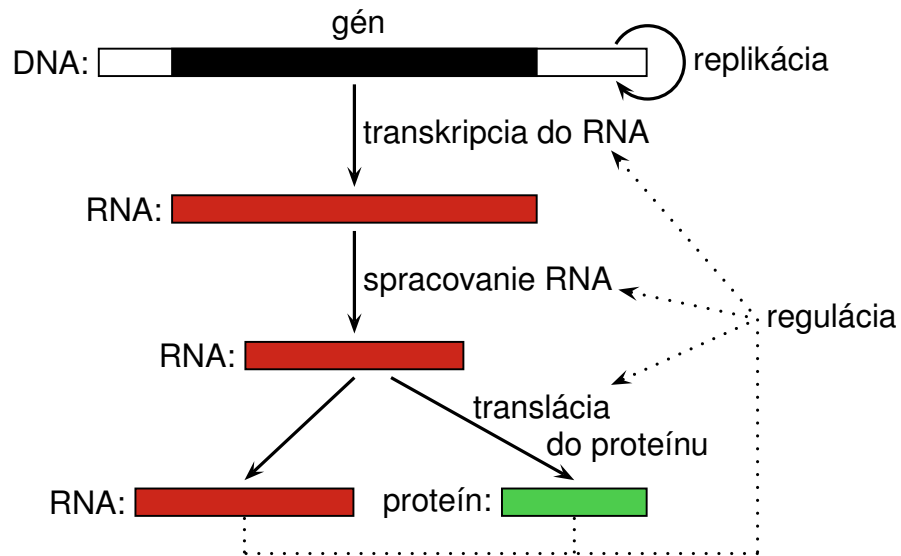
Hľadanie génov

Tomáš Vinař

24.10.2024



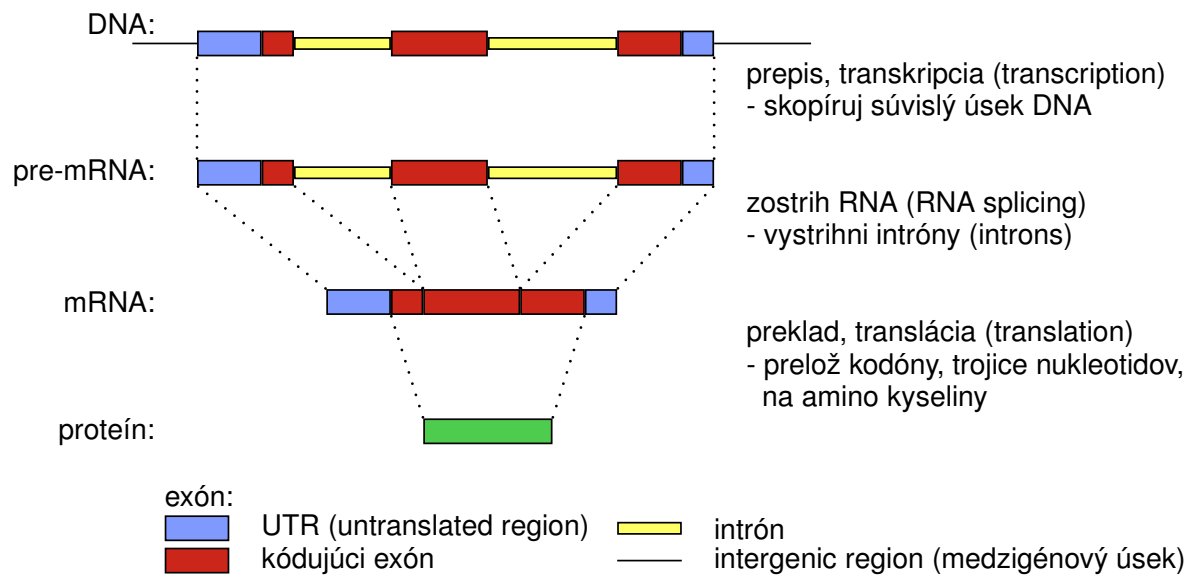
Čo s osekvenovanými genómami?



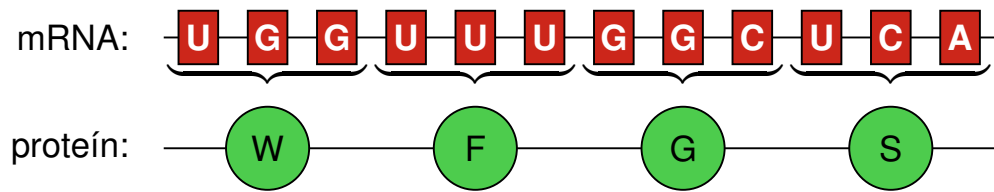
- gény kódujúce proteíny (dnešná prednáška)
- RNA gény
- signály pre reguláciu transkripcie, zostrihu, atď
- pseudogény (nefunkčné kópie génov)
- repetitívne sekvencie, opakovania (sequence repeats)

Štruktúra eukaryotických génov

Proces syntézy proteínov:



Translácia: tri bázy mRNA (kodón) → aminokyselina proteínu



L'udský genóm

Gény kódujúce proteíny

- cca 20 000, pokrývajú 40% genómu
- cca 10 exónov v géne
- exóny pokrývajú 2% genómu
- kódujúce exóny 1.2% genómu

Repetitívne sekvencie

- pokrývajú 49% genómu

Príklad: gén IGF1R zaberá 315 569nt, z toho kóduje 4101nt v 21 exónoch



Bioinformatický problém: hľadanie génov

Cieľ: nájsť všetky gény kódujúce proteíny v genóme.

