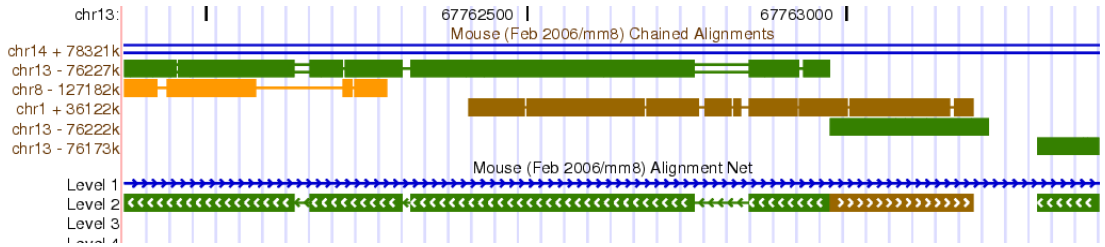
# Zarovnávanie sekvencií (sequence alignment) 1/2

**Tomáš Vinař**

**6.10.2022**



## Problém: Lokálne zarovnávanie (local alignment)

```
ggcccttggagttgactgtcctgctgctccttgagg
ccattctcagagagaggaagtggcctcattttaatc
cgcttcccacagccttgtcctttccagacccatggg
agagggaggggctgagggtgtggctgagcccaccca
agtcacgcgtcactctgcaggtccctctcccccaag
gccgtggccttgggagcccgtggatcccagtgagtg
acgcctccacccccgccctactcgggcagtttaac
ccttgttgttcacttgcagacatcgtgaacacggcc
cggcccgacgagaaggccataatgacctatgtgtcc
agcttctaccatgccttttcaggagcgcagaaggta
ccgagcagggccaggcaggccctcctcgccgccacc
gcgcaatgccgccgctgcctctcgcctcccgtgctc
acctcatttctcttgcagacggcagtggcctctctc
caactggaagccacccccagctccct...
```

```
tgatgccgaggatgtgttcgtcgagcatccggacga
gaagtccatcacctacgtggtcacctactatcacta
ctttagcaaactcaagcaggagacggtgcagggcat
aagcgtatcggtaaggtggtcggcattgccatggag
aacgacaaaatggtccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgaccatccag
tcgctgggcgagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcaccatcgagaagccgcccaagtttgtggaaaag
ggcaacctcgaggtgctcctttcaccctgcagtcc
aagatgcgggccaacaaccagaagccctacacaccc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggccctgcgcgaggagctcatccg...
```

**Vstup:** dve sekvencie

## Problém: Lokálne zarovnávanie (local alignment)

```
ggcccttggagttgactgtcctgctgctccttgagg   tgatgccgaggatgtgttcgtcgagcatccggacga
ccattctcagagagaggaagtggcctcattttaatc   gaagtccatcacctacgtggtcacctactatcacta
cgcttcccacagccttgtcctttccagacccatggg   ctttagcaaactcaagcaggagacggtgcagggcat
agagggaggggctgagggtgtggctgagcccaccca   aagcgtatcggtaaggtggtcggcattgccatggag
agtcacgcgtcactctgcaggtccctctcccccaag   aacgacaaaatggtccacgactacgagaacttcaca
gccgtggccttgggagcccgtggatcccagtgagtg   agcgatctgctcaagtggatcgaaacgaccatccag
acgcctccacccccgccctactcgggcagtttaac   tcgctgggcgagcgggagttcgaaaactcgctggcc
ccttgttgttcacttgcagacatcgtgaacacggcc   ggcgtccaagggcagttggcccagttctccaactac
cggcccgacgagaaggccataatgacctatgtgtcc   cgcaccatcgagaagccgcccaagtttgtggaaaag
agcttctaccatgcctttcaggagcgcagaaggta   ggcaacctcgaggtgctcctttcaccctgcagtcc
ccgagcagggccaggcaggccctcctcgccgccacc   aagatgcgggccaacaaccagaagccctacacaccc
gcgcaatgccgccgctgcctctcgcctcccgtgctc   aaagagggcaagatgatttcggacatcaacaaggcc
acctcatttctcttgcagacggcagtggcctctctc   tgggagcgtctggagaaggccgagcacgaacgcgaa
caactggaagccacccccagctccct...          ttggccctgcgcgaggagctcatccg...
```

