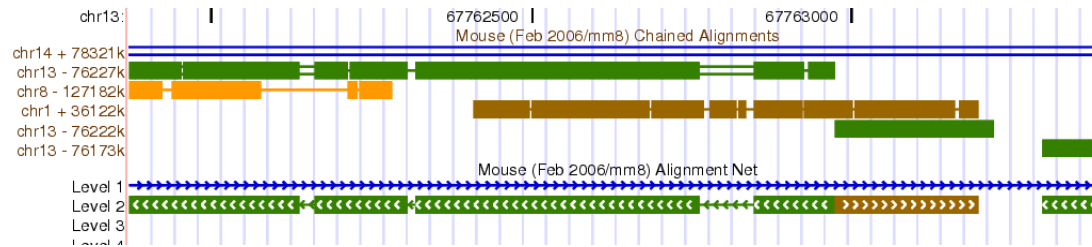


Zarovňavanie sekvencií (sequence alignment) 1/2

Broňa Brejová

5.10.2023



Problém: Lokálne zarovnanie (local alignment)

ggccttggagttgactgtcctgctgctccttgagg
ccattctcagagagaggaagtggcctcattttaatc
cgcttcccacagccttgtcctttccagacccatggg
agagggaggggctgaggggtgtggctgagcccacca
agtcacgcgtcactctgcaggtccctctcccccaag
gccgtggccttgggagcccgtggatcccagtgagtg
acgcctccacccccgcctactcgggcagtttaac
ccttgttgttcaacttgcagacatcgtgaacacggcc
cggcccgcagagaaggccataatgacctatgtgtcc
agcttctaccatgccttttcaggagcgcagaaggta
ccgagcagggccaggcaggccctcctcgccgccacc
gcgcaatgccgccgctgctctcgctcccgtgctc
acctcatttctcttgcagacggcagtggcctctctc
caactggaagccacccccagctcct...

tgatgccgaggatgtgttcgtcgagcatccggacga
gaagtccatcacctacgtggtcacctactatcacta
cttagcaaactcaagcaggagacgggtgcagggcat
aagcgtatcggtaaggtggcggcattgccatggag
aacgacaaaatgggtccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgaccatccag
tcgctgggcgagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcaccatcgagaagccgcccaagtttgtggaaaag
ggcaacctcgaggtgctccttttcacctgcagtcc
aagatgcgggccaacaaccagaagccctacacacc
aaagagggcaagatgatctcgacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggcctgcgcgaggagctcatccg...

Vstup: dve sekvencie

Problém: Lokálne zarovnávanie (local alignment)

ggcccttggagttgactgtcctgctgctccttgagg
ccattctcagagagaggaagtggcctcattttaatc
cgcttcccacagccttgtcctttccagacccatggg
agagggaggggctgaggggtgtggctgagcccacca
agtcacgcgtcactctgcaggtccctctcccccaag
gccgtggccttgggagcccgtggatcccagtgagtg
acgcctccaccccccgccctactcgggcagtttaac
ccttgttgttcacttgcagacatcgtgaacacggcc
cggcccgacgagaaggccataatgacctatgtgtcc
agcttctaccatgccttttcaggagcgcagaaggta
ccgagcagggccagggcaggccctcctcgccgccacc
gcgcaatgccgcccgtgctctcgctcccgtgctc
acctcatttctcttgcagacggcagtggcctctctc
caactggaagccacccccagctccct...

tgatgccgaggatgtgttcgctcgagcatccggacga
gaagtccatcacctacgtgggtcacctactatcacta
cttttagcaaactcaagcaggagacgggtgcagggcat
aagcgtatcggtaaggtggtcggcattgccatggag
aacgacaaaatgggtccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgaccatccag
tcgctgggfcgagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcaccatcgagaagccgcccagtttgtggaaaag
ggcaacctcgaggtgctccttttcaccctgcagttcc
aagatgcgggccaacaaccagaagccctacacacc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggccctgcgcgaggagctcatccg...

Výstup: podobné úseky (zarovnaná, 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 zarovnanie?

- **Orientácia v obrovských databázach.**

Genbank WGS má vyše 22 TB sekvencií.

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

- **Prekryvy čítaní** pri skladaní genómov, mapovanie čítaní

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

Podobné sekvencie často majú rovnakú/podobnú funkciu.

