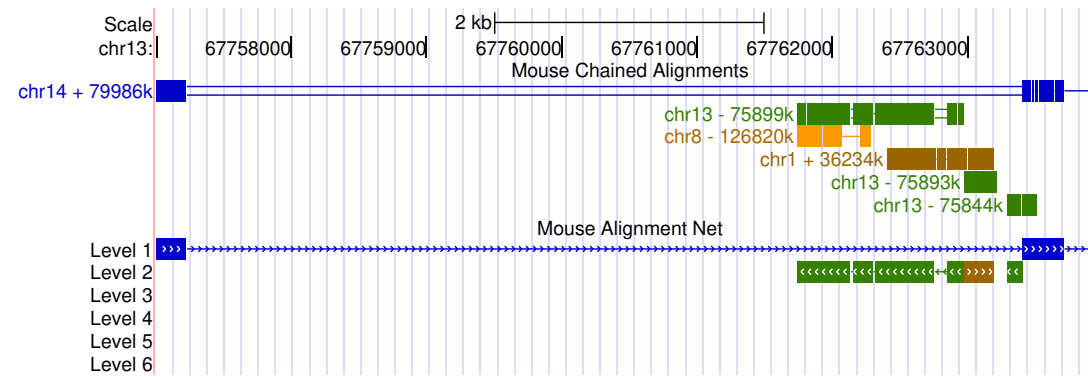


## Oznamy

- Výber článku na journal club formulárom na stránke do budúcej stredy 22.10. 22:00.
- Domáca úloha 1 bude zverejnená budúci týždeň
- Nezabudnite na pravidelné kvízy, ak je niečo nejasné, pýtajte sa.

## Zarovňavanie sekvencií 2/2 (sequence alignment)

Broňa Brejová  
16.10.2025



## Zhrnutie z minulej prednášky

- **Problém globálneho a lokálneho zarovnania**

**Vstup:** sekvencie  $X = x_1x_2 \dots x_n$  a  $Y = y_1y_2 \dots y_m$ .

**Výstup:**

globálne: zarovnanie  $X$  a  $Y$  s najvyšším skóre

lokálne: zarovnania podreťazcov  $x_i \dots x_j$  a  $y_k \dots y_\ell$  s najvyšším skóre.

- **Správny algoritmus na riešenie**

dynamické programovanie

- **Realistické skórovacie schémy**

## Máme správny algoritmus na zarovnávanie, čo viac nám chýba?

**Časová zložitosť:**  $O(n^2)$  na dvoch sekvenciách dĺžky  $n$ .

### Koľko je to času v skutočnosti?

(jednoduchá implementácia, náhodné sekvencie dĺžky  $n$ ,  
bežný počítač)

$n$	čas výpočtu
1,000	0.02s
10,000	1.5s
100,000	2.5 minúty (*)
1,000,000	4 hodiny (*)
10,000,000	17 dní (*)
100,000,000	5 rokov (*)
1,000,000,000	476 rokov (*)

### Potrebuje 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 a Lipman 1988)

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

## Ako nájdeme jadrá, 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  $w = 11$  a viac.)

		C	A	G	T	C	C	T	A	G	A
C	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	1	1	0	0	0	0
T	0	0	2	1	0	0	0	0	1	0	0
G	0	0	0	1	2	1	0	1	0	0	0
T	0	0	0	0	2	1	1	0	0	0	0
C	0	1	0	0	0	4	3	0	0	0	0
A	0	0	2	1	0	3	3	2	1	0	1
T	0	0	1	1	2	2	2	4	3	2	1
A	0	0	1	0	1	1	1	3	5	4	3

1. nájdi zhodné úseky

2. rozšír bez medzier

3. spoj medzerami

## Rýchlosť heuristického algoritmu

### Algoritmus:

- Nájdí 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.

**Ako zvýšenie  $w$  ovplyvní rýchlosť?**



## 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 jadier?

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

Jadro, t.j.  $w$  zhôd za sebou s pravdepodobnosťou  $4^{-w}$ .

