Metódy v bioinformatike

Zarovnávanie sekvencií

Jana Černíková

FMFI UK

10/10/2024

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
ggcaacctcgaggtgctccttttcaccctgcagtcc
aagatgcgggccaacaaccagaagccctacacaccc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
ttggccctgcgcgaggagctcatccg...
```

Vstup: dve sekvencie

Problém: Lokálne zarovnávanie (local alignment)

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

```
tgatgccgaggatgtgttcgtcgagcatccggacga
gaagtccatcacctacgtggtcacctactatcacta
ctttagcaaactcaagcaggagacggtgcagggcat
aagcgtatcggtaaggtggtcggcattgccatggag
aacgacaaaatggtccacgactacgagaacttcaca
agcgatctgctcaagtggatcgaaacgaccatccag
tcgctgggcgagcgggagttcgaaaactcgctggcc
ggcgtccaagggcagttggcccagttctccaactac
cgcaccatcgagaagccgcccaagtttgtggaaaag
ggcaacctcgaggtgctccttttcaccctgcagtcc
aagatgcgggccaacaaccagaagccctacacaccc
aaagagggcaagatgatttcggacatcaacaaggcc
tgggagcgtctggagaaggccgagcacgaacgcgaa
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 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 inzerciám a deléciám, zarovnané bázy zachovaným bázam a substitúciám.

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, nájdi zarovnanie s najlepším skóre
  (Optimalizačná fáza je téma dnešnej prednášky.)

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 $X = x_1 x_2 \dots x_n$ a $Y = y_1 y_2 \dots y_m$.
Výstup: zarovnanie X a Y s najvyšším skóre.

Problém 2: lokálne zarovnanie (local alignment)
Vstup: sekvencie $X = x_1 x_2 \dots x_n$ a $Y = y_1 y_2 \dots y_m$.
Výstup: zarovnania 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 zarovnania reťazcov $x_1x_2\ldots x_i$ a $y_1y_2\ldots y_j$.

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										

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

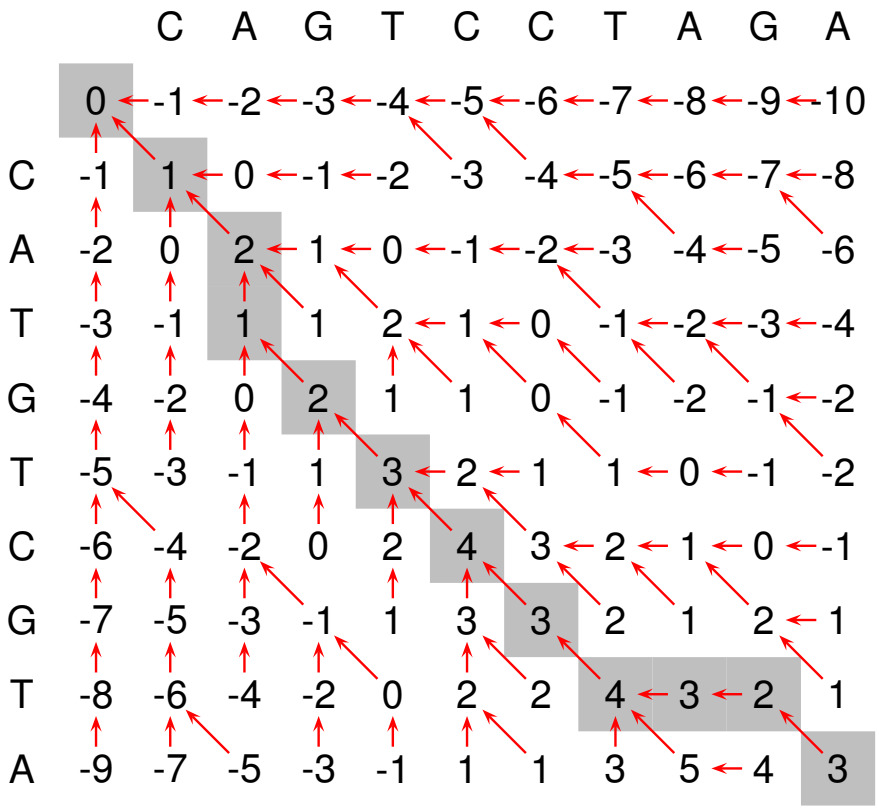
Príklad globálneho zarovnania

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

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

Ako získať zarovnanie?




