Oznamy

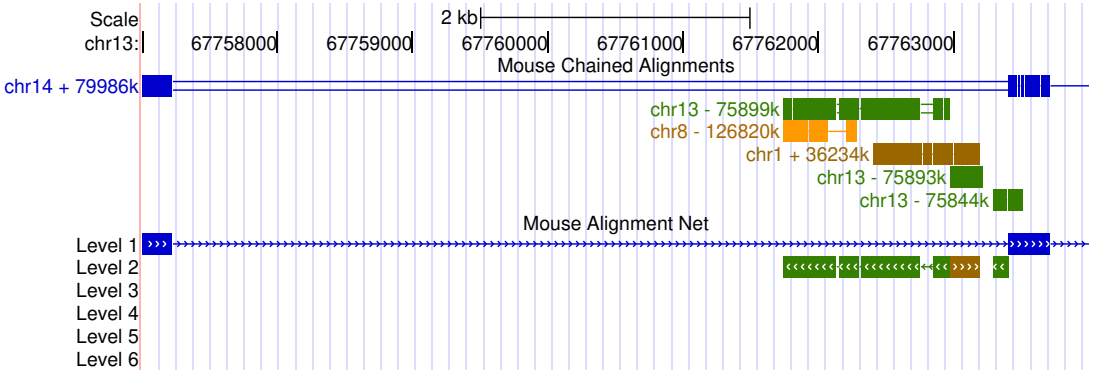
- Výber článku na journal club formulárom na stránke do budúcej stredy 19.10. 22:00.
- Domáca úloha 1 je zverejnená na stránke, odovzdávajte do utorka 8.11. (pdf cez Moodle)
- Domáce úlohy neodpisujte. Môžete sa rozprávať, ale nerobte si pritom poznámky, neukazujte si navzájom svoje riešenia. Každý by mal napísať riešenie samostatne.
- Otázky k zadaniam a všeobecnú diskusiu k predmetu píšte do MS Teams (kanál General), môžete spolužiakom aj odpovedať. Otázky k vašim riešeniam posielajte vyučujúcim e-mailom alebo cez MS Teams.
- Nezabudnite na pravidelné kvízy, ak je niečo nejasné, pýtajte sa.

Zarovnávanie sekvencií 2/2

(sequence alignment)

Askar Gafurov

13.10.2022



Zhrnutie z minulej prednášky

- **Problém globálneho a lokálneho zarovnania**

  Vstup: sekvencie $X = x_1 x_2 \ldots x_n$ a $Y = y_1 y_2 \ldots y_m$.

  Výstup: zarovnanie X a Y s najvyšším skóre

  resp. zarovnania podreťazcov $x_i \ldots x_j$ a $y_k \ldots y_\ell$ s najvyšším skóre.

- **Správny algoritmus na riešenie**

  dynamické programovanie

- **Realistické skórovacie schémy**

Máme správny algoritmus na zarovnávanie, čo viac nám chýba?

Časová zložitosť: $O(nm)$ na sekvenciách dĺžky n a m.

Koľko je to času v skutočnosti?

(jednoduchá implementácia, náhodné sekvencie dĺžky n,

bežný počítač)

n	čas výpočtu
100	0.0008s
1,000	0.08s
10,000	8s
100,000	13 minút (*)
1,000,000	22 hodín (*)
10,000,000	3 mesiace (*)
100,000,000	25 rokov (*)

Potrebujeme efektívnejší algoritmus,

najmä ak chceme pracovať s celými genómami

Pamäť: základný algoritmus $O(n^2)$, dá sa zlepšiť na $O(n)$.

Heuristické lokálne zarovnávanie

- Nie je zaručené, že nájdeme najlepšie zarovnanie, ale program pobeží rýchlejšie.
- Prehľadá iba “sľubné” časti dyn. prog. matice.

Napríklad: BLASTN [Altschul et al., 1990],
FASTA [Pearson and Lipman, 1988]

- Nájdi krátke zhodujúce sa úseky dĺžky w (jadrá zarovnania).
- Rozšír každé jadro pozdĺž uhlopriečky na zarovnanie bez medzier.
- Spoj zarovnania na neďalekých uhlopriečkach medzerami.
- Lokálne vylepši zarovnanie dynamickým programovaním (možno vynechať).

Ako nájdeme zhodujúce sa úseky?

- Vybudujeme “slovník” úsekov dĺžky w z prvej sekvencie.
- Nájdeme každý úsek z druhej sekvencie v slovníku.