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

Zvýšenie  $w$  o 1 zníži jadier cca 4 krát.

### Má zvýšenie $w$ nejakú nevýhodu?

## Senzitivita heuristického algoritmu

### Algoritmus:

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

**Nenájdene zarovnania:** vysoké skóre, ale **nemajú jadro dĺžky  $w$**

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

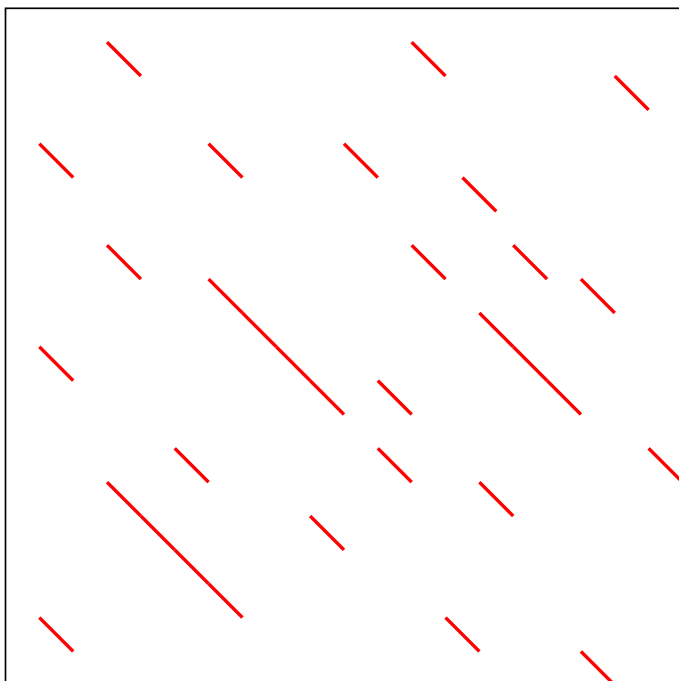
  | |   | | |   | |  
CATGTCATA

**Senzitivita:** aká časť **skutočných zarovnaní** obsahuje jadro, t.j. zhodu dĺžky  $w$

## Rýchlosť vs. senzitivita

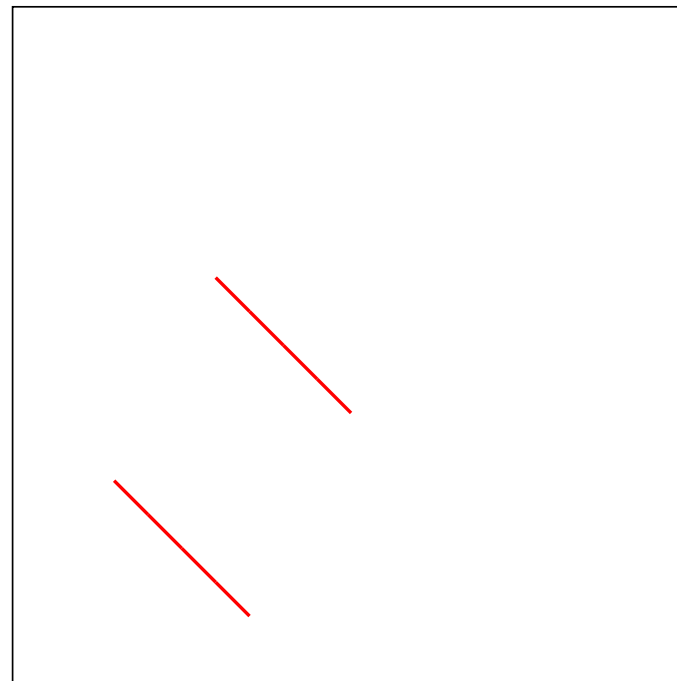
**Malé  $w$**

veľa náhodných zhôd, pomalé