- **Štúdium evolúcie.**

Hľadáme homológy: sekvencie, ktoré sa vyvinuli z tej istej sekvencie v spoločnom predkovi.

V ideálnom prípade medzery zodpovedajú inzerciam a deléciám, zarovnané bázy zachovaným báзам a substitúciám.

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

Zarovňavanie sekvencií ako optimalizačný problém

Cieľ: nájsť 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, nájsť zarovnanie s najlepším skóre
dôležitá je výpočtová a pamäťová zložitosť algoritmu

Formulácia problému

Skórovanie zarovnaní: napr. zhoda +1, nezhoda -1, medzera -1.

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

22 zhôd, 6 nezhôd, 3 medzery → skóre 13.

V praxi zložitejšie skórovanie.

Problém 1: globálne zarovnanie (global alignment)

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

Výstup: zarovnanie X a Y s najvyšším skóre.

Problém 2: lokálne zarovnanie (local alignment)

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

Výstup: zarovnanie podreťazcov $x_i \dots x_j$ a $y_k \dots y_\ell$ s najvyšším skóre.

Dynamické programovanie pre globálne zarovnanie (Needleman, Wunsch 1970)

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

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 zarovnanania reťazcov $x_1x_2 \dots x_i$ a $y_1y_2 \dots y_j$.

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 $x = y$ $s(x, y) = -1$ 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										

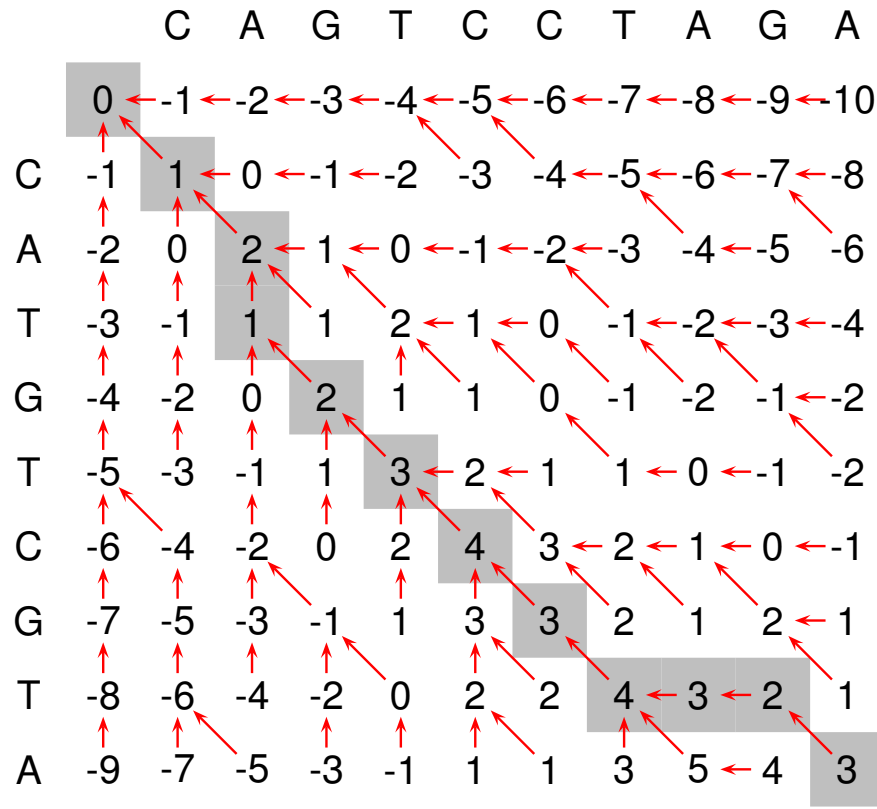
$$A[i, j] = \max \begin{cases} A[i - 1, j - 1] + s(x_i, y_j), \\ A[i - 1, j] - 1, \\ A[i, j - 1] - 1 \end{cases}$$

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?