Príklad: CAGTCCTAGA vs CATGTCATA

Slovník:

```
AG 2, 8
CA 1
CC 5
CT 6
GA 9
GT 3
TA 7
TC 4
```

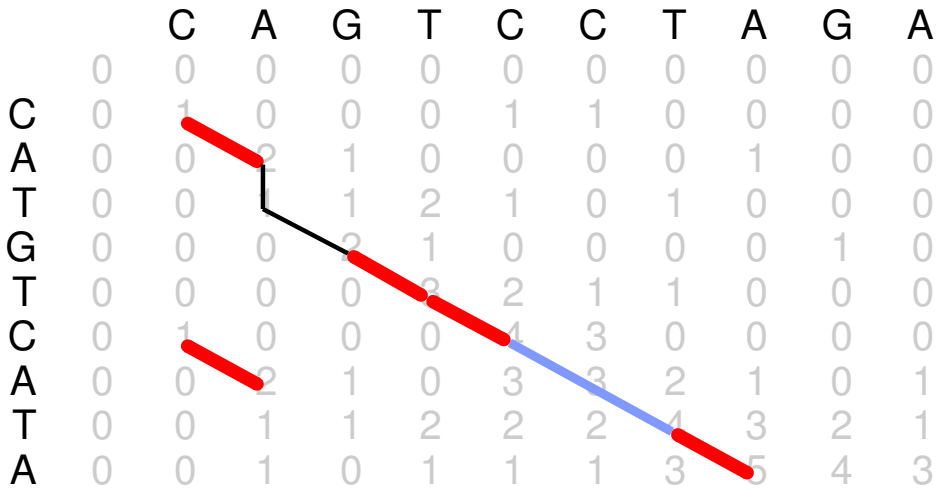
Hľadaj:

```
CA → 1
AT → -
TG → -
GT → 3
TC → 4
CA → 1
AT → -
TA → 7
```

Heuristické lokálne zarovnávanie

Príklad: začíname z jadier dĺžky $w = 2$

(V praxi sa používa $w = 10$ a viac.)



1. nájdi zhodné úseky
2. rozšír bez medzier
3. spoj medzerami

Rýchlosť heuristického algoritmu

Algoritmus:

- Nájdi jadrá zarovnaní (krátke zhodujúce sa úseky dĺžky w).
- **Drahý krok:** Rozširovanie/spájanie jadier do väčších zarovnaní.

Náhodné zhody dĺžky w: nie sú časťou zarovnania s vysokým skóre.

Vyfiltrujeme ich pri rozširovaní, ale spomaľujú program.

Koľko náhodných zhôd?

Dva nukleotidy sa zhodujú s pravdepodobnosťou $1/4$.

w zhôd za sebou s pravdepodobnosťou 4^{-w}.

Stredná hodnota počtu zhôd $nm4^{-w}$.

Zvýšenie w o 1 zníži počet zhôd cca 4 krát.

Senzitivita heuristického algoritmu

Algoritmus:

- Nájdi jadrá zarovnaní (krátke zhodujúce sa úseky dĺžky w).
- **Drahý krok:** Rozširovanie/spájanie jadier do väčších zarovnaní.