**Veľké  $w$**

nenájdeme veľa zarovnaní



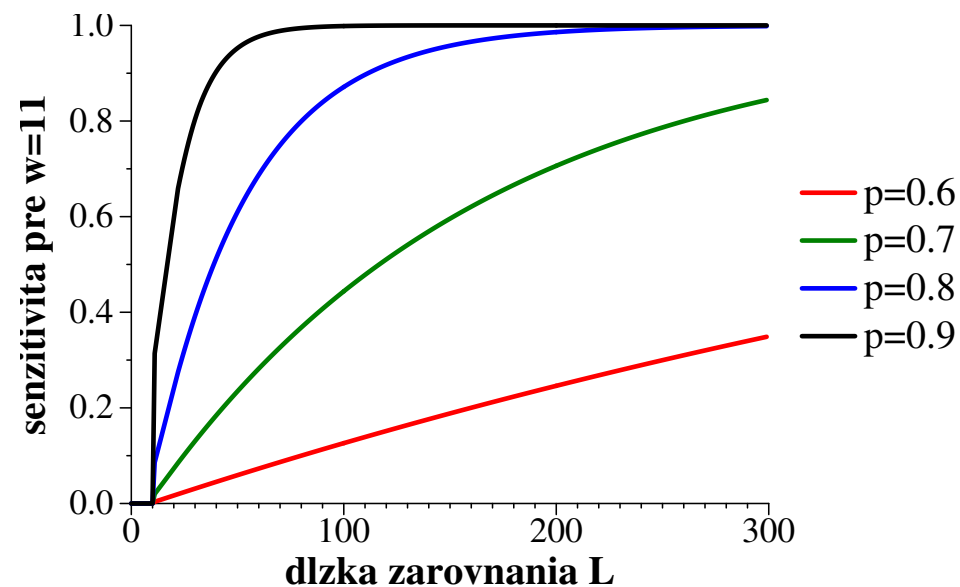
## Senzitivita heuristického algoritmu

### 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 na rôzne účely

**NCBI BLAST:** `blastn` pre DNA/RNA, `blastp` pre proteíny, `tblastx` preloží DNA do proteínu a použije `blastp`

**UCSC Blat:** pomerne rýchle vyhľadávanie veľmi podobných sekvencií, napr. kde je daná sekvencia v genóme

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

**Minimap2:** mapuje dlhé čítania na genóm alebo porovnáva dva príbuzné genómy

- veľmi rýchly, používa techniku minimizerov na ušetrenie pamäti (neukladá všetky úseky dĺžky  $w$ , vid' cvičenia pre inf.)