Tým získame katalóg všetkých proteínov.

Zjednodušenia:

- neuvažujeme alternatívny zostrih, prekrývajúce sa gény
- nehľadáme neprekladané oblasti (UTRs) na začiatku a konci génu

Bioinformatický problém: hľadanie génov

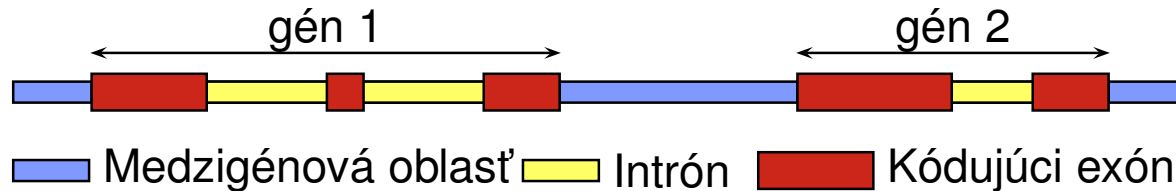
Vstup: sekvencia DNA

cggtgaaactgcacgattggtgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatattgcgctagtgttgggtggtccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcgggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtcactgaactgcttattc
gtaaactaccttccagcgcctggtgcgcgagattgcgcaggactttaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgctcactatcatgccaagg
acatccagctcgcccgccgcatccgcgagagagggcggtgattactgtggtctctctgac

Bioinformatický problém: Hľadanie génov

Cieľ: označ každú bázu ako intrón/exón/medzigénovú oblasť

```
cggtgaaactgcacgattggtgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtggtccgctgtgctgtttttccgctcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgcctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccggtatcagaagtccactgaactgcttattc
gtaaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgctcactatcatgccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgatttactgtggtctctctgac
```



Bioinformatický problém: hľadanie génov

Vstup: sekvencia DNA

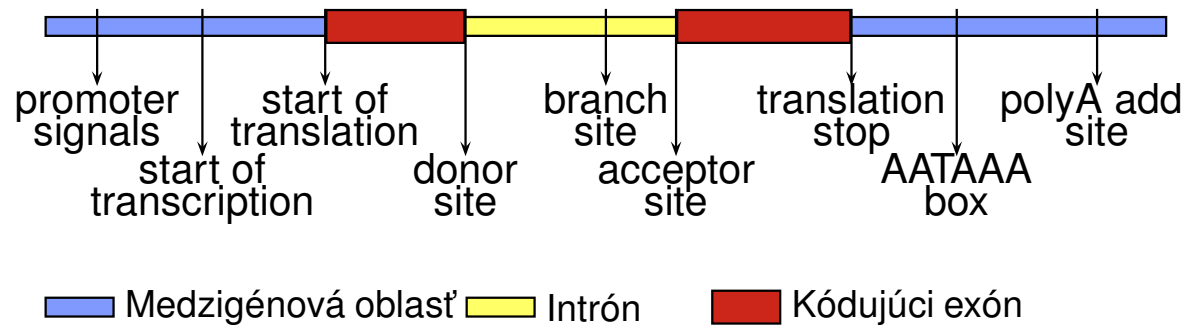
Cieľ: označ každú bázu ako intrón/exón/medzigénovú oblasť (anotácia)

- Toto nie je dobre definovaný problém!
Ako spoznáme, čo je gén?

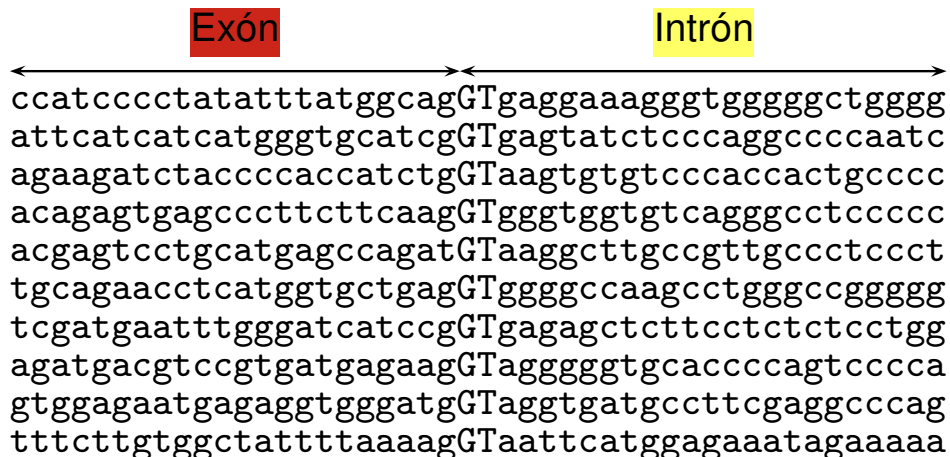
Ako spoznáme gény?

Signály na hraniciach exónov:

krátke reťazce, kde sa viažu komplexy zúčastnujúce sa na expresii génu



Príklad signálu: miesto zostrihu



Ako spoznáme gény?

Zloženie sekvencie:

- iná frekvencia k -tic báz v kódujúcich a nekódujúcich oblastiach,
- kódujúce oblasti sú 3-periodické,
- stop kodóny (TAA, TGA, TAG) len na konci posledného kódujúceho exónu.