Nenájdené zarovnania: vysoké skóre, ale nemajú jadro dĺžky w

Príklad: `CA-GTCCTA` nenájdeme pre $w \geq 4$

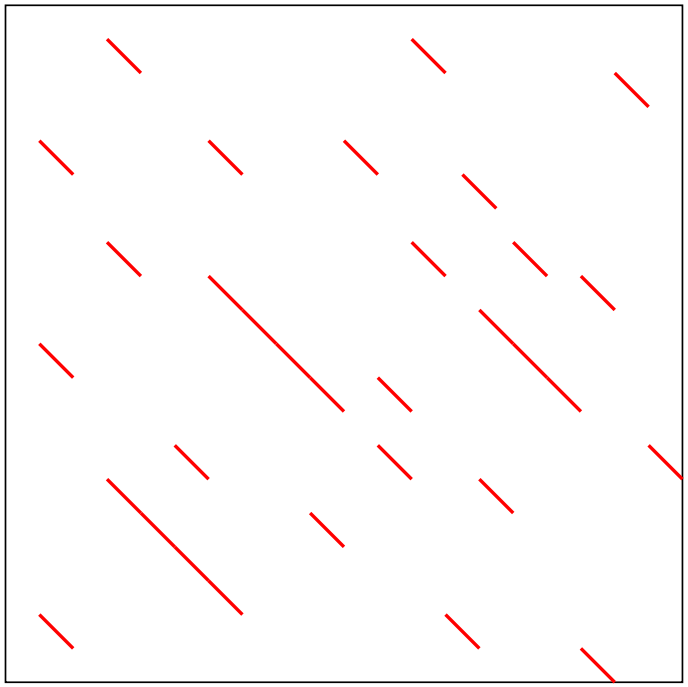
`CATGTCATA`

Senzitivita: aká časť skutočných zarovnaní obsahuje zhodu dĺžky w

Rýchlosť vs. senzitivita

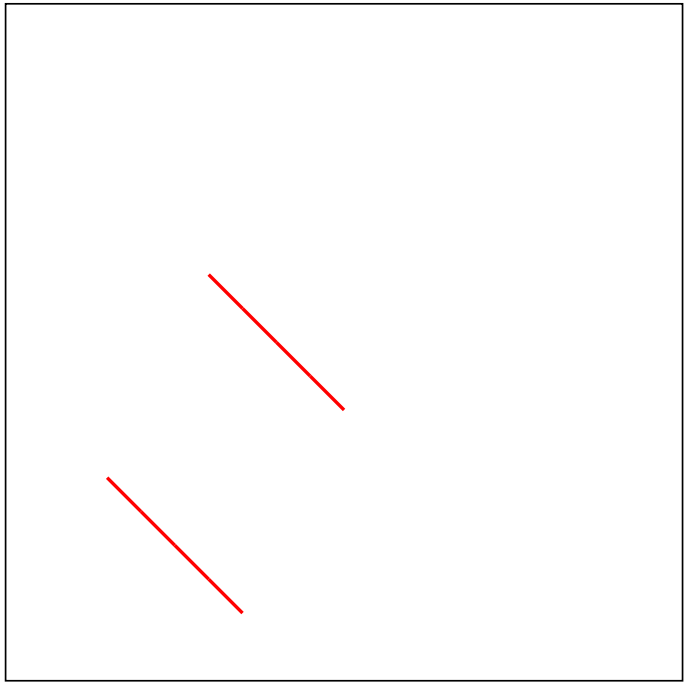
Malé w

veľa náhodných zhôd, pomalé



Veľké w

nenájdeme veľa zarovnaní



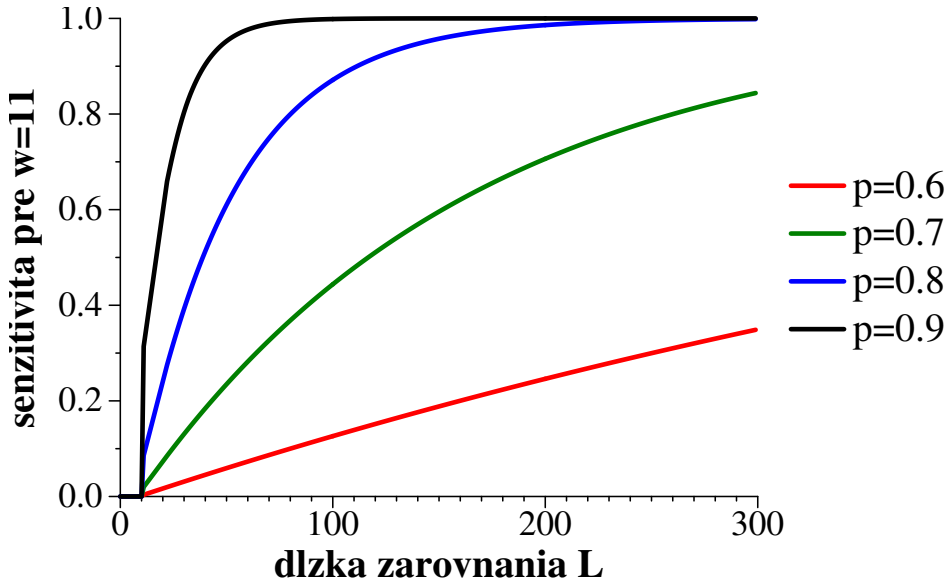
Senzitivita heuristického algoritmu

Odhad senzitivity:

Predpokladáme zarovnanie bez medzier, dĺžky L

Každá pozícia je zhoda s pravdepodobnosťou p

$f(L,p) = \Pr(\text{zarovnanie obsahuje } w \text{ zhôd za sebou})$



(človek-myš: $p \approx 0.7$)

BLAST algoritmus pre proteíny

BLOSUM62 skórovacia matica pre proteíny

```
   A  R  N  D  C  Q  E  G  H  I ...
A  4 -1 -2 -2  0 -1 -1  0 -2 -1
R -1  5  0 -2 -3  1  0 -2  0 -3
N -2  0  6  1 -3  0  0  0  1 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1
Q -1  1  0  0 -3  5  2 -2  0 -3
E -1  0  0  2 -4  2  5 -2  0 -3
G  0 -2  0 -1 -3 -2 -2  6 -2 -4
H -2  0  1 -1 -3  0  0 -2  8 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
```

Proteínový BLAST namiesto zhody dĺžky w vyžaduje 3 aminokyseliny so skóre aspoň 13