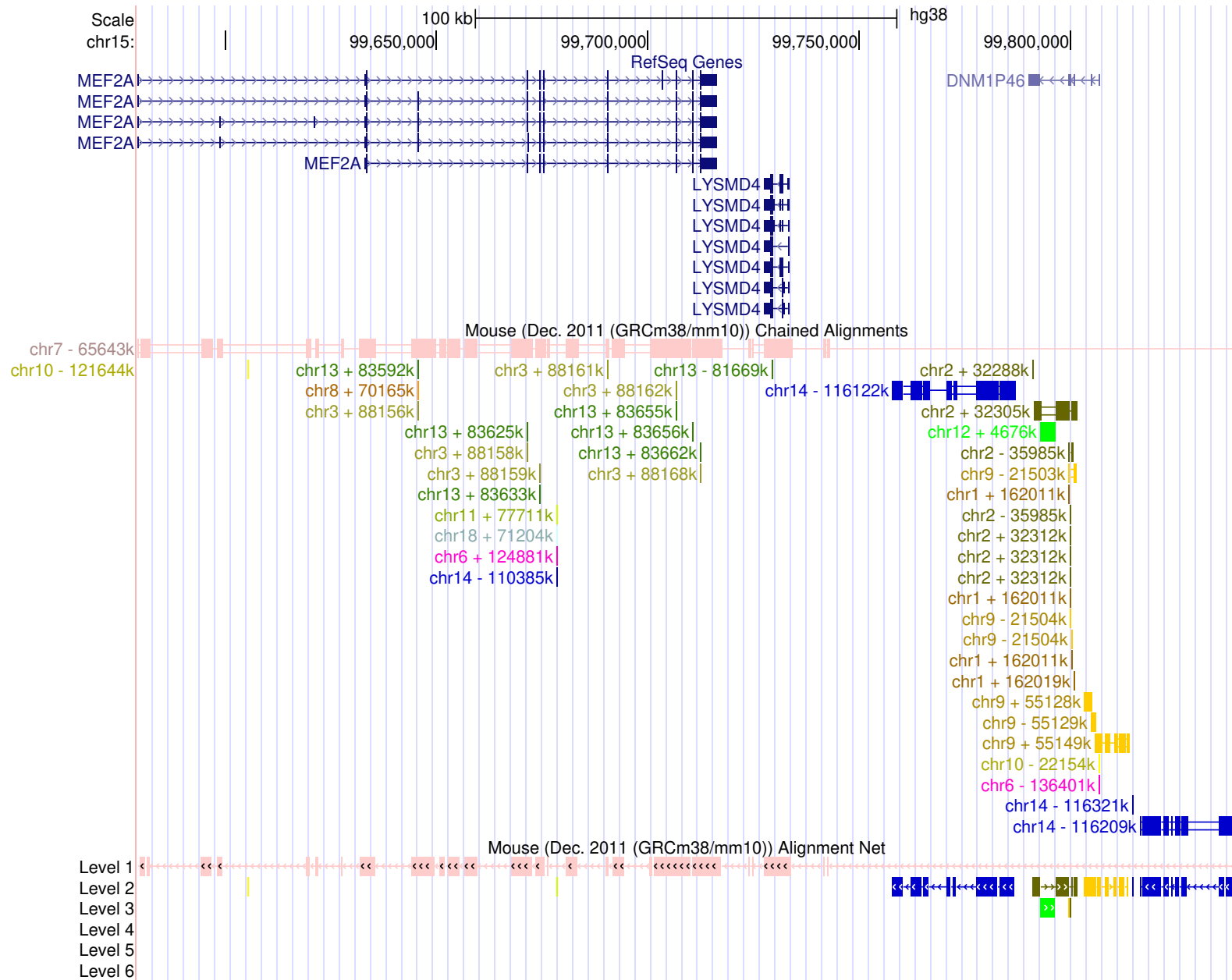
**BWA-MEM/BWA-MEM2:** mapuje krátke čítania na genóm

- namiesto jadier fixnej dĺžky používa maximálne presné zhody
- zložitejšie dátové štruktúry (BWT z informatických cvičení)

## Genomické zarovnanie (whole-genome alignments)

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

- Lokálne zarovnanie 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 zarovnanie, 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 (ortológ)





## Viacnásobné zarovnanie, multiple sequence alignment

Cieľ: Zarovnaj viacero sekvencií.

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

Ako by ste riešili skórovanie, aký by ste použili algoritmus?

## Viacnásobné zarovnanie, multiple sequence alignment

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

**Skórovanie:** napr. súčet párových skór všetkých dvojíc sekvencií.

V každej dvojici vyhodím stĺpce s dvomi pomlčkami.

**Zložitosť dynamického programovania:**  $O(2^k n^k)$  pre  $k$  sekvencií dĺžky  $n$ .

Pre všeobecné  $k$  NP-ťažké.

**Heuristické algoritmy,** napr. CLUSTAL-W, MUSCLE, TBA, MAFFT.

Často zarovnávajú hierarchicky vždy dve skupiny do jednej väčšej.