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

Časová zložitosť celého algoritmu $O(nm)$

Dynamické programovanie pre lokálne zarovnanie (Smith, Waterman 1981)

Podproblém: $A[i, j]$: najvyššie skóre lokálneho zarovnania reťazcov $x_1 x_2 \ldots x_i$ a $y_1 y_2 \ldots y_j$, 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 $A[i, j] = 0$

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.

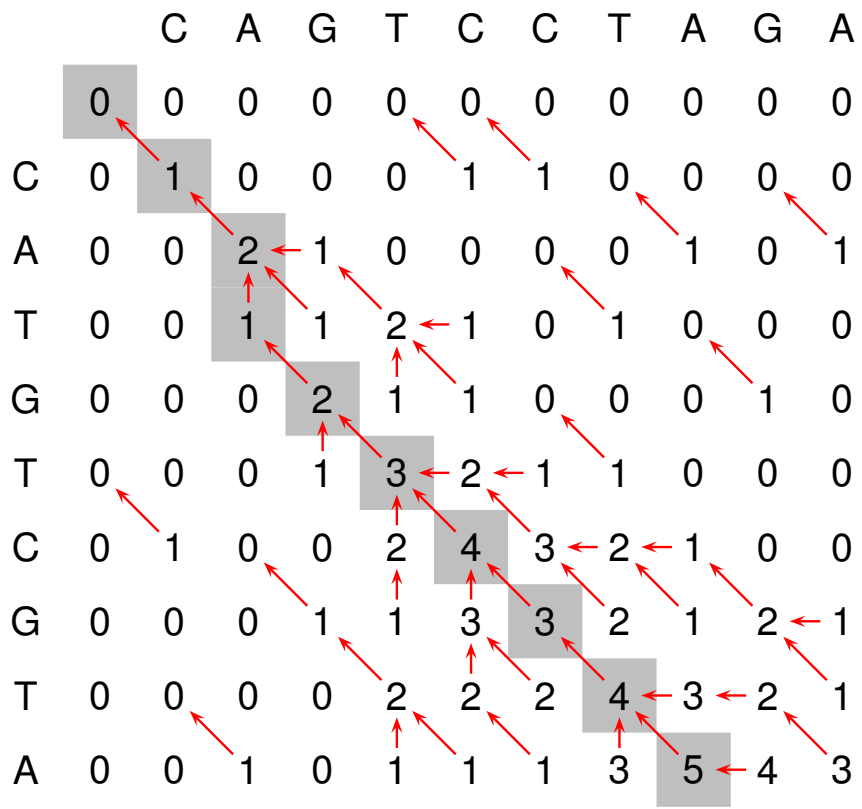
Všeobecný prípad, $i > 0$**,** $j > 0$**:**

- 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 zarovnania* od *horších zarovnaní*:
  - Ktoré usporiadania pomlčiek dávajú väčší zmysel
- Chceme vedieť, či dané zarovnanie *má biologický význam*:
  - Ide o homológy, alebo sú zarovnané len náhodou?

Povedali sme si:

- Globálne a lokálne zarovania
- Needlemanov-Wunschov a Smithov-Watermanov algoritmus

Pokračovanie prednášky

`https://youtu.be/0GkhkRiqbl4?feature=shared&t=2227`

- Skórovanie zarovnaní pomocou porovnávania modelov
- Proteínové BLOSUM matice
- Afínne skórovanie medzier

Metódy v bioinformatike

CB #3 Zarovnávanie sekvencií

Jana Černíková

FMFI UK

10/10/2024

Globálne zarovnanie

Uvažujme skórovanie zhoda +3, nezhoda -1, medzera -2
Reťazce TAACGG a CACACT

$$A[i,j] = \max \begin{cases} A[i-1,j-1] + s(x_i, y_j), \\ A[i-1,j] - 2, \\ A[i,j-1] - 2 \end{cases}$$

$s(x_i, y_j) = 3$ ak $x_i = y_j$,
$s(x_i, y_j) = -1$ ak $x_i \neq y_j$

$A[i,0] = -2i$,
$A[0,j] = -2j$

Globálne zarovnanie

		C	**A**	**C**	**A**	**C**	**T**
	0	-2	-4	-6	-8	-10	-12
T	-2						
A	-4						
A	-6						
C	-8						
G	-10						
G	-12						

Globálne zrovnanie

		0	1	2	3	4	5	6
			C	A	C	A	C	T
0		0	-2	-4	-6	-8	-10	-12
1	T	-2	-1	-3	-5	-7	-9	-7
2	A	-4	-3	2	0	-2	-4	-6
3	A	-6	-5	0	1	3	1	-1
4	C	-8	-3	-2	3	1	6	4
5	G	-10	-5	-4	1	2	4	5
6	G	-12	-7	-6	-1	0	2	3