Príklad: ak uvažujeme len jednotlivé bázy, exóny majú viac C a G (ľudský genóm)

| | | a | c | g | t |
|---------------|----------|------|------|------|------|
| kódujúci exón | 0 | 0.26 | 0.26 | 0.32 | 0.16 |
| | 1 | 0.30 | 0.24 | 0.20 | 0.26 |
| | 2 | 0.17 | 0.32 | 0.31 | 0.20 |
| intrón | | 0.26 | 0.22 | 0.22 | 0.30 |
| medzig. | | 0.27 | 0.23 | 0.23 | 0.27 |

Bioinformatický problém: hľadanie génov

Vstup: sekvencia DNA

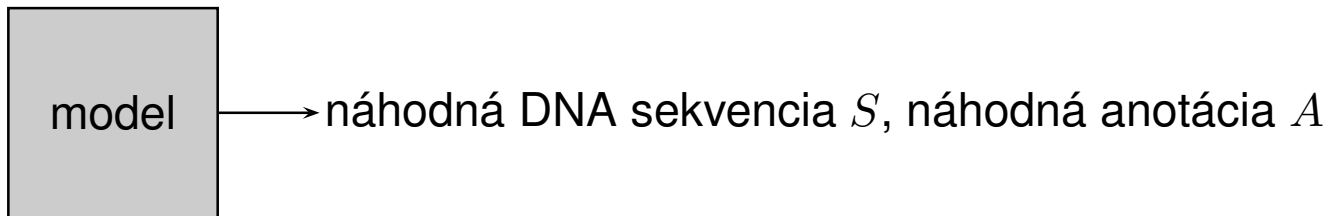
Cieľ: označ každú bázu ako intrón/exón/medzigénovú oblasť (anotácia)

- Toto nie je dobre definovaný problém!
Ako spoznáme, čo je gén?
- Žiadna informácia nám neumožňuje jednoznačne určiť, čo je gén.
- Chceme **skórovací systém**, ktorý povie, ako dobre potenciálna anotácia zodpovedá našim znalostiam.
- Potom hľadáme anotáciu (alebo: segmentáciu pôvodnej sekvencie na neprekrývajúce sa regióny, ktoré reprezentujú intróny, exóny a medzigénové úseky) **s maximálnym skóre.**
- Na definíciu skórovacieho systému použijeme **pravdepodobnostné modely.**

Pravdepodobnostný model génov

Žiadna informácia nám neumožňuje jednoznačne určiť, čo je gén.

Skombinujeme dostupnú informáciu pravdepodobnostným modelom.



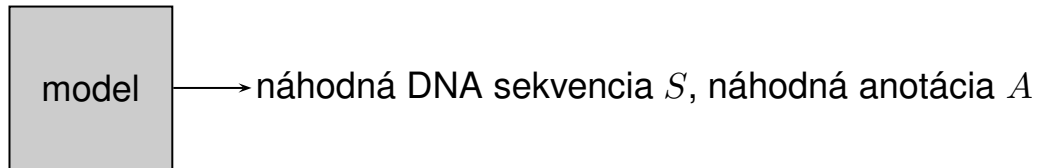
$\Pr(S, A)$ – pravdepodobnosť, že model vygeneruje pár (S, A) .

Model zostavíme tak, aby páry s vlastnosťami podobnými skutočným génom mali veľkú pravdepodobnosť.

Použitie: pre novú sekvenciu S nájdí najpravdepodobnejšiu anotáciu

$$A = \arg \max_A \Pr(A|S)$$

Pravdepodobnostný model génov



Použitie: pre sekvenciu S nájsi najpravdepodobnejšiu anotáciu A

Hračkársky príklad modelu: sekvencie dĺžky 2

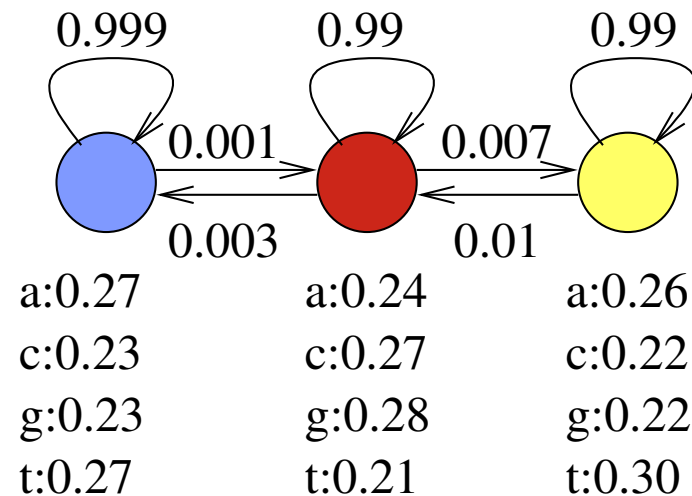
Tabuľka pravdepodobností pre 16 sekvencií, 9 anotácií (súčet 1)

Najpravdepodobnejšia anotácia pre $S = aa$ je **aa**.

| | | | | | | |
|----|-------|-----|-------|-----|--------|-----|
| aa | 0.008 | ac | 0.009 | ag | 0.0085 | ... |
| aa | 0 | ac | 0 | ... | | |
| aa | 0.011 | ... | | | | |
| aa | 0 | | | | | |
| aa | 0.009 | | | | | |
| aa | 0 | | | | | |
| aa | 0.007 | | | | | |
| aa | 0 | | | | | |
| aa | 0.010 | | | | | |