**Výstup:** podobné úseky (zarovnania, alignments).

```
CCCGACGAGAAGGCCATAATGACCTATGTGTCCAGCTTCTACCATGCCTTT
||  |||||||||| ||||     ||||| |||  ||  ||  ||| ||    ||||
CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT
```

Vlož pomlčky (medzery, gaps) tak, aby rovnaké bázy boli pod sebou.

Dobré zarovnanie má veľa zarovnaných rovnakých báz, málo medzier.

## Na čo sú dobré zarovnania?

- **Orientácia v obrovských databázach.**

  Genbank WGS má vyše 3 TB sekvencií.

  Napr. z ktorého genómu (a odkiaľ) pochádza daná sekvencia?

- **Určovanie funkcie (napr. proteínu).**

  Podobné sekvencie často majú rovnakú/podobnú funkciu.

- **Štúdium evolúcie.**

  Hľadáme homológy: sekvencie, ktoré sa vyvinuli z toho istého spoločného predka.

  V ideálnom prípade medzery zodpovedajú inzerciám a deléciám, zarovnané bázy zachovaným bázam a substitúciám.

- **Hľadanie génov a iných funkčných prvkov.**

  Menia sa pomalšie ako ostatné sekvencie.

## Zarovnávanie sekvencií ako optimalizačný problém

**Cieľ:** nájdi páry homologických sekvencií

(tých, čo pochádzajú z rovnakého spoločného predka)

**Modelovacia fáza:** vytvor skórovaciu schému, ktorá

– skutočným homologickým párom dáva vysoké skóre

– falošne pozitívnym párom dáva nízke skóre

**Optimalizačná fáza:**

pre dané dve vstupné sekvencie, nájde zarovnanie s najlepším skóre

dôležitá je výpočtová a pamäťová zložitosť algoritmu

## Formulácia problému

**Skórovanie zarovnania:** napr. zhoda +1, nezhoda -1, medzera -1.

```
GAGAAGGCCATAATGACCTATGTGTCCAGCT
||||||  ||||     |||||  |||   ||  ||
GAGAAGTCCAT---CACCTACGTGGTCACCT
```

22 zhôd, 6 nezhôd, 3 medzery $\rightarrow$ skóre 13.

V praxi zložitejšie skórovanie.

### Problém 1: globálne zarovnanie (global alignment)

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.

### Problém 2: lokálne zarovnanie (local alignment)

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

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

## Dynamické programovanie pre globálne zarovnanie

## (Needleman, Wunsch 1970)

**Podproblém:** $A[i, j]$: najvyššie skóre globálneho zarovnania reťazcov $x_1 x_2 \ldots x_i$ a $y_1 y_2 \ldots y_j$.

**Jeden z reťazcov dĺžky 0:** druhý reťazec je zarovnaný s medzerou.

$A[0, j] = -j$, $A[i, 0] = -i$.

**Všeobecný prípad, $i > 0$, $j > 0$:**

ak $x_i = y_j$ sú zarovnané $A[i, j] = A[i-1, j-1] + 1$

ak $x_i \neq y_j$ sú zarovnané $A[i, j] = A[i-1, j-1] - 1$

ak $x_i$ je zarovnané s medzerou $A[i, j] = A[i-1, j] - 1$

ak $y_j$ je zarovnané s medzerou $A[i, j] = A[i, j-1] - 1$

## Dynamické programovanie pre globálne zarovnanie

**Podproblém:** $A[i,j]$: najvyššie skóre globálneho zarovnania reťazcov $x_1x_2\ldots x_i$ a $y_1y_2\ldots y_j$.

**Všeobecný prípad, $i>0$, $j>0$:**

ak $x_i = y_j$ sú zarovnané $A[i,j] = A[i-1,j-1]+1$

ak $x_i \neq y_j$ sú zarovnané $A[i,j] = A[i-1,j-1]-1$

ak $x_i$ je zarovnané s medzerou $A[i,j] = A[i-1,j]-1$

ak $y_j$ je zarovnané s medzerou $A[i,j] = A[i,j-1]-1$

**Rekurencia:**

$$A[i,j] = \max \begin{cases} A[i-1,j-1] + s(x_i,y_j), \\ A[i-1,j]-1, \\ A[i,j-1]-1 \end{cases}$$

