

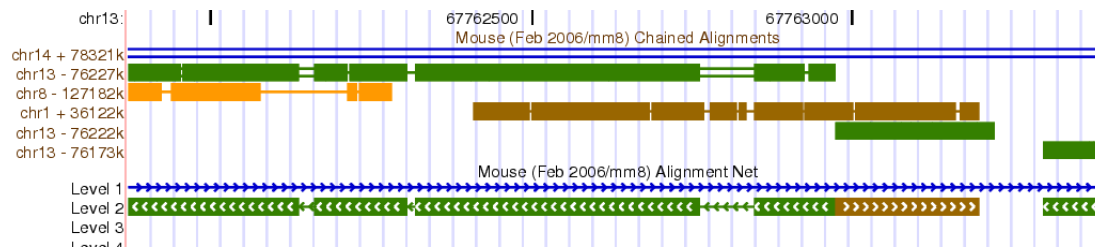
Oznamy

- Výber článku na journal club formulárom na stránke do utorka 23.10.
- Prvá domáca úloha: zadanie na budúci týždeň, čas na vypracovanie cca 2 týždne (detaily nabudúce)
- Ďalšie info <http://compbio.fmph.uniba.sk/vyuka/mbi/>

Zarovnávane sekvencií (sequence alignment) 1/2

Tomás Vinar

11.10.2018



[Durbin et al., 1998, kapitola 2]

Problém: Lokálne zarovnanie (local alignment)

ggcccttggagttgactgtcctgctgctccttgagg
ccattctcagagagaggaagtggcctcattttaatc
cgcttcccacagccttgtcctttccagacccatggg
agagggaggggctgagggtgtggctgagcccacca
agtcacgcgctcactctgcaggtccctctccccaag
gccgtggccttgggagcccgtggatcccagtgagtg
acgcctccacccccgacctactcgggcagtttaac
ccttgttgttcacttgcagacatcgtgaacacggcc
cggcccgacgagaaggccataatgacctatgtgtcc
agcttctaccatgccttttcaggagcgcagaaggta
ccgagcagggccaggcaggccctcctcgccgccacc
gcgcaatgccgcccgtgcctctcgctcccgtgctc
acctcatttctcttgcagacggcagtggcctctctc
caactggaagccacccccagctccct...

tgatgccgaggatgtgttcgctcgagcatccggacga
gaagtcacacctacgtggtcacactactatcacta
cttagcaaaactcaagcaggagacggcagggcat
aagcgtatcggttaagggtggcattgccatggag
aacgacaaaatggccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgacctccag
tcgctgggagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcacctcgagaagccgccaagtttgtggaaaag
ggcaacctcgaggtgctccttttcacctgcagtc
aagatgcgggccaacaaccagaagccctacacacc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggccctgcgcgaggagctcatccg...

Vstup: dve sekvencie

Problém: Lokálne zarovnávanie (local alignment)

```
ggccttggagttgactgtcctgctgctccttgagg
ccattctcagagagaggaagtggcctcattttaatc
cgcttcccacagccttgtcctttccagacccatggg
agagggaggggctgagggtgtggctgagcccacca
agtcacgcgtcactctgcaggtccctctccccaag
gccgtggccttgggagcccgtggatcccagtgagtg
acgcctccacccccgcctactcgggcagtttaac
ccttgttgttcacttgcagacatcgtgaacacggcc
cggcccgacgagaaggccataatgacctatgtgtcc
agcttctaccatgccttttcaggagcgcagaaggta
ccgagcagggccaggcaggccctcctcgccgccacc
gcgcaatgccgccgctgcctctgcctcccgtgctc
acctcatttctcttgagacggcagtggcctctctc
caactggaagccacccccagctccct...
```

```
tgatgccgaggatgtgttcgtcgagcatcggacga
gaagtccatcacctacgtggtcacctactactacta
cttagcaaaactcaagcaggagacggcgcagggcat
aagcgtatcggtaagggtggcggcattgccatggag
aacgacaaaatgggtccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgaccatccag
tcgctgggagcagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcaccatcgagaagccgccaagtttgtggaaaag
ggcaacctcgaggtgctccttttcacctgcagtcc
aagatgcgggccaacaaccagaagccctacacacc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggcctgcgcgaggagctcatccg...
```

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 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 inzerciam a deléciám, zarovnané bázy zachovaným bázam a substitúciám.

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

Menia sa pomalšie ako ostatné sekvencie.

Formulácia problému

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

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

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

V praxi zložitejšie skórovanie. Chceme nastaviť tak, aby homológy mali vysoké skóre, náhodné zarovnaní nízke.

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: zarovnaní podreťazcov $x_i \dots x_j$ a $y_k \dots y_l$ 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.