Áno:
```
N I R
N L R

6+2+5=13
```

Nie:
```
A I L
A I L

4+4+4=12
```

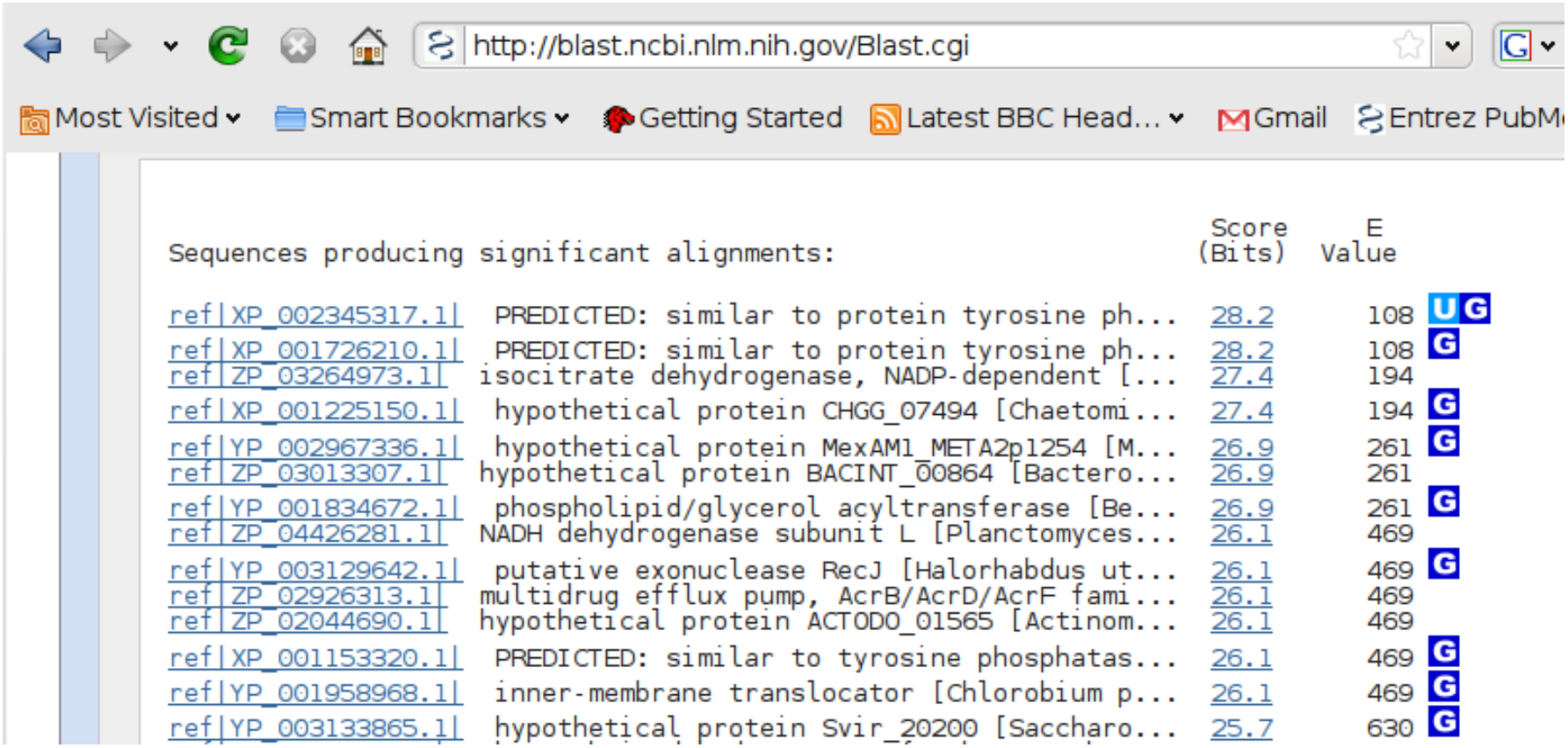
Príklady programov

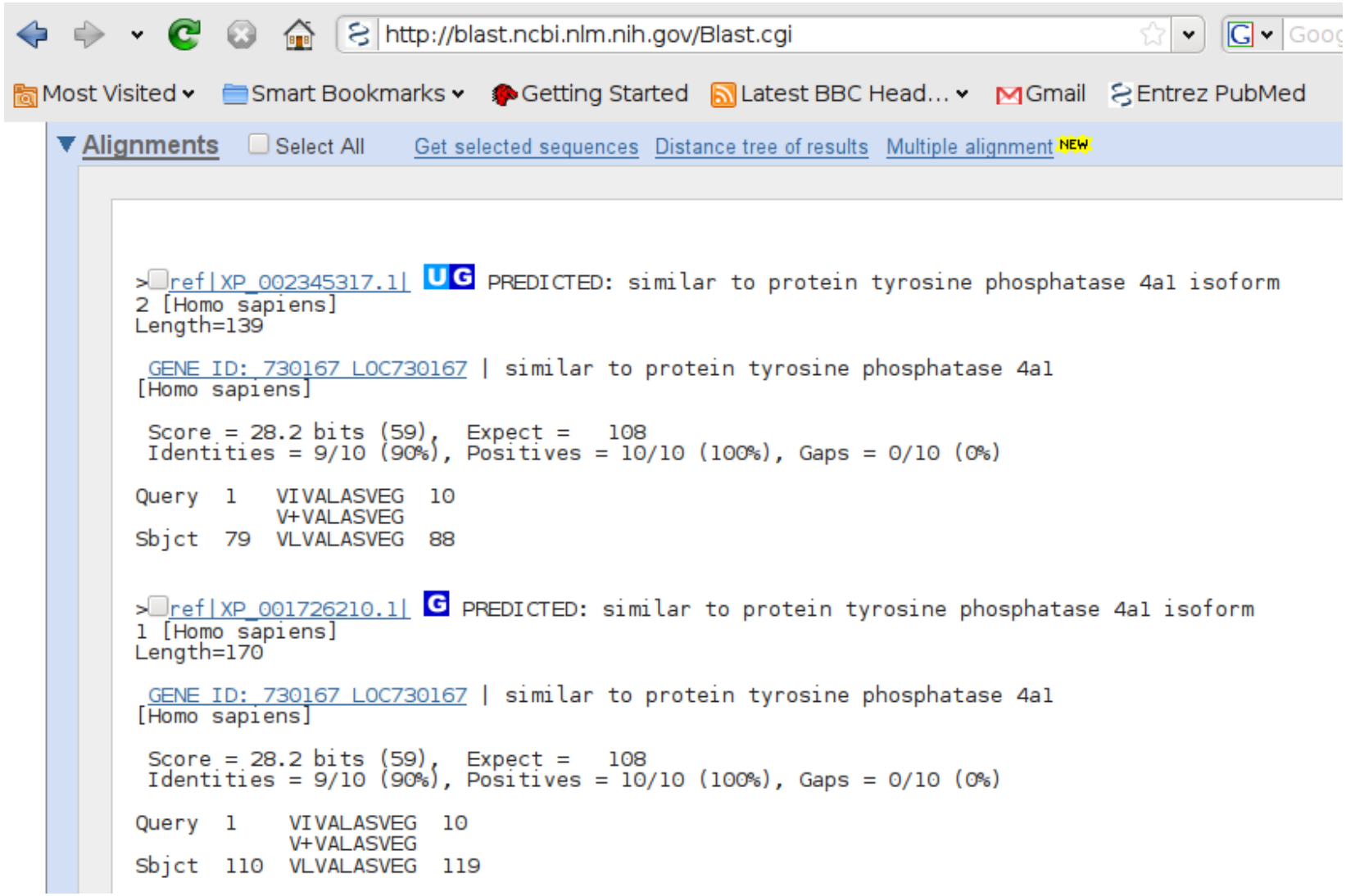
NCBI BLAST: blastn pre DNA/RNA, blastp pre proteíny,