Sequences producing significant alignments									
Download ▾ Select columns ▾ Show 100 ▾ ?									
<input checked="" type="checkbox"/> select all 100 sequences selected           GenPept Graphics Distance tree of results Multiple alignment MSA Viewer									
	Description ▾	Scientific Name ▾	Max Score ▾	Total Score ▾	Query Cover ▾	E value ▾	Per. Ident ▾	Acc. Len ▾	Accession
<input checked="" type="checkbox"/>	<a href="#">hypothetical protein [Collimonas pratensis]</a>	<a href="#">Collimonas pratensis</a>	31.6	31.6	100%	17	91.67%	102	<a href="#">WP_150119746.1</a>
<input checked="" type="checkbox"/>	<a href="#">DNA mismatch repair protein [Mycena indigotica]</a>	<a href="#">Mycena indigotica</a>	30.3	30.3	91%	46	90.91%	968	<a href="#">XP_037221711.1</a>
<input checked="" type="checkbox"/>	<a href="#">DJ-1/Pfpl family protein [Rhodococcus sp. ACS1]</a>	<a href="#">Rhodococcus sp. ACS1</a>	30.3	30.3	91%	46	90.91%	217	<a href="#">WP_095863293.1</a>
<input checked="" type="checkbox"/>	<a href="#">DJ-1/Pfpl family protein [Rhodococcus koreensis]</a>	<a href="#">Rhodococcus koreensis</a>	30.3	30.3	91%	46	90.91%	217	<a href="#">WP_072942975.1</a>
<input checked="" type="checkbox"/>	<a href="#">MFS transporter [Brevibacterium ihuae]</a>	<a href="#">Brevibacterium ihuae</a>	29.9	29.9	91%	65	81.82%	393	<a href="#">WP_245873555.1</a>
<input checked="" type="checkbox"/>	<a href="#">MgtC/SapB family protein [Paenibacillus montanisoli]</a>	<a href="#">Paenibacillus montani...</a>	29.9	29.9	83%	66	90.00%	246	<a href="#">WP_112883223.1</a>
<input checked="" type="checkbox"/>	<a href="#">MgtC/SapB family protein [Paenibacillus montanisoli]</a>	<a href="#">Paenibacillus montani...</a>	29.9	29.9	83%	66	90.00%	246	<a href="#">WP_308637993.1</a>
<input checked="" type="checkbox"/>	<a href="#">cation diffusion facilitator family transporter [Spongiibacter sp. IMCC21906]</a>	<a href="#">Spongiibacter sp. IMC...</a>	29.5	29.5	83%	93	90.00%	302	<a href="#">WP_047012794.1</a>
<input checked="" type="checkbox"/>	<a href="#">cation diffusion facilitator family transporter [Zhongshania sp.]</a>	<a href="#">Zhongshania sp.</a>	29.5	29.5	83%	93	90.00%	302	<a href="#">WP_296435489.1</a>
<input checked="" type="checkbox"/>	<a href="#">tRNA modification GTPase [Lachnellula hyalina]</a>	<a href="#">Lachnellula hyalina</a>	29.1	29.1	100%	131	75.00%	581	<a href="#">XP_031004415.1</a>

Výsledok programu BLAST voči RefSeq proteínovej databáze na serveroch NCBI  
<https://blast.ncbi.nlm.nih.gov/>

[Download](#)  [GenPept](#) [Graphics](#)

### hypothetical protein [Collimonas pratensis]

Sequence ID: [WP\\_150119746.1](#) Length: 102 Number of Matches: 1

Range 1: 20 to 31 [GenPept](#) [Graphics](#)

 [Next Match](#)  [Previous Match](#)

Score	Expect	Identities	Positives	Gaps
31.6 bits(67)	17	11/12(92%)	11/12(91%)	0/12(0%)
Query	1	VIVALASVEGAS	12	
		VIVALASV GAS		
Sbjct	20	VIVALASVIGAS	31	

[Download](#)  [GenPept](#) [Graphics](#)

### DNA mismatch repair protein [Mycena indigotica]

Sequence ID: [XP\\_037221711.1](#) Length: 968 Number of Matches: 1

Range 1: 482 to 492 [GenPept](#) [Graphics](#)

 [Next Match](#)  [Previous Match](#)

Score	Expect	Identities	Positives	Gaps
30.3 bits(64)	46	10/11(91%)	10/11(90%)	0/11(0%)
Query	2	IVALASVEGAS	12	
		IVALASVE AS		
Sbjct	482	IVALASVEDAS	492	

## Ako rozlíšiť, či ide o významné zarovnanie?

Dĺžka dotazu  $m$ . Veľkosť databázy  $n$ .