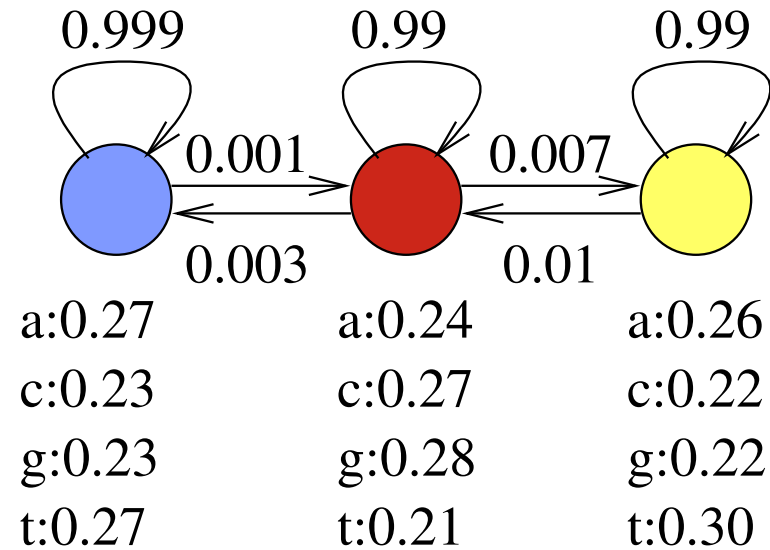
Skrytý Markovov model, hidden Markov model (HMM)

Spôsob, ako zdefinovať model pre dlhšie sekvencie.



- Konečný automat, stavy napr. exón, intrón, medzigénová oblasť
- Sekvenciu aj anotáciu generuje bázu po báze
- V každom kroku je v jednom stave a náhodne vygeneruje jednu bázu podľa tabuľky v stave
- Potom sa presunie do ďalšieho stavu podľa pravdepodobností na hranách

Skrytý Markovov model (HMM)



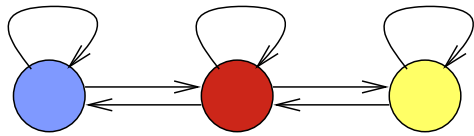
Predpokladajme, že model vždy začína v modrom stave.

Príklad:

$$\Pr(\text{aca}) = 0.27 \cdot 0.001 \cdot 0.27 \cdot 0.99 \cdot 0.24 = 0.000017$$

$$\Pr(\text{aca}) = 0.27 \cdot 0.999 \cdot 0.23 \cdot 0.999 \cdot 0.27 = 0.017$$

Matematické označenie



Sekvencia S_1, \dots, S_n










Anotácia A_1, \dots, A_n

Parametre modelu:

Prechodová pravdepodobnosť $a(u, v) = \Pr(A_{i+1} = v | A_i = u)$,

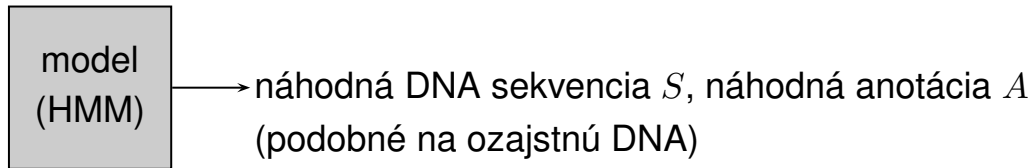
Emisná pravdepodobnosť $e(u, x) = \Pr(S_i = x | A_i = u)$,

Počiatočná pravdepodobnosť $\pi(u) = \Pr(A_1 = u)$.

| a |  |  |  | e | a | c | g | t |
|---|---|---|---|---|------|------|------|------|
|  | 0.99 | 0.007 | 0.003 |  | 0.24 | 0.27 | 0.28 | 0.21 |
|  | 0.01 | 0.99 | 0 |  | 0.26 | 0.22 | 0.22 | 0.30 |
|  | 0.001 | 0 | 0.999 |  | 0.27 | 0.23 | 0.23 | 0.27 |

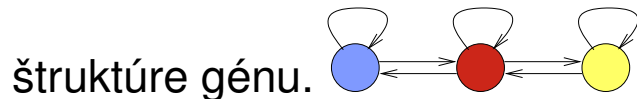
Výsledná pravdepodobnosť: $\Pr(A_1, \dots, A_n, S_1, \dots, S_n) = \pi(A_1)e(A_1, S_1) \prod_{i=2}^n a(A_{i-1}, A_i)e(A_i, S_i)$

Hľadanie génov s HMM



$\Pr(S, A)$ – pravdepodobnosť, že model vygeneruje pár (S, A) .

- **Určenie stavov a prechodov v modeli:** ručne, na základe poznatkov o



- **Tréovanie parametrov:** emisné a prechodové pravdepodobnosti určíme na základe sekvencií so známymi génmi (**trénovacia množina**).