tblastx preloží DNA do proteínu a použije blastp

UCSC Blat: veľmi rýchle vyhľadávanie veľmi podobných sekvencií, napr. sekvenačné čítania ku genómu

- používa veľké w
- vie nájsť zarovnania s veľkými medzerami (napr. intróny pri mRNA)


http://blast.ncbi.nlm.nih.gov/Blast.cgi
Most Visited
Smart Bookmarks
Getting Started
Latest BBC Head...
Gmail
Entrez PubM
Sequences producing significant alignments:
Score (Bits)
E Value
ref|XP_002345317.1| PREDICTED: similar to protein tyrosine ph... 28.2 108
ref|XP_001726210.1| PREDICTED: similar to protein tyrosine ph... 28.2 108
ref|ZP_03264973.1| isocitrate dehydrogenase, NADP-dependent [... 27.4 194
ref|XP_001225150.1| hypothetical protein CHGG_07494 [Chaetomi... 27.4 194
ref|YP_002967336.1| hypothetical protein MexAM1_META2p1254 [M... 26.9 261
ref|ZP_03013307.1| hypothetical protein BACINT_00864 [Bactero... 26.9 261
ref|YP_001834672.1| phospholipid/glycerol acyltransferase [Be... 26.9 261
ref|ZP_04426281.1| NADH dehydrogenase subunit L [Planctomyces... 26.1 469
ref|YP_003129642.1| putative exonuclease RecJ [Halorhabdus ut... 26.1 469
ref|ZP_02926313.1| multidrug efflux pump, AcrB/AcrD/AcrF fami... 26.1 469
ref|ZP_02044690.1| hypothetical protein ACTODO_01565 [Actinom... 26.1 469
ref|XP_001153320.1| PREDICTED: similar to tyrosine phosphatas... 26.1 469
ref|YP_001958968.1| inner-membrane translocator [Chlorobium p... 26.1 469
ref|YP_003133865.1| hypothetical protein Svir_20200 [Saccharo... 25.7 630


http://blast.ncbi.nlm.nih.gov/Blast.cgi
Most Visited
Smart Bookmarks
Getting Started
Latest BBC Head...
Gmail
Entrez PubMed
Alignments
Select All
Get selected sequences
Distance tree of results
Multiple alignment
NEW
>ref|XP_002345317.1| PREDICTED: similar to protein tyrosine phosphatase 4a1 isoform 2 [Homo sapiens]
Length=139
GENE ID: 730167 LOC730167 | similar to protein tyrosine phosphatase 4a1 [Homo sapiens]
Score = 28.2 bits (59), Expect = 108
Identities = 9/10 (90%), Positives = 10/10 (100%), Gaps = 0/10 (0%)
Query 1 VIVALASVEG 10
V+VALASVEG
Sbjct 79 VLVALASVEG 88
>ref|XP_001726210.1| PREDICTED: similar to protein tyrosine phosphatase 4a1 isoform 1 [Homo sapiens]
Length=170
GENE ID: 730167 LOC730167 | similar to protein tyrosine phosphatase 4a1 [Homo sapiens]
Score = 28.2 bits (59), Expect = 108
Identities = 9/10 (90%), Positives = 10/10 (100%), Gaps = 0/10 (0%)
Query 1 VIVALASVEG 10
V+VALASVEG
Sbjct 110 VLVALASVEG 119

Ako rozlíšiť, či ide o významné zarovnanie?

Dĺžka dotazu m. Veľkosť databázy n.

Zarovnanie so skóre S.

P-hodnota: Pravdepodobnosť, že pre náhodný dotaz dĺžky m v náhodnej databáze dĺžky n nájdeme zarovnanie so skóre aspoň S.