CACACT-
TA-ACGG
alebo
CACAC-T
TA-ACGG

Lokálne zarovnanie

Uvažujme skórovanie zhoda +3, nezhoda -1, medzera -2
Reťazce TAACGG a CACACT

$$A[i,j] = \max \begin{cases} 0, \\ A[i-1,j-1] + s(x_i, y_j), \\ A[i-1,j] - 2, \\ A[i,j-1] - 2 \end{cases}$$

$s(x_i, y_j) = 3$ ak $x_i = y_j$,
$s(x_i, y_j) = -1$ ak $x_i \neq y_j$

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

Lokálne zarovnanie

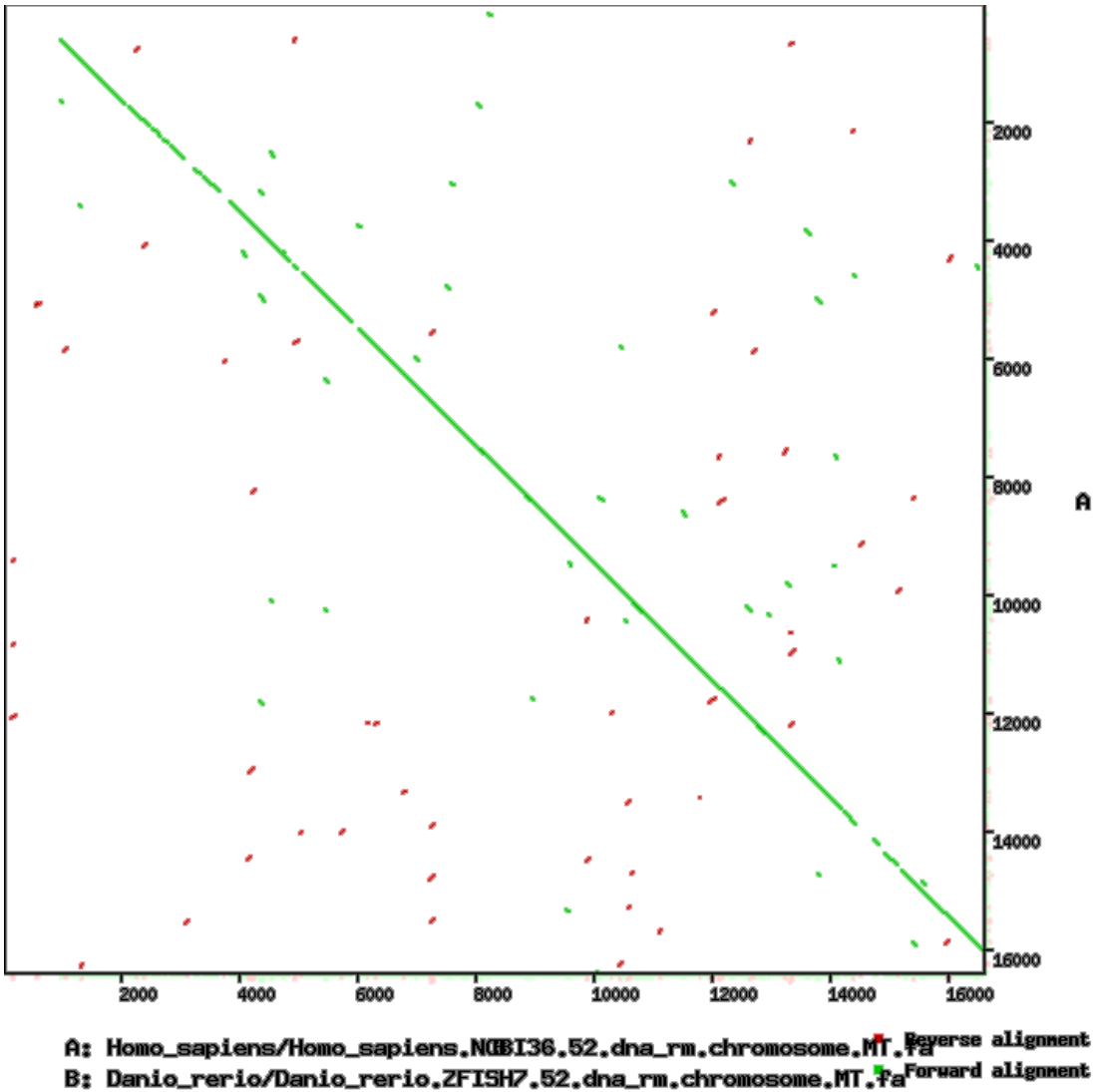
		C	**A**	**C**	**A**	**C**	**T**
	0	0	0	0	0	0	0
T	0						
A	0						
A	0						
C	0						
G	0						
G	0						