CATGTCGT--A

CA-GTCCTAGA

Časová zložitosť celého algoritmu $O(nm)$

Dynamické programovanie pre lokálne zarovnanie

(Smith, Waterman 1981)

Podproblém: $A[i, j]$: najvyššie skóre lokálneho zarovnania reťazcov $x_1x_2 \dots x_i$ a $y_1y_2 \dots y_j$, 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 $A[i, j] = A[i - 1, j] - 1$

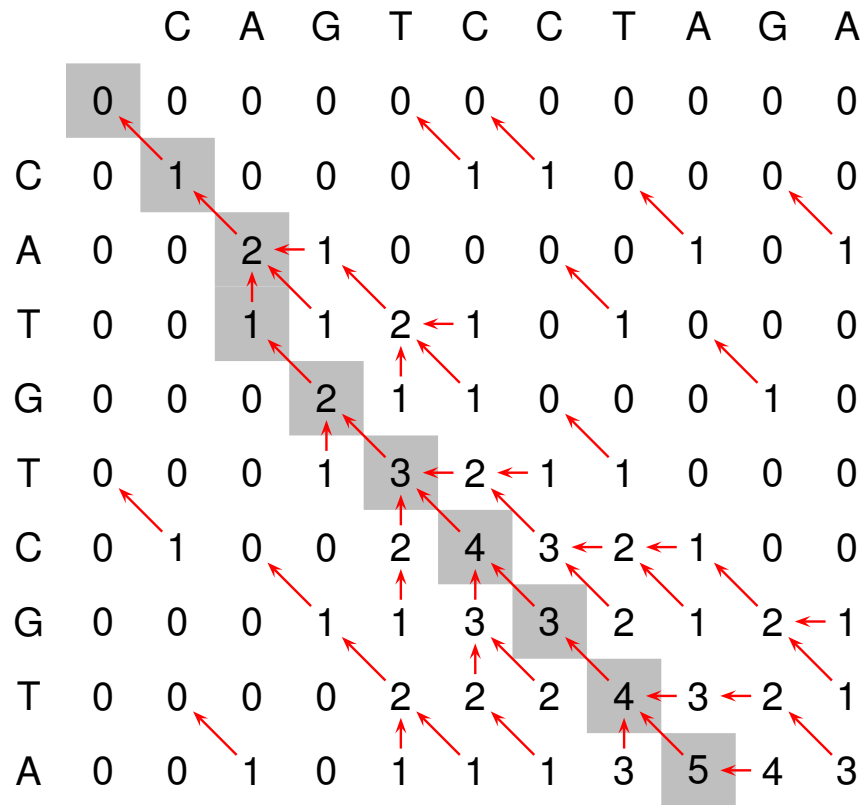
ak y_j je zarovnané s medzerou $A[i, j] = A[i, j - 1] - 1$

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



CATGTCGTA

CA-GTCCTA

Časová zložitosť celého algoritmu $O(nm)$

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

Skórovanie založené na dvoch pravdepodobnostných modeloch

Majme zarovnané sekvencie X a Y

Aké má zarovnanie dostať skóre?

Pre jednoduchosť teraz neuvažujme medzery v zarovnaní.

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

Model generuje nezávisle **jednotlivé stĺpce** zarovnaní.

Dané pravdepodobnosti $p(x, y)$, že v stĺpci budú pod sebou x a y .

Pravdepodobnosť vygenerovania X zarovnaného s Y :

$$\Pr(X, Y | H) = \prod_{i=1}^n p(x_i, y_i) = p(x_1, y_1) \cdot p(x_2, y_2) \cdot \dots \cdot p(x_n, y_n)$$

Príklad: Uvažujme iba bázy A, T

Model H: $p(A, A) = 0.4$, $p(A, T) = 0.15$, $p(T, A) = 0.15$, $p(T, T) = 0.3$

$$\Pr(AAAT, ATAT | H) = 0.4 \cdot 0.15 \cdot 0.4 \cdot 0.3 = 0.0072$$

Skórovanie založené na dvoch pravdepodobnostných modeloch

Majme zarovnané sekvencie X a Y

Aké má zarovnanie dostať skóre?

Pre jednoduchosť teraz neuvažujme medzery v zarovnaní.

Model R: Sekvencie X a Y **nijako spolu nesúvisia** (random)

Model generuje nezávisle **jednotlivé bázy** v oboch sekvenciách.

Dané pravdepodobnosti $p(x)$, že vygeneruje bázu x .