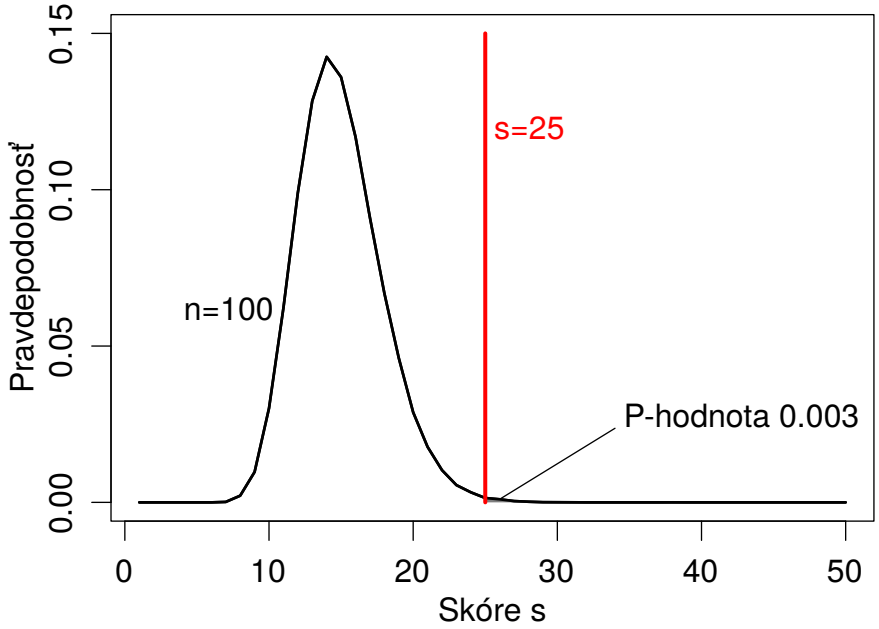
E-hodnota: Očakávaný počet zarovnaní so skóre aspoň S nájdených pre náhodný dotaz dĺžky m v náhodnej databáze dĺžky n.

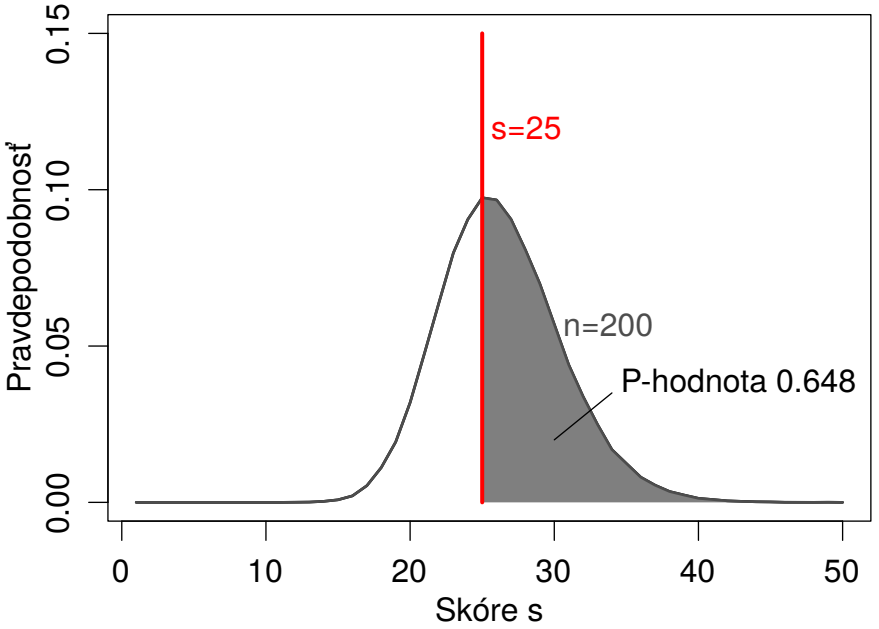
Pri veľmi malých hodnotách sú E-hodnota a P-hodnota takmer identické.

[Karlin and Altschul, 1990, Dembo et al., 1994]