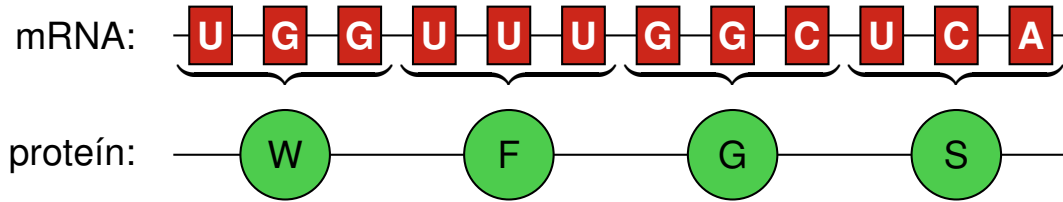
- **Použitie:** pre novú sekvenciu S nájdí najpravdepodobnejšiu anotáciu

$$A = \arg \max_A \Pr(A|S)$$

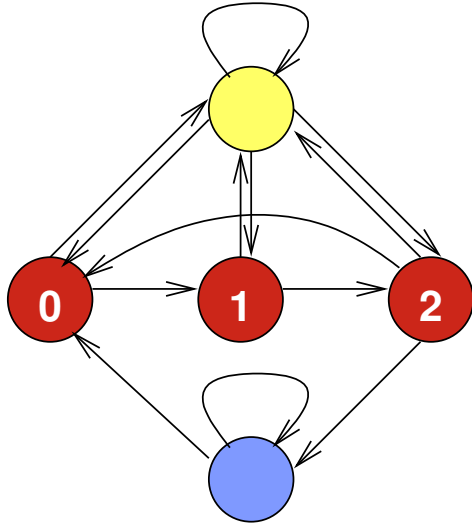
Viterbiho algoritmus v čase $O(nm^2)$ (dynamické programovanie)

HMM na hľadanie génov: 3-periodické exóny

Kodón (trojica báz) → jedna aminokyselina



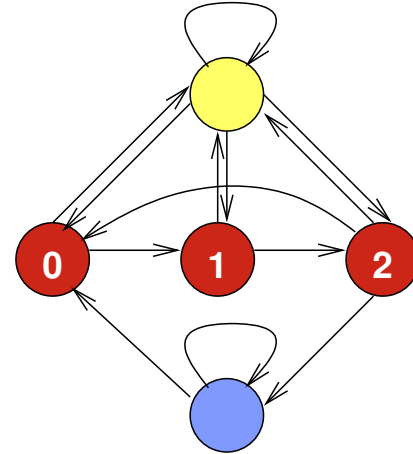
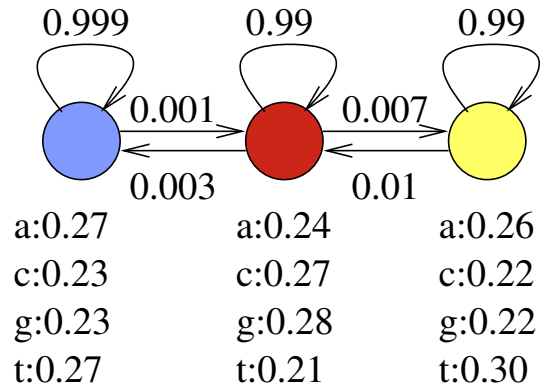
Namiesto jedného stavu pre exón použijeme tri stavy v cykle.



| a | 0 | 1 | 2 | Yellow | Blue |
|----------|----------|----------|----------|--------|------|
| 0 | 0 | | 0 | | 0 |
| 1 | 0 | 0 | | | 0 |
| 2 | | 0 | 0 | | |
| Yellow | | | | | 0 |
| Blue | | 0 | 0 | 0 | |

$\Pr(A_i|A_{i-1})$

Nové stavy mají odlišné emisné pravdepodobnosti

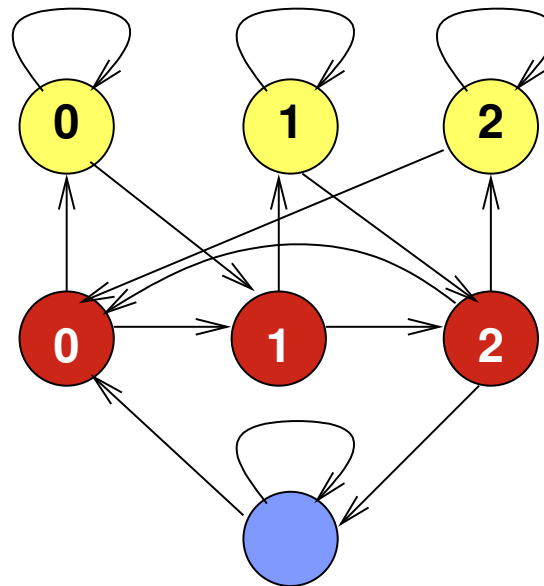
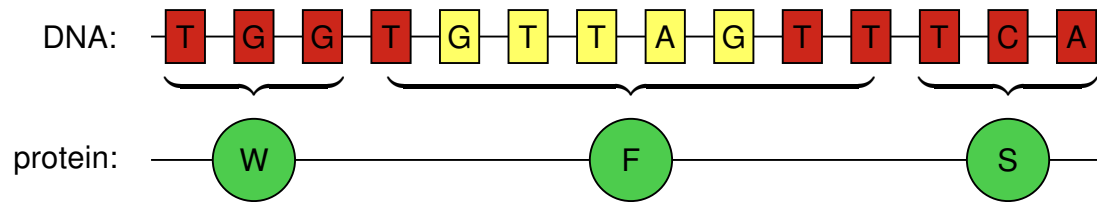


| <i>e</i> | a | c | g | t |
|----------|------|------|------|------|
| ■ | 0.24 | 0.27 | 0.28 | 0.21 |
| ■ | 0.26 | 0.22 | 0.22 | 0.30 |
| ■ | 0.27 | 0.23 | 0.23 | 0.27 |

| <i>e</i> | a | c | g | t |
|----------|------|------|------|------|
| 0 | 0.26 | 0.26 | 0.32 | 0.16 |
| 1 | 0.30 | 0.24 | 0.20 | 0.26 |
| 2 | 0.17 | 0.32 | 0.31 | 0.20 |
| ■ | 0.26 | 0.22 | 0.22 | 0.30 |
| ■ | 0.27 | 0.23 | 0.23 | 0.27 |