Lokálne zarovnanie

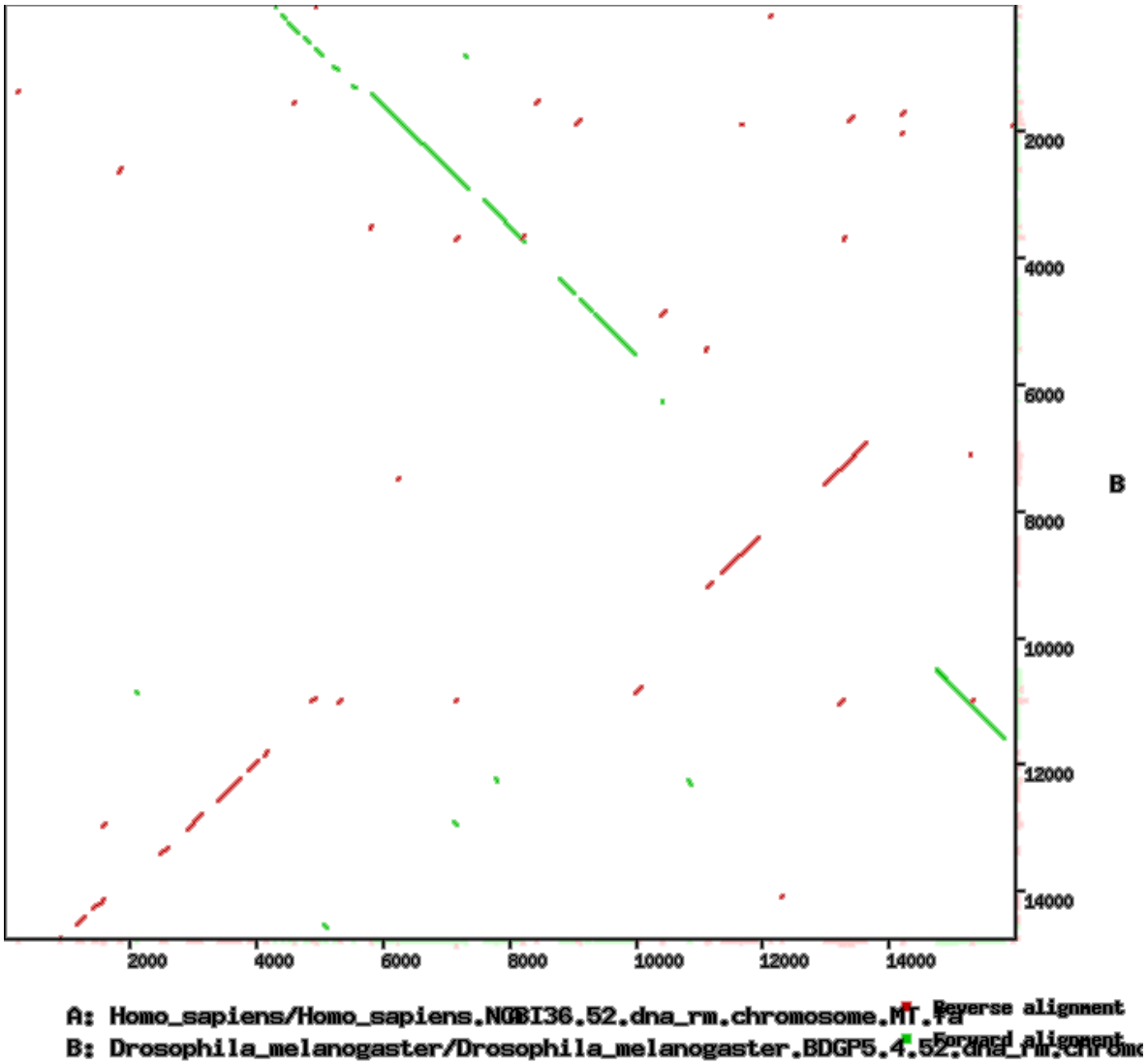
		0	1	2	3	4	5	6
			C	A	C	A	C	T
0		0	0	0	0	0	0	0
1	T	0	0	0	0	0	0	3
2	A	0	0	3	1	3	1	1
3	A	0	0	3	2	4	2	0
4	C	0	3	1	6	4	7	5
5	G	0	1	2	4	5	5	6
6	G	0	0	0	2	3	4	4