kde $s(x,y)=1$ ak $x=y$ $\quad s(x,y)=-1$ ak $x\neq y$

## Príklad globálneho zarovnania

CATGTCGTA vs CAGTCCTAGA

| | | C | A | G | T | C | C | T | A | G | A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 | -10 |
| C | -1 | 1 | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 |
| A | -2 | 0 | 2 | 1 | 0 | -1 | -2 | -3 | -4 | -5 | -6 |
| T | -3 | -1 | 1 | 1 | ? | | | | | | |
| G | -4 | | | | | | | | | | |
| T | -5 | | | | | | | | | | |
| C | -6 | | | | | | | | | | |
| G | -7 | | | | | | | | | | |
| T | -8 | | | | | | | | | | |
| A | -9 | | | | | | | | | | |

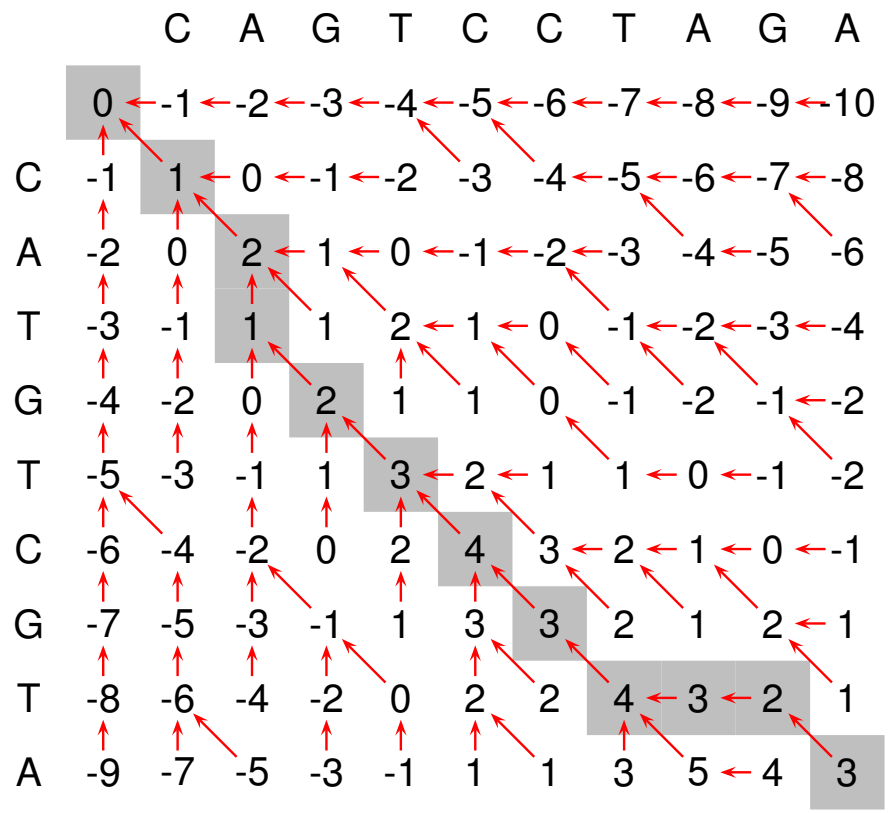
$$A[i,j] = \max \begin{cases} A[i-1,j-1] + s(x_i, y_j), \\ A[i-1,j] - 1, \\ A[i,j-1] - 1 \end{cases}$$

