Metódy v bioinformatike

CB05: HMM, E-value

Jana Černíková

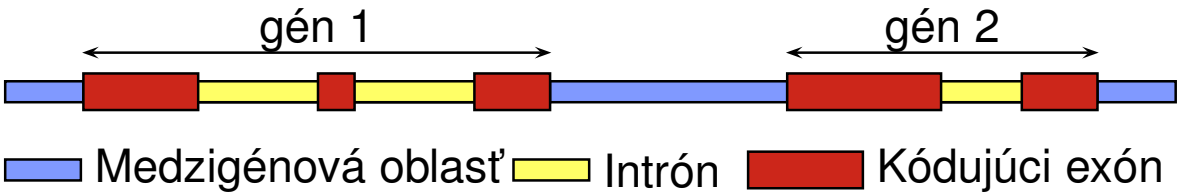
FMFI UK

23/10/2025

Bioinformatický problém: Hľadanie génov

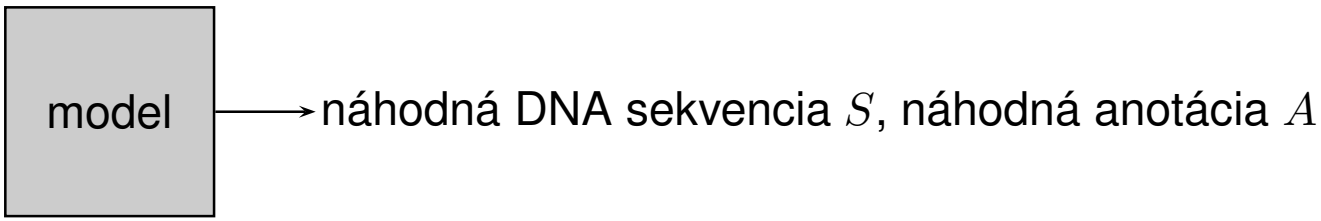
Vstup: DNA sekvencia
Cieľ: označ každú bázu ako intrón/exón/medzigénovú oblasť (anotácia)
Výstup: anotácia s maximálnym skóre
(segmentácia pôvodnej sekvencie na neprekrývajúce sa regióny, ktoré reprezentujú intróny, exóny a medzigénové úseky, pre ktorú dostaneme max. skóre na základe pravdepodobnostného modelu)

cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac



Pravdepodobnostný model génov

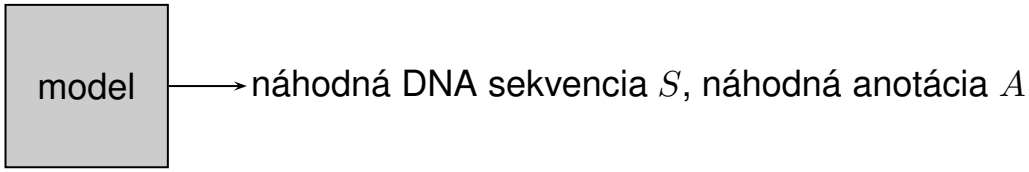
Žiadna informácia nám neumožňuje jednoznačne určiť, čo je gén.
Skombinujeme dostupnú informáciu pravdepodobnostným modelom.



$\Pr(S, A)$ – 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 nájdi najpravdepodobnejšiu anotáciu $A = \arg\max_A \Pr(A|S)$

Pravdepodobnostný model génov



Použitie: pre sekvenciu S nájdi najpravdepodobnejšiu anotáciu A

Hračkársky príklad modelu: sekvencie dĺžky 2

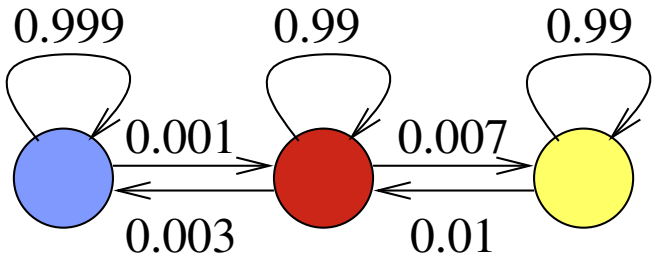
Tabuľka pravdepodobností pre 16 sekvencií, 9 anotácií (súčet 1)

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)

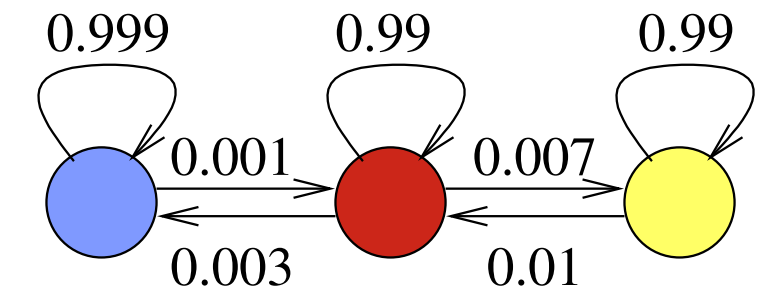
Spôsob, ako zadefinovať model pre dlhšie sekvencie.



a:0.27	a:0.24	a:0.26
c:0.23	c:0.27	c:0.22
g:0.23	g:0.28	g:0.22
t:0.27	t:0.21	t:0.30

- 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:

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

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

Príklady stavových automatov pre HMM

Uvažujme HMM
so špeciálnym začiatočným stavom b a koncovým stavom e,
ktoré nič negenerujú.