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

Všeobecný prípad, $i > 0, j > 0$:

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

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

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

Rekurencia:

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

Príklad globálneho zarovnania

CATGTCATA 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										

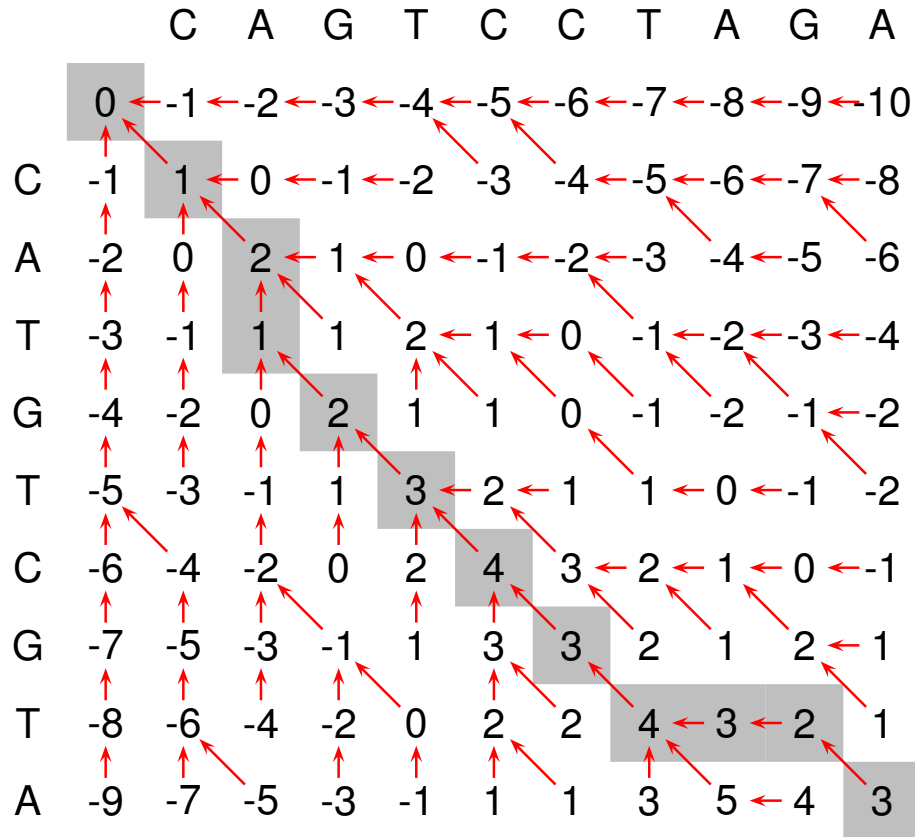
$$A[i, j] = \max \begin{cases} A[i - 1, j - 1] + s(x_i, y_j), \\ A[i - 1, j] - 1, \\ A[i, j - 1] - 1 \end{cases}$$

Príklad globálneho zarovnania

CATGTCATA 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

CATGTCAT--A

Dynamické programovanie pre lokálne zarovnanie (Smith, Waterman 1981)

Podproblém: $A[i, j]$: najvyššie skóre lokálneho zarovnanie 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 zarovnanie s kladným skóre $A[i, j] = 0$

Rekurencia: $A[i, j] = \max \left\{ \begin{array}{l} 0, \\ A[i - 1, j - 1] + s(x_i, y_j), \\ A[i - 1, j] - 1, \\ A[i, j - 1] - 1 \end{array} \right.$

Príklad lokálneho zarovnania

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

CA-GTCCTA

CATGTCATA

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?

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

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

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

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

Zložitejšie skórovanie: dva pravdepodobnostné modely

(Pre jednoduchosť teraz neuvažujme medzery)

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

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

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

Model R: Sekvencie X a Y nijako spolu nesúvisia

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

$p(x_i)$: pravdepodobnosť výskytu bázy x_i

Porovnanie modelov H a R: “log likelihood”

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

Zložitejšie skórovanie: dva pravdepodobnostné modely

Porovnanie modelov H a R : “log likelihood”

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

- Dve sekvencie sú **homológy**
 - ⇒ pomer pravdepodobností je oveľa väčší ako 1
 - ⇒ **veľmi kladné skóre**
- Dve sekvencie **nesúvisia**
 - ⇒ pomer pravdepodobností je oveľa menší ako 1
 - ⇒ **veľmi zaporné skóre**

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

BLOSUM62 skórovacia matica pre proteíny

BLOcks of aminoacid **S**Ubstitution **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 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** : $\log \frac{p(x, y)}{p(x)p(y)}$
- 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

→ 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 porovnávaní 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?