Zarovnanie so skóre  $S$ .

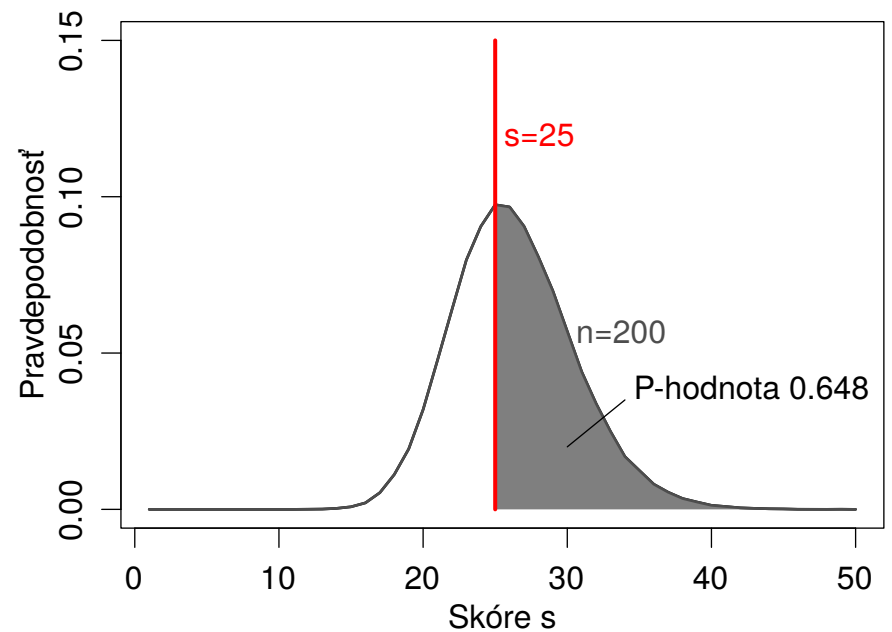
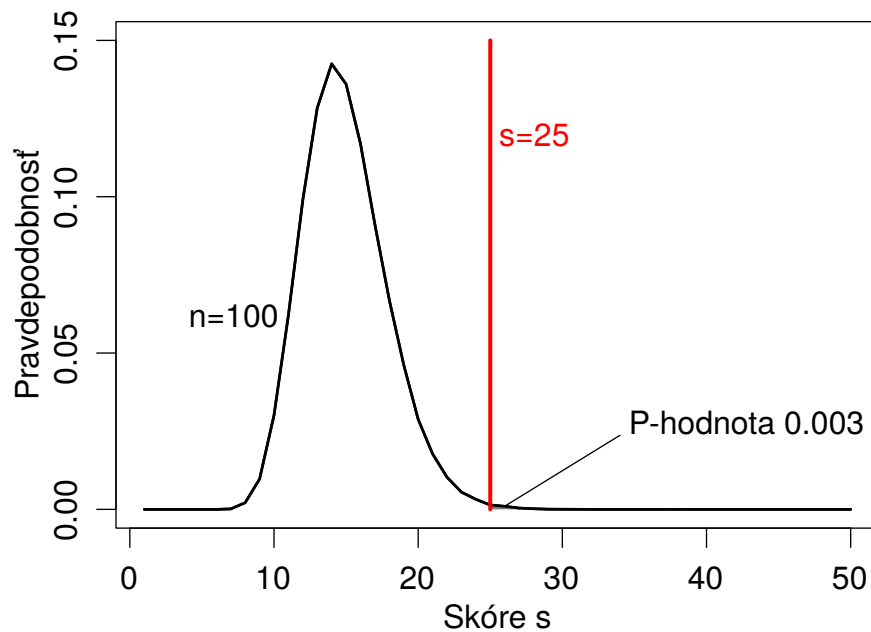
**$P$ -hodnota:** Pravdepodobnosť, že pre náhodný dotaz dĺžky  $m$  v náhodnej databáze dĺžky  $n$  nájdeme zarovnanie so skóre aspoň  $S$ .