## Príklad globálneho zarovnania

CATGTCGTA vs CAGTCCTAGA

| | | C | A | G | T | C | C | T | A | G | A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 | -10 |
| C | -1 | 1 | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 |
| A | -2 | 0 | 2 | 1 | 0 | -1 | -2 | -3 | -4 | -5 | -6 |
| T | -3 | -1 | 1 | 1 | 2 | 1 | 0 | -1 | -2 | -3 | -4 |
| G | -4 | -2 | 0 | 2 | 1 | 1 | 0 | -1 | -2 | -1 | -2 |
| T | -5 | -3 | -1 | 1 | 3 | 2 | 1 | 1 | 0 | -1 | -2 |
| C | -6 | -4 | -2 | 0 | 2 | 4 | 3 | 2 | 1 | 0 | -1 |
| G | -7 | -5 | -3 | -1 | 1 | 3 | 3 | 2 | 1 | 2 | 1 |
| T | -8 | -6 | -4 | -2 | 0 | 2 | 2 | 4 | 3 | 2 | 1 |
| A | -9 | -7 | -5 | -3 | -1 | 1 | 1 | 3 | 5 | 4 | 3 |

## Ako získať zarovnanie?




```
CA-GTCCTAGA
CATGTCGT--A
```

## Dynamické programovanie pre lokálne zarovnanie

## (Smith, Waterman 1981)

**Podproblém:** $A[i,j]$: najvyššie skóre lokálneho zarovnania reťazcov $x_1x_2\ldots x_i$ a $y_1y_2\ldots y_j$, ktoré obsahuje bázy $x_i$ a $y_j$, alebo je prázdne.

**Jeden z reťazcov dĺžky 0:** prázdne zarovnanie $A[0,j] = A[i,0] = 0$

**Všeobecný prípad, $i > 0$, $j > 0$:**

ak $x_i$ a $y_j$ sú zarovnané $A[i,j] = A[i-1,j-1] + s(x_i, y_j)$

ak $x_i$ je zarovnané s medzerou $A[i,j] = A[i-1,j] - 1$

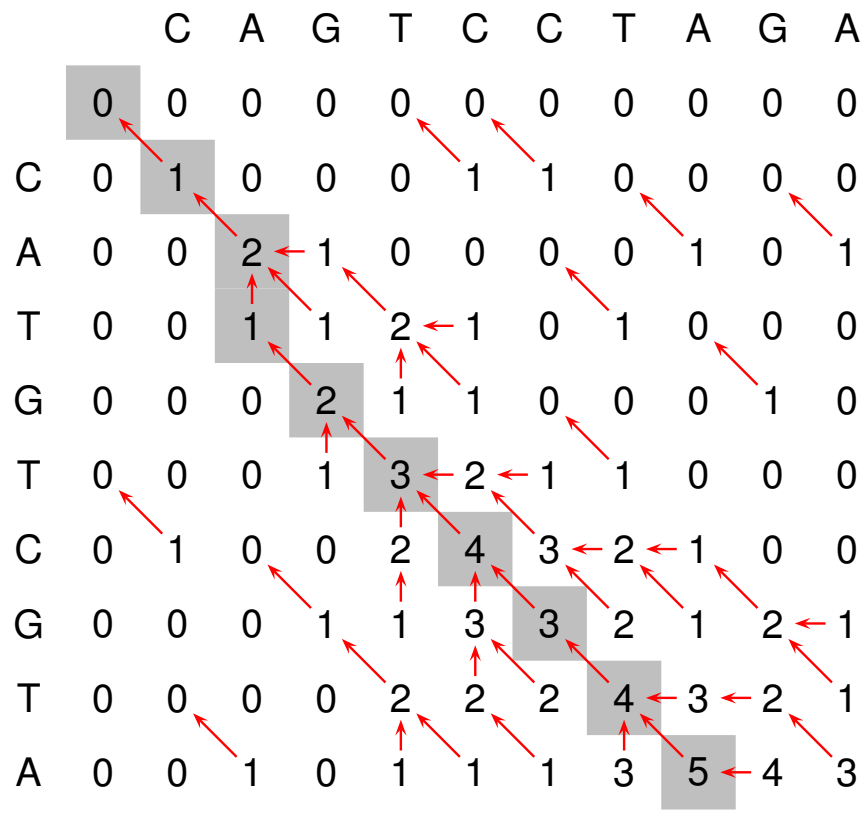
ak $y_j$ je zarovnané s medzerou $A[i,j] = A[i,j-1] - 1$