HMM na hľadanie génov: konzistentné kodóny

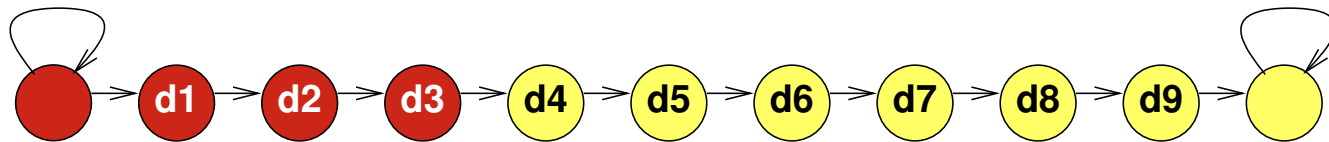
Intrón môže prerušiť kodón uprostred, chceme pokračovať, kde sme prestali.



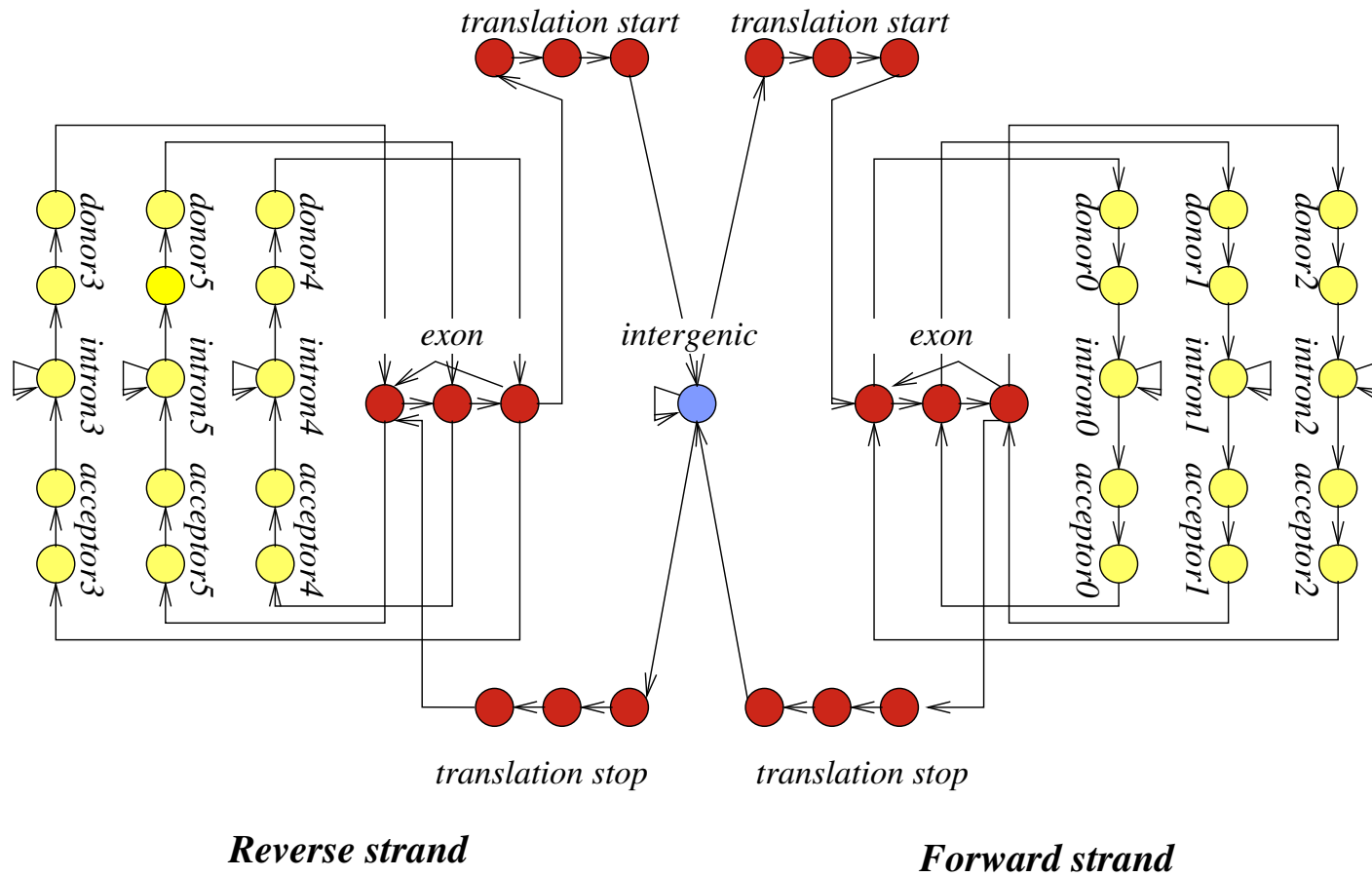
HMM na hľadanie génov: signály



Pridaj sériu stavov medzi exón a intrón:





HMM na hľadanie génov: celkový model



Stavy vyšších rádov

Rád 0: emisná tabuľka e určuje $\Pr(S_i|A_i)$

Rád 1: e určuje $\Pr(S_i|A_i, S_{i-1})$

| A_i | S_{i-1} | a | c | g | t |
|---|-----------|------|------|------|------|
|  | a | 0.24 | 0.23 | 0.34 | 0.19 |
| | c | 0.30 | 0.31 | 0.13 | 0.26 |
| | g | 0.27 | 0.28 | 0.28 | 0.17 |
| | t | 0.13 | 0.28 | 0.38 | 0.21 |
|  | a | 0.30 | 0.18 | 0.27 | 0.25 |
| | c | 0.32 | 0.28 | 0.06 | 0.35 |
| | g | 0.27 | 0.22 | 0.27 | 0.24 |
| | t | 0.20 | 0.21 | 0.26 | 0.33 |

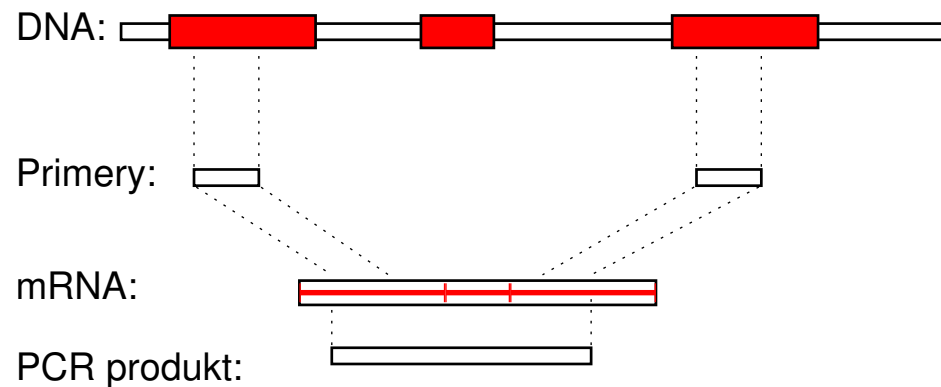
...

Na charakterizovanie exónov, intrónov atď používame rád 4-5.

Experimentálne overovanie génov

Overenie transkripcie a zotrihu

- **RNA-Seq:** sekvenovanie častí mRNA extrahovaných z bunky. Nie je cielené na konkrétny gén.
- **RT PCR:** cielené over konkrétny predpovedaný gén pomocou špecifických primerov.



Problémy: ťažko nájsť gény s expresiou iba za zvláštnych podmienok (napr. v embryu), kontaminácia genómovou DNA, nejednoznačné namapovanie na genóm.

Experimentálne overovanie génov

Overenie translácie, prítomnosti proteínu

- Hmotnostná spektrometria (mass spectrometry) dokáže detegovať prítomnosť proteínu izolovaného napr. z 2D gélu.
- Metódy založené na protilátkach (antibody), prípadne špecifické techniky podľa typu proteínu.

Príklady programov na hľadanie génov

Len na základe sekvencie DNA:

HMMGene [Krogh 1997] (autor je priekopníkom HMM v bioinf.),
Genscan [Burge a Karlin 1997] (po mnohé roky štandard),
GeneZilla [Majoros a kol. 2004], ExonHunter [Brejová a kol. 2005], Augustus [Stanke a Waack 2003] (novšie programy založené na zovšeobecnených HMM).
CONTRAST [Gross 2007], CONRAD [DeCaprio 2007] (programy založené na conditional random fields, obmena HMM)

Prokaryotické genómy:

GeneMark [Lukashin a Borodovsky 1998], Glimmer [Delcher a kol. 1999] a ďalšie.

Vybrané programy na hľadanie génov

Porovnávaním viacerých sekvencií:

Twinscan [Korf a kol. 2001]

(prvý úspešný gene finder s dvoma genómami),

Exoniphy [Siepel a Haussler 2004]

(viacero genómov, nehľadá celé gény),

N-SCAN [Gross a Brent 2006]

(rozšírenie Twinscanu na viacero genómov).

Iná informácia: (napr. RNA-seq, príbuzné proteíny a pod.)

ExonHunter [Brejová a kol. 2005], Augustus [Stanke a kol. 2006],

Jigsaw [Allen a Salzberg 2005], Fgenes++ [Solovyev 2006].

Augustus patrí dodnes medzi často používané programy.

Obmedzenia hľadacov génov

- Alternatívny zostrih (alternative splicing): jeden gén môže vyprodukovať viacero mRNA molekúl. Programy väčšinou hľadajú iba jednu.

Retained intron:



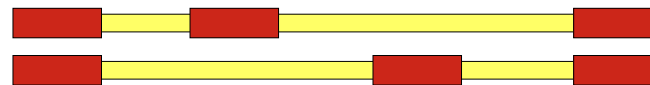
Skipped exon:



Alternative donor or acceptor:



Mutually exclusive exons:



- Pretínajúce sa gény, resp. gény v intrónoch.
- Netypické gény (neobvyklé signály, veľmi krátke alebo dlhé exóny alebo intróny atď.)
- Hľadanie UTR a začiatku/konca transkripcie.