Príklady stavových automatov pre HMM

Uvažujme HMM
so špeciálnym začiatočným stavom b a koncovým stavom e,
ktoré nič negenerujú.

- **Úloha 1:** Nakreslite HMM (stavový diagram), ktorý generuje sekvencie, ktoré začíajú niekoľkými červenými písmenami a potom obsahujú niekoľko modrých

Príklady stavových automatov pre HMM

Uvažujme HMM
so špeciálnym začiatočným stavom b a koncovým stavom e,
ktoré nič negenerujú.

- **Úloha 1:** Nakreslite HMM (stavový diagram), ktorý generuje sekvencie, ktoré začíajú niekoľkými červenými písmenami a potom obsahujú niekoľko modrých
- **Úloha 2:** Ako treba zmeniť HMM, aby dovoloval ako "niekoľko" aj nula?

Príklady stavových automatov pre HMM

Uvažujme HMM
so špeciálnym začiatočným stavom b a koncovým stavom e,
ktoré nič negenerujú.

- **Úloha 1:** Nakreslite HMM (stavový diagram), ktorý generuje sekvencie, ktoré začíajú niekoľkými červenými písmenami a potom obsahujú niekoľko modrých
- **Úloha 2:** Ako treba zmeniť HMM, aby dovoloval ako "niekoľko" aj nula?
- **Úloha 3:** Ako treba zmeniť HMM, aby počet červených aj modrých bol vždy parne číslo?

Príklady stavových automatov pre HMM

Uvažujme HMM
so špeciálnym začiatočným stavom b a koncovým stavom e,
ktoré nič negenerujú.

- **Úloha 1:** Nakreslite HMM (stavový diagram), ktorý generuje sekvencie, ktoré začíajú niekoľkými červenými písmenami a potom obsahujú niekoľko modrých
- **Úloha 2:** Ako treba zmeniť HMM, aby dovoloval ako "niekoľko" aj nula?
- **Úloha 3:** Ako treba zmeniť HMM, aby počet červených aj modrých bol vždy parne číslo?
- **Úloha 4:** Ako zmeniť HMM, aby sa striedali červené a modré kusy párnej dĺžky?

Príklady stavových automatov pre HMM

V ďalších príkladoch uvažujeme aj to, ktoré písmená su v ktorom stave povolené (pravdepodobnosť emisie > 0) a ktoré sú zakázané

Príklady stavových automatov pre HMM

V ďaľších príkladoch uvažujeme aj to, ktoré písmená su v ktorom stave povolené (pravdepodobnosť emisie > 0) a ktoré sú zakázané

- **Úloha 5:** Model generujúci červené sekvencie dĺžky dva, ktoré začínajú na A

Príklady stavových automatov pre HMM

V ďaľších príkladoch uvažujeme aj to, ktoré písmená su v ktorom stave povolené (pravdepodobnosť emisie > 0) a ktoré sú zakázané

- **Úloha 5:** Model generujúci červené sekvencie dĺžky dva, ktoré začínajú na A
- **Úloha 6:** Model generujúci červené sekvencie dĺžky dva, ktoré môzu byť čokoľvek iné ako AA

Príklady stavových automatov pre HMM

V ďalších príkladoch uvažujeme aj to, ktoré písmená su v ktorom stave povolené (pravdepodobnosť emisie > 0) a ktoré sú zakázané

- **Úloha 5:** Model generujúci červené sekvencie dĺžky dva, ktoré začínajú na A
- **Úloha 6:** Model generujúci červené sekvencie dĺžky dva, ktoré môzu byť čokoľvek iné ako AA
- **Úloha 7:** Rozšírte predošlý model na sekvencie dĺžky 3 bázy, tak aby to nemohli byť stop kodóny TAA, TAG, TGA

Príklady stavových automatov pre HMM

V ďalších príkladoch uvažujeme aj to, ktoré písmená su v ktorom stave povolené (pravdepodobnosť emisie > 0) a ktoré sú zakázané

- **Úloha 5:** Model generujúci červené sekvencie dĺžky dva, ktoré začínajú na A
- **Úloha 6:** Model generujúci červené sekvencie dĺžky dva, ktoré môzu byť čokoľvek iné ako AA
- **Úloha 7:** Rozšírte predošlý model na sekvencie dĺžky 3 bázy, tak aby to nemohli byť stop kodóny TAA, TAG, TGA

  toto sa dá rozšíriť na HMM, ktorý reprezentuje ORF (open reading frame): začína štart kodónom, potom niekoľko bežných kodónov, ktoré nie sú stop kodónom a na koniec stop kodón

Iný príklad použitia HMM: Topológia transmembránových proteínov

Chceme označiť aminokyseliny v proteíne – vonku z bunky, v membráne, vo vnútri bunky
Nie každá postupnosť označení dáva zmysel – napr. vonku—vnútri alebo vonku—v_membráne—vonku

Iný príklad použitia HMM: Topológia transmembránových proteínov

Chceme označiť aminokyseliny v proteíne – vonku z bunky, v membráne, vo vnútri bunky
Nie každá postupnosť označení dáva zmysel – napr. vonku—vnútri alebo vonku—v_membráne—vonku

Čo by reprezentovali stavy v tomto prípade?