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

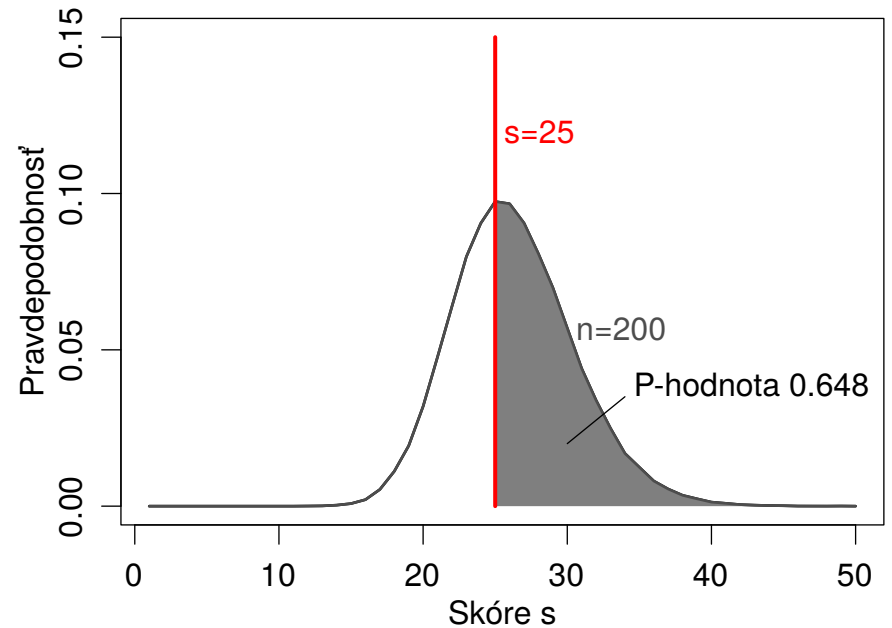
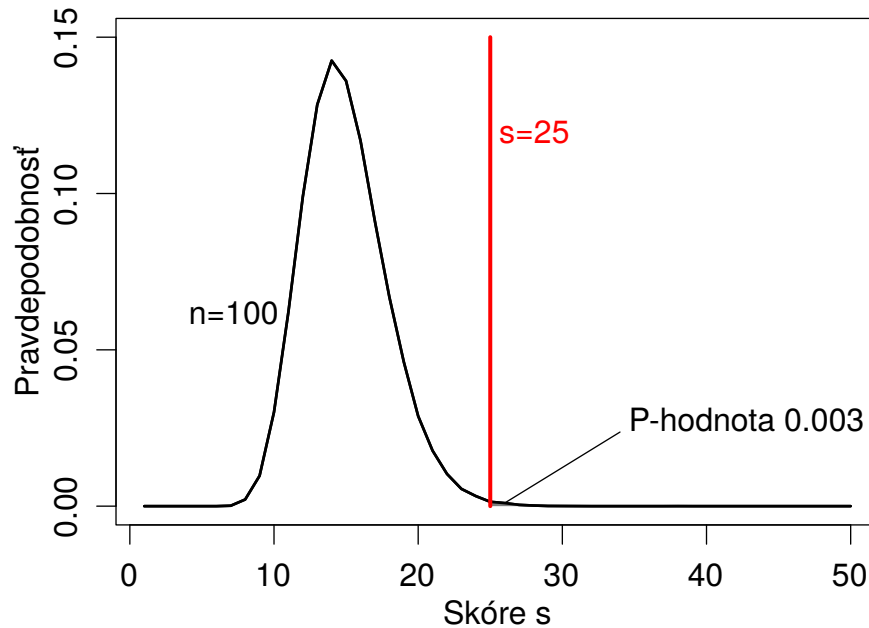
Pri veľmi malých hodnotách sú  $E$ -hodnota a  $P$ -hodnota takmer identické.

## 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 matematické odhady rozdelenia.

Karlin and Althschul 1990, Dembo et al. 1994

## 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