Pravdepodobnosť vygenerovania X zarovnaného s Y :

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

Príklad: Uvažujme iba bázy A, T

Model R: $p(A) = 0.55$, $p(T) = 0.45$

$$\Pr(AAAT, ATAT | R) = 0.55^5 \cdot 0.45^3 \approx 0.00459$$

Skórovanie založené na dvoch pravdepodobnostných modeloch

Porovnáваме dva modely H a R : skóre bude logaritmus podielu

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

- Ak sú sekvencie X a Y **homológy**
⇒ logaritmus podielu oveľa väčšieho ako 1 ⇒ **veľmi pozitívne skóre**
- Ak sekvencie X a Y **nesúvisia**
⇒ logaritmus podielu oveľa menšieho ako 1 ⇒ **veľmi negatívne skóre**

Príklad:

$$\Pr(\text{AAAT}, \text{ATAT} | H) = 0.0072$$

$$\Pr(\text{AAAT}, \text{ATAT} | R) \approx 0.00459$$

$$\log_2 \frac{\Pr(X, Y | H)}{\Pr(X, Y | R)} \approx \log_2 \frac{0.0072}{0.00459} \approx \log_2 1.5699 \approx 0.6507$$

Základ logaritmu môžeme zvoliť hocijaký, tu 2.

Skórovanie založené na dvoch pravdepodobnostných modeloch

Porovnáваме dva modely H a R : skóre bude logaritmus podielu

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

- Ak sú sekvencie X a Y **homológy**
⇒ logaritmus podielu oveľa väčšieho ako 1 ⇒ **veľmi pozitívne skóre**
- Ak sekvencie X a Y **nesúvisia**
⇒ logaritmus podielu oveľa menšieho ako 1 ⇒ **veľmi negatívne skóre**

Príklad:

$$\Pr(\text{AAAT}, \text{TTAA} | H) = 0.00135$$

$$\Pr(\text{AAAT}, \text{TTAA} | R) \approx 0.00459$$

$$\log_2 \frac{\Pr(X, Y | H)}{\Pr(X, Y | R)} \approx \log_2 \frac{0.000135}{0.00459} \approx \log_2 0.2944 \approx -1.7643$$

Ako použiť takéto skóre v N.-W. a S.-W. algoritmoch?

Potrebuje ho rozbiť na **súčet** skór pre zarovnané dvojice písmen.

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

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

$$\begin{aligned} \log \frac{\Pr(X, Y | H)}{\Pr(X, Y | R)} &= \log \frac{\prod_{i=1}^n p(x_i, y_i)}{\left(\prod_{i=1}^n p(x_i)\right) \left(\prod_{i=1}^n p(y_i)\right)} \\ &= \log \prod_{i=1}^n \frac{p(x_i, y_i)}{p(x_i) \cdot p(y_i)} = \sum_{i=1}^n \log \frac{p(x_i, y_i)}{p(x_i) \cdot p(y_i)} \end{aligned}$$

Skóre zarovnanie bázy x a bázy y :

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

Ako použiť takéto skóre v N.-W. a S.-W. algoritmoch?

Skóre zarovnanie bázy x a bázy y :

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

Príklad:

Model H: $p(A, A) = 0.4$, $p(A, T) = 0.15$, $p(T, A) = 0.15$, $p(T, T) = 0.3$

Model R: $p(A) = 0.55$, $p(T) = 0.45$

$$s(A, A) = \log_2 \frac{0.4}{0.55 \cdot 0.55} \approx 0.4031$$

$$s(A, T) = \log_2 \frac{0.15}{0.55 \cdot 0.45} \approx -0.7225$$

$$s(T, A) = \log_2 \frac{0.15}{0.45 \cdot 0.55} \approx -0.7225$$

$$s(T, T) = \log_2 \frac{0.3}{0.45 \cdot 0.45} \approx 0.5670$$

Pre AAAT, ATAT skóre $0.4031 - 0.7225 + 0.4031 + 0.5670 = 0.6507$

Pre AAAT, TTAA skóre $-0.7225 - 0.7225 + 0.4031 - 0.7225 = -1.7644$

BLOSUM62 skórovacia matica pre proteíny

BLOcks of aminoacid SUbstitution Matrix; 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 zarovnanie** proteínov (BLOCKS)
- Páry s nanajvyš 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** : $\log \frac{p(x, y)}{p(x)p(y)}$

- pre násobíme 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

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

Zhrnutie

- Globálne a lokálne zarovania
- Needlemanov-Wunschov a Smithov-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?