E-value: Hračkársky prípad

Dotaz: ATGCTCAAAC (dĺžka $m = 10$)
Databáza: (dĺžka $n = 300$)

accacttgcgcacgatttccagattcggtttccctgggcgcacgaagggc
ccacgaagcgGCTCAACccggagccttagttagaaggggggtctccgtca
agagagacggtaagttggagggtcactagcggtggactccgaatggaaac
actgaatagtggcagaacctaaacctcgttttggatttcctgaaaaaggc
aggcgctagaggaagaggcacgactgtgctagagataatcacttgtaaga
ccttggggatgggcttcgtatgcagaacgcgataaggtatcgaaaacgtg

Skórovacia schéma: zhoda $+1$, nezhoda -1, medzera -1
Lokálne zarovnanie so skóre $S = 6$

GCTCAAAC
GCTCA-AC

E-value: koľko očakávame lokálnych zarovnaní so skóre aspoň S v náhodnej databáze dĺžky n pri náhodnom dotaze dĺžky m

Náhodný dotaz a databáza

Dotaz: GTGCCTGCAG

Databáza:

cctctgatagccttgaaccgggcgagactcatacagacagtgctcctcgg
gcgataaccatgagatgacaggtccgatgctaatgttaacggacctacag
tgacatgttaaagtgtccattaagtttataccggaatcaacgagtgtccc
ccagcgcggcgaccgatggagccCCTGCAGgtatactcacttcaaggatt
accgctcggtgtaagttagtgttcagtcagactatactaagtattcagtt
atagagcgttagtaggtcgaccatgagcgggtaggGTGCCGAGatgtgaa

Počet výskytov: 2

Náhodný dotaz a databáza

Dotaz: `TCGACCGAAA`

Databáza:

```
tactccattagggattataacgactaaagcccgtcgtggcgggatcactt
tgagattcaactttaacgcatcacagaggaatctgagacaaagcaaaacc
gatcataatgatcgatccaggtaataagtctccttgatggcgttagactg
gaaataacagttgacttccgactatagtttaatgaacgttcgtaattaga
cgatcgtgtaacttaaccaaaggctgcccccaaactagctgagtaatagc
tcgtcctgagcatgtaagagtcagcctccacggaacactgcaacgttctt
```

Počet výskytov: 0

Náhodný dotaz a databáza

Dotaz: CCCGTCGTAG

Databáza:

cagcattagccccgttatttCGTCGTtctccaacgggtctgcctttctgg
aacgtggcgaaccttcacaggtcagtctgtcatcgcctgcgcttagagcg
gacggtactcgaaaggtcggttcagtgtggcgctggaaagaagaatagca
acacatgcactaatggaaggtcccagtggtgtgggacattctggaCCCGT
GTgtgccaacctatgtgagctccggcgttgactcggaggatgttaacaag
atcaagctgtaggcgacgatccccgccgggtttcctctactgcctcgagc

Počet výskytov: 2

Náhodný dotaz a databáza

Dotaz: AGGATGAGGA

Databáza:

ttatcgattctccggtgcgccagtacagcacaaggctcggatcctgtaaa
acactacaccttaaaaactaagtcAGGATGtgatctcccttaaGATGAGa
cagtctctaatgcggcgtagtgggaccctcgtgaccgagctaagcagttc
acaatgggcgctctgagcgattggctggagaccttgacttcccggtaggt
gtggtgttagttctgtgcccagagataaccatccaccgtaatggatctcg
taactttacGATGAAGAccggcatcatctcagttatatttctaggacggg

Počet výskytov: 3

Celkovo opakujeme 100 krát

$S = 6$, $m = 10$, $n = 300$, obsah GC 50%

Počet výskytov: 2, 0, 2, 3, 3, 1, 0, 1, 1, 1, 0, 0, 4, 2, 0, 1, 0, 1, 0, 0, 1, 0, 0, 4, 3, 1, 1, 0, 0, 0, 2, 3, 0, 0, 2, 1, 1, 1, 0, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, 1, 2, 2, 2, 0, 0, 2, 0, 1, 1, 0, 1, 2, 2, 1, 0, 0, 1, 1, 2, 0, 1, 0, 0, 1, 0, 3, 2, 0, 2, 2, 1, 0, 0, 2, 0, 0, 1, 2, 1, 1, 3, 2, 2, 1, 1, 0, 2, 0, 1, 3

Priemerný počet výskytov: 1.05

Keď celé opakujeme viackrát, dostávame hodnoty 0.99, 1.15, 1.02, 1.07, 0.98, . . .

Správna hodnota **E-value:** 0.99