ak $x_i$ a $y_j$ nie sú časťou zarovnania s kladným skóre $A[i,j] = 0$

**Rekurencia:**

$$A[i,j] = \max \begin{cases} 0, \\ A[i-1,j-1] + s(x_i, y_j), \\ A[i-1,j] - 1, \\ A[i,j-1] - 1 \end{cases}$$

# Príklad lokálneho zarovnania




CA-GTCCTA

CATGTCGTA

## Zložitejšie skórovanie

### Problémy $+1, -1$ skórovania:

- Je skutočne jedna nezhoda alebo medzera až taká zlá v porovnaní s jednou zhodou?
- Čo urobíme pre zarovnávanie proteínov?
  (20 prvková abeceda $\approx$ 200 parametrov)

### Úloha skórovacej schémy:

- Chceme vedieť rozlíšiť **lepšie zarovnania** od **horších zarovnaní**:
  - Ktoré usporiadania pomlčiek dávajú väčší zmysel
- Chceme vedieť, či dané zarovnanie **má biologický význam**:
  - Ide o homológy, alebo sekvencie nesúvisia?

## Zložitejšie skórovanie: prvý pokus

Nech $X$ a $Y$ sú **správne zarovnané homológy**

$a =$ pravdepodobnosť, že sa dve bázy **zhodujú**

$b =$ pravdepodobnosť, že sa **nezhodujú**

$c =$ pravdepodobnosť, že báza je **zarovnaná s medzerou**

$$a + b + c = 1$$

**Pravdepodobnosť zarovnania $A$:**

```
GAGAAGGCCATAATGACCTATGTGTCCAGCT
|||||| ||||     ||||| |||  || ||
GAGAAGTCCAT---CACCTACGTGGTCACCT
```

$$\Pr(A) = a^{22}b^6c^3$$

**Ktoré je pravdepodobnejšie?**

```
CACA
|  |
CCAA
```

$\Pr(A) = a^2b^2$

```
CACA-
|  ||
C-CAA
```

$\Pr(A) = a^3c^2$

## Zložitejšie skórovanie: prvý pokus

Zlogaritmujeme: násobenie sa zmení na sčítavanie

môžeme použiť S.-W. alebo N.-W. dyn. prog. algoritmy

$$\Pr(A) = a^{22} b^6 c^3$$

$$\log \Pr(A) = 22 \log a + 6 \log b + 3 \log c$$

**Skóre:** Zhoda: $\log a$ Nezhoda: $\log b$ Medzera: $\log c$

## Nevýhody takejto schémy:

- Vždy záporné skóre $\Rightarrow$ čo s lokálnymi zarovnaniami?
- Neužitočné pre porovnávanie rôznych párov sekvencií

## Skórovanie založené na dvoch pravdepodobnostných modeloch

**Porovnávame dva modely $H$ a $R$:** logaritmus podielu

$$\log \frac{\Pr(X, Y \mid H)}{\Pr(X, Y \mid R)}$$

- Ak sú sekvencie $X$ a $Y$ **homológy**

  $\Rightarrow$ logaritmus podielu oveľa väčší ako 1 $\Rightarrow$ **veľmi pozitívne skóre**

- Ak sekvencie $X$ a $Y$ **nesúvisia**

  $\Rightarrow$ logaritmus podielu oveľa menší ako 1 $\Rightarrow$ **veľmi negatívne skóre**

## Zložitejšie skórovanie: dva pravdepodobnostné modely

(Pre jednoduchosť teraz neuvažujme medzery)