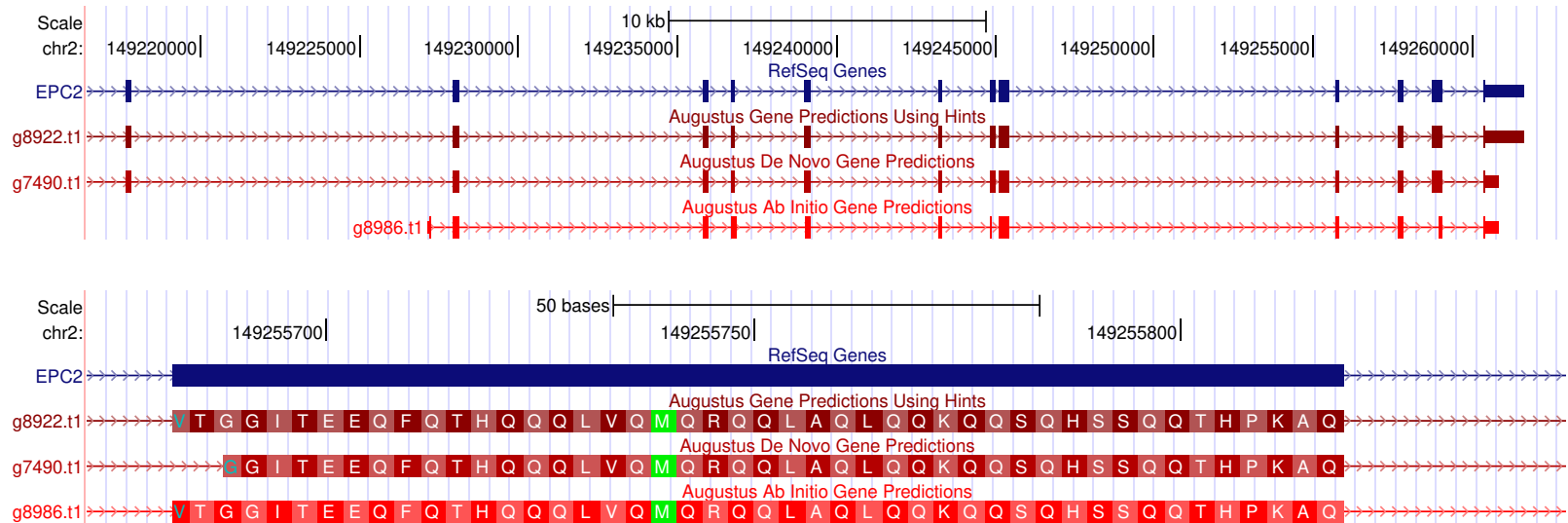
Hľadače génov robia často chyby

Najlepšie metódy v 2005 na ľudskom genóme: [Guigo et al 2006]

20% génov, 60% exónov správne iba na základe DNA

35% génov, 65% exónov správne komparatívne

70% génov, 85% exónov správne s ďalšou informáciou



Koľko g3nov m3 3lovek?

Do 2001: R3zne odhady: **50 000–140 000** g3nov

2001: predbeŹn3 verzia ľudsk3ho gen3mu: **30 000–40 000** g3nov

2004: sekvencia ľudsk3ho gen3mu: **20 000–25 000** g3nov

2007: v katal3goch Ensembl, RefSeq a VEGA spolu **24 500** g3nov

[Clamp a kol. 2007] tvrdia, Źe iba **20 500** z nich je spr3vnych

Ale s3 g3ny, o ktor3ch eŹte nevieme?

2010: RefSeq m3 **22 333** g3nov

St3le neistota ± 1000 [Perteau, Salzberg 2010]

R3zni ľudia sa m3Źu l3iŹ v desiatkach g3nov

2012: Projekt ENCODE odhaduje **20 687** g3nov k3duj3cich prote3ny,

v priemere 6 altern3tvnych transkriptov na g3n,

plus 8 800 kr3tkych a 9 600 dlh3ch RNA g3nov

Zhrnutie

- Novo osekvenované genómy treba anotovať:
určovať funkcie jednotlivým oblastiam sekvencie
- Príkladom anotácie je hľadanie génov kódujúcich proteíny
- Na hľadanie génov sa hodia skryté Markovove modely
- Modely robia veľa chýb, ale dajú nám základnú predstavu o polohe a počte génov, môžeme študovať ich funkciu

Journal club: skupiny

- Zoznam skupín na stránke predmetu, každej skupine tiež príde spoločný email, aby ste sa vedeli navzájom nakontaktovať
- Niektoré skupiny sú anglicky hovoriace

Journal club: prvé stretnutie - pred 22.11.

- Každý si najprv prečíta článok, potom sa koná stretnutie, kde o článku diskutujete, vysvetlíte si navzájom nejasnosti, plánujete písanie správy
- Dátum, čas a miesto oznámte aspoň 24 hodín vopred v diskusii na Moodli
- Po stretnutí napíšte krátku správu zo stretnutia do príslušnej položky v Moodli (kto sa zúčastnil, čo sa dohodlo, či sú nejaké problémy, stačí pár viet)
- Ak treba, dohodnite si s nami konzultácie

Správa zo journal clubu

- Vlastnými slovami hlavné metódy a výsledky článku
- Pochopiteľná pre študentov tohto predmetu (inf aj bio)
- Vysvetlite pojmy, ktoré sú nad rámec tohto predmetu
- Netreba pokryť všetko, môžete využiť aj iné zdroje
- Podrobne vysvetliť aspoň jednu bioinformatickú metódu a aspoň jeden biologický výsledok (alebo overovanie správnosti metódy na dátach)
- Ako článok súvisí s učivom preberaným na predmete
- Nájdite zopár citujúcich prác, ktoré výsledky využili alebo vylepšili
- Rozsah cca 1-2 strany na osobu, jeden ucelený text
- V správe vymenujte členov skupiny, ktorí sa podieľali na jej spísaní, dostanú rovnako bodov