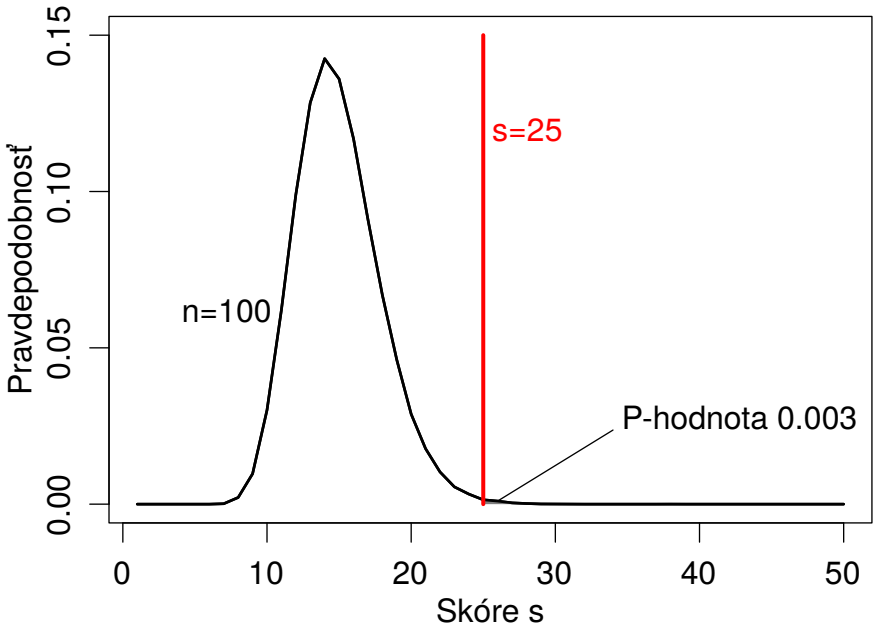
Výpočet P-hodnoty simuláciou

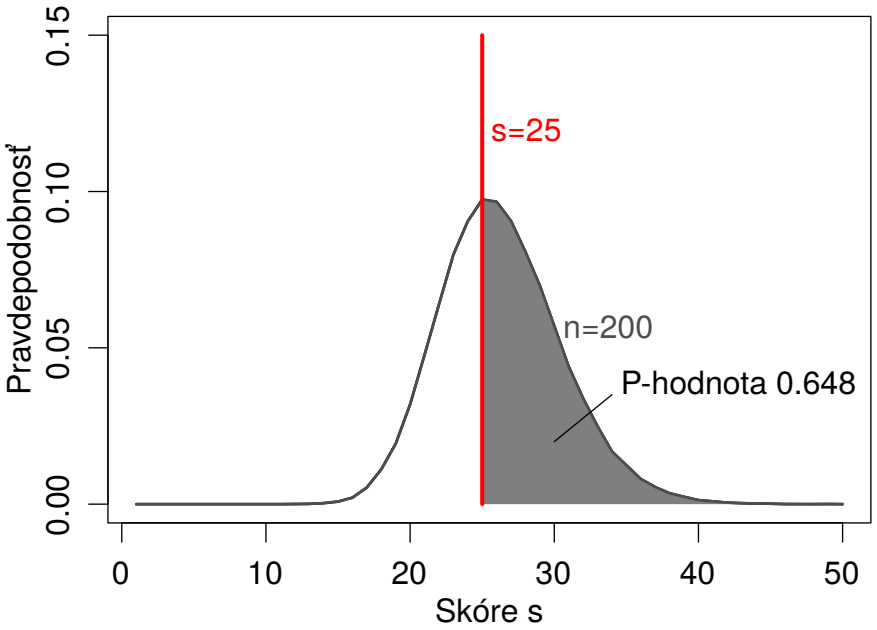
- Vygenerujeme náhodne dve sekvencie dĺžky n
- Spočítame ich najlepšie lokálne zarovnanie (schéma +1/-1)
- Zaznamenáme si výsledné skóre
- Opakujeme veľa krát





Výpočet P-hodnoty simuláciou (pokr.)





P-hodnota pre skóre 25:

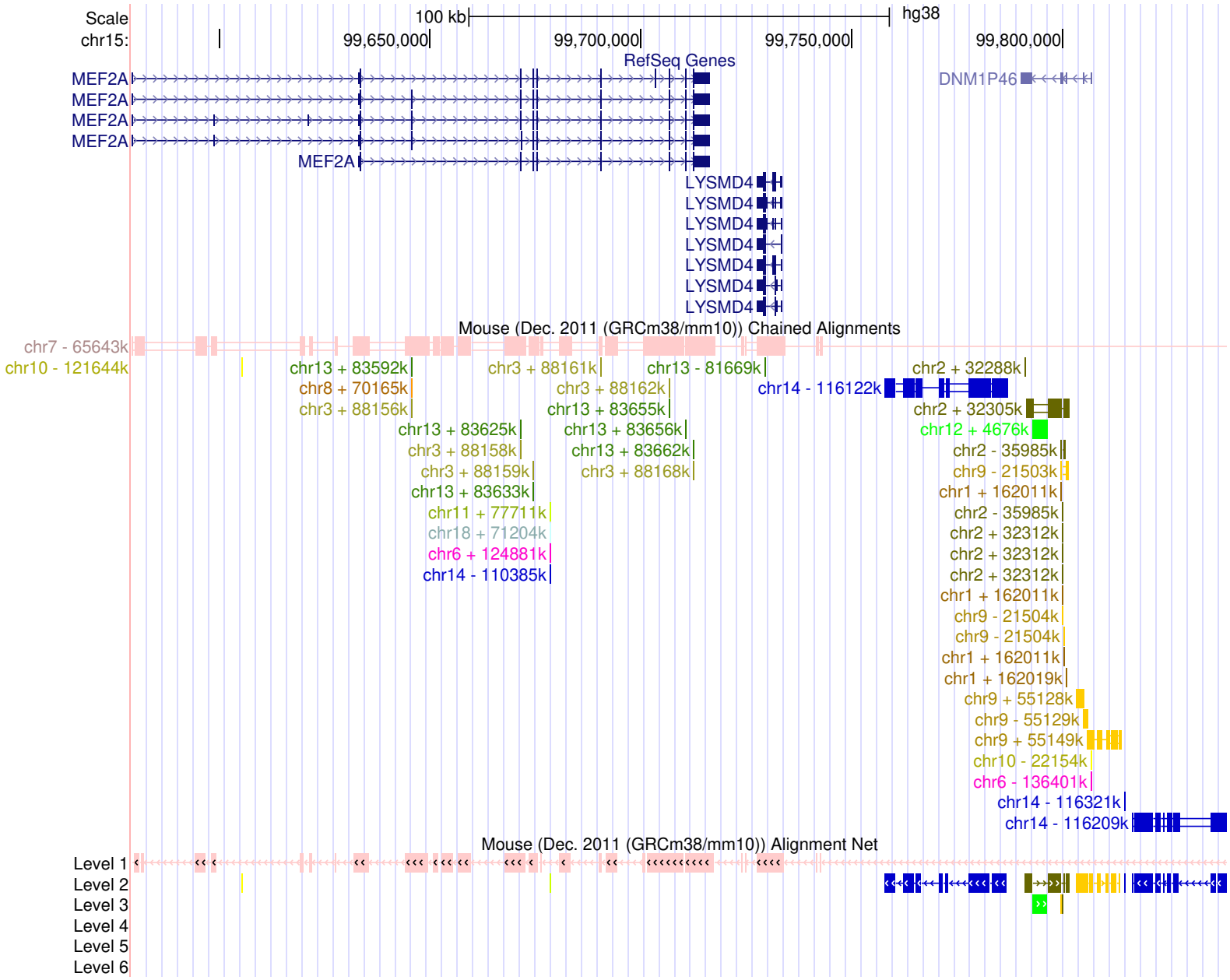
Aká časť zarovnaní má skóre 25 alebo vyššie?