**Model H:** Sekvencie $X$ a $Y$ sú **správne zarovnané homológy**

$\Pr(X, Y \mid H) = \prod_{i=1}^{n} p(x_i, y_i)$

$p(x_i, y_i)$ : pravdepodobnosť, že vidíme zarovnané práve bázy $x_i$ a $y_i$

**Model R:** Sekvencie $X$ a $Y$ **nijako spolu nesúvisia**

$\Pr(X, Y \mid R) = \left(\prod_{i=1}^{n} p(x_i)\right) \left(\prod_{i=1}^{n} p(y_i)\right)$

$p(x_i)$ : pravdepodobnosť výskytu bázy $x_i$

**Porovnávame dva modely $H$ a $R$:** logaritmus podielu

$$\log \frac{\Pr(X, Y \mid H)}{\Pr(X, Y \mid R)}$$

$$\Pr(X, Y \mid H) = \prod_{i=1}^{n} p(x_i, y_i)$$
$$\Pr(X, Y \mid R) = \left(\prod_{i=1}^{n} p(x_i)\right)\left(\prod_{i=1}^{n} p(y_i)\right)$$

$$\log \frac{\Pr(X, Y \mid H)}{\Pr(X, Y \mid R)} = \log \frac{\prod_{i=1}^{n} p(x_i, y_i)}{\left(\prod_{i=1}^{n} p(x_i)\right)\left(\prod_{i=1}^{n} p(y_i)\right)} = \sum_{i=1}^{n} \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}$$

**Skóre zarovnanie bázy $x$ a bázy $y$:**

$s(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$

## BLOSUM62 skórovacia matica pre proteíny

**BLO**cks of aminoacid **SU**bstitution **M**atrix; Henikoff, Henikoff 1992

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

- Vyber **biologicky relevantné zarovnania** proteínov (BLOCKS)
- Páry s nanajvýš 62% identitou
- $p(x, y)$ : ako často vidíme aminokyseliny $x$ a $y$ zarovnané
- $p(x)$ : ako často sa vyskytuje aminokyselina $x$

- **skóre pre dvojicu aminokyselín $x$ a $y$**: $\log \dfrac{p(x,y)}{p(x)p(y)}$
- prenásobíme konštantou a zaokrúhlime:
  - aby sme neurobili príliš veľkú chybu
  - aby sa s číslami lepšie počítalo

## Zložitejšie skórovanie: afínne skóre medzier

```
CCCGACGAGAAGGCCATAATGACCTATGTGTCCAGCTTCTACCATGCCTTT
||  |||||||||  ||||    |||||  |||  ||  ||  |||  ||   ||||
CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT
```

Niekoľko medzier za sebou asi nevzniklo nezávisle, možno jedna mutácia.

Penalta za začatie medzery (gap opening cost) $o$,

Penalta za rozšírenie medzery o jedna (gap extension cost) $e$.

Medzera dĺžky $g$ má penaltu $o + e(g - 1)$.

Zvolíme $o < e$ (t.j. $|o| > |e|$).

Základné nastavenia blastn: zhoda +2, nezhoda -3, $o = -5$, $e = -2$.

Príklad vyššie: 22 zhôd, 6 nezhôd, 1 medzera dĺžky 3

$\rightarrow$ skóre $2 \cdot 22 - 3 \cdot 6 - 5 - 2 \cdot 2 = 16$.

## Zhrnutie

- Globálne a lokálne zarovania
- Needleman-Wunschov a Smith-Watermanov algoritmus
- Skórovanie zarovnaní pomocou porovnávania modelov
- Proteínové BLOSUM matice
- Afínne skórovanie medzier

## Problémy na zamyslenie

1. **Časová zložitosť Smith-Waterman:** $O(nm)$

   $n$ - veľkosť prvej sekvencie

   $m$ - veľkosť druhej sekvencie

   **Čo robiť ak chceme porovnať ľudský genóm s myšacím genómom?**

2. Povedzme, že nájdeme zarovnanie so skóre $14$

   **Je toto skóre dobré, alebo ide o niečo, čo vidíme náhodou?**