ACAC
A-AC

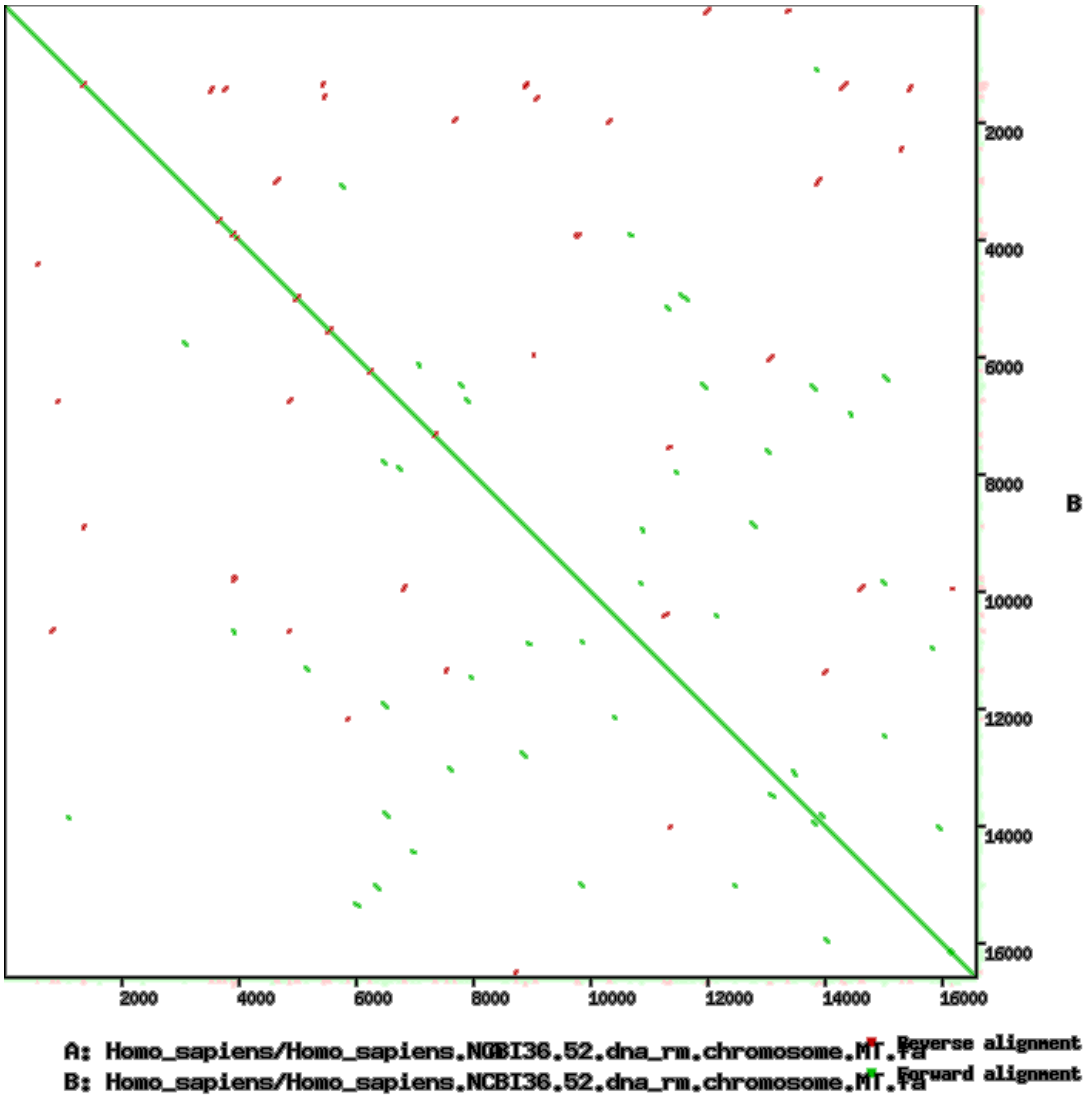
Mitochondriálny genóm človeka vs. ryba Danio rerio