(V praxi je simulácia pomalá, existujú odhady rozdelenia.)

Genomické zarovnania (whole-genome alignments)

Ku každému úseku ľudského genómu nájsť zodpovedajúcu časť z myši, psa, sliepky, atď. (predpočítané v UCSC browseri)

- Lokálne zarovnania nájdu exóny a iné zachované časti, sú však úseky, ktoré sa príliš zmenili.
- Pri duplikovaných úsekoch nevieme rozhodnúť, ktoré dvojice úsekov patria k sebe.
- **Synténia (synteny):** lokálne zarovnania, ktoré sa nachádzajú v dvoch genómoch v tom istom poradí a orientácii.
  Pomáha nám určiť, ktoré dvojice úsekov vznikli z tej istej oblasti v spoločnom predkovi (ortológy)


Scale
chr15:
100 kb
hg38
99,650,000
99,700,000
99,750,000
99,800,000
RefSeq Genes
MEF2A
MEF2A
MEF2A
MEF2A
MEF2A
DNM1P46
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
Mouse (Dec. 2011 (GRCm38/mm10)) Chained Alignments
chr7 - 65643k
chr10 - 121644k
chr13 + 83592k
chr8 + 70165k
chr3 + 88156k
chr3 + 88161k
chr13 - 81669k
chr3 + 88162k
chr13 + 83655k
chr13 + 83625k
chr3 + 88158k
chr3 + 88159k
chr13 + 83633k
chr11 + 77711k
chr18 + 71204k
chr6 + 124881k
chr14 - 110385k
chr13 + 83656k
chr13 + 83662k
chr3 + 88168k
chr14 - 116122k
chr2 + 32288k
chr2 + 32305k
chr12 + 4676k
chr2 - 35985k
chr9 - 21503k
chr1 + 162011k
chr2 - 35985k
chr2 + 32312k
chr2 + 32312k
chr2 + 32312k
chr1 + 162011k
chr9 - 21504k
chr9 - 21504k
chr1 + 162011k
chr1 + 162019k
chr9 + 55128k
chr9 - 55129k
chr9 + 55149k
chr10 - 22154k
chr6 - 136401k
chr14 - 116321k
chr14 - 116209k
Mouse (Dec. 2011 (GRCm38/mm10)) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6

Viacnásobné zarovnanie, multiple sequence alignment

Zarovnaj viacero sekvencií.

Zložitosť: $O(2^k n^k)$ pre k sekvencií dĺžky n

Pre všeobecné k NP-ťažké.

```
      Human  ctccatagcaatgt-cagagatagggcagagcggat------ggtggtgac
     Rhesus  ctccatggcaatgt-cagagatagggcagagcggat------gctggtgac
      Mouse  ttt--tgacaaca--tagagac-tgagatagaaaat-------atgctgac
        Dog  -tccccgctaatgtacaaagatggggcag-gaaga--a----tgtgctgaa
      Horse  -tccacggcaatac-tggagatggggcagagcaga--agat-ggtgatgaa
  Armadillo  ctgcatagaaatct-cagagatgggggaaagcaga------agacattcat
    Opossum  atccatggaaacat-cagaagtgggagaaatagaaga----tggcaatga-
   Platypus  acccggggaagggg-aagaggaagggccggccg-----------------
```

Heuristické algoritmy, napr. CLUSTAL-W [Higgins et al., 1996], MUSCLE [Edgar, 2004] a TBA [Blanchette et al., 2004].

Zhrnutie

- Zarovnávanie (alignment) je základný nástroj bioinformatiky
- Formulácia problému: voľba skórovacej schémy
- Riešenie problému: presné ale pomalé algoritmy a rýchlejšie heuristiky, ktoré nie vždy nájdu všetko
- Odhad štatistickej významnosti (E-hodnota, P-hodnota) je dôležitý nástroj na rozpoznávanie reálnych zarovnaní od tých, čo sa vyskytli náhodou
- Špecializované programy na rôzne úlohy súvisiace so zarovnávaním
  – Informatici na ďalších cvičeniach ďalšie finty na zlepšenie jadier
  – Biológovia ukážky použitia programov

Literatúra

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[Blanchette et al., 2004] Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715.

[Dembo et al., 1994] Dembo, A., Karlin, S., and Zeitouni, O. (1994). Limit distributions of maximal non-aligned two-sequence segmental score. *The Annals of Probability*, 22:2022–2039.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5:113.

[Higgins et al., 1996] Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996).

Using CLUSTAL for multiple sequence alignments. *Methods in enzymology*, 266:383–402.

[Karlin and Altschul, 1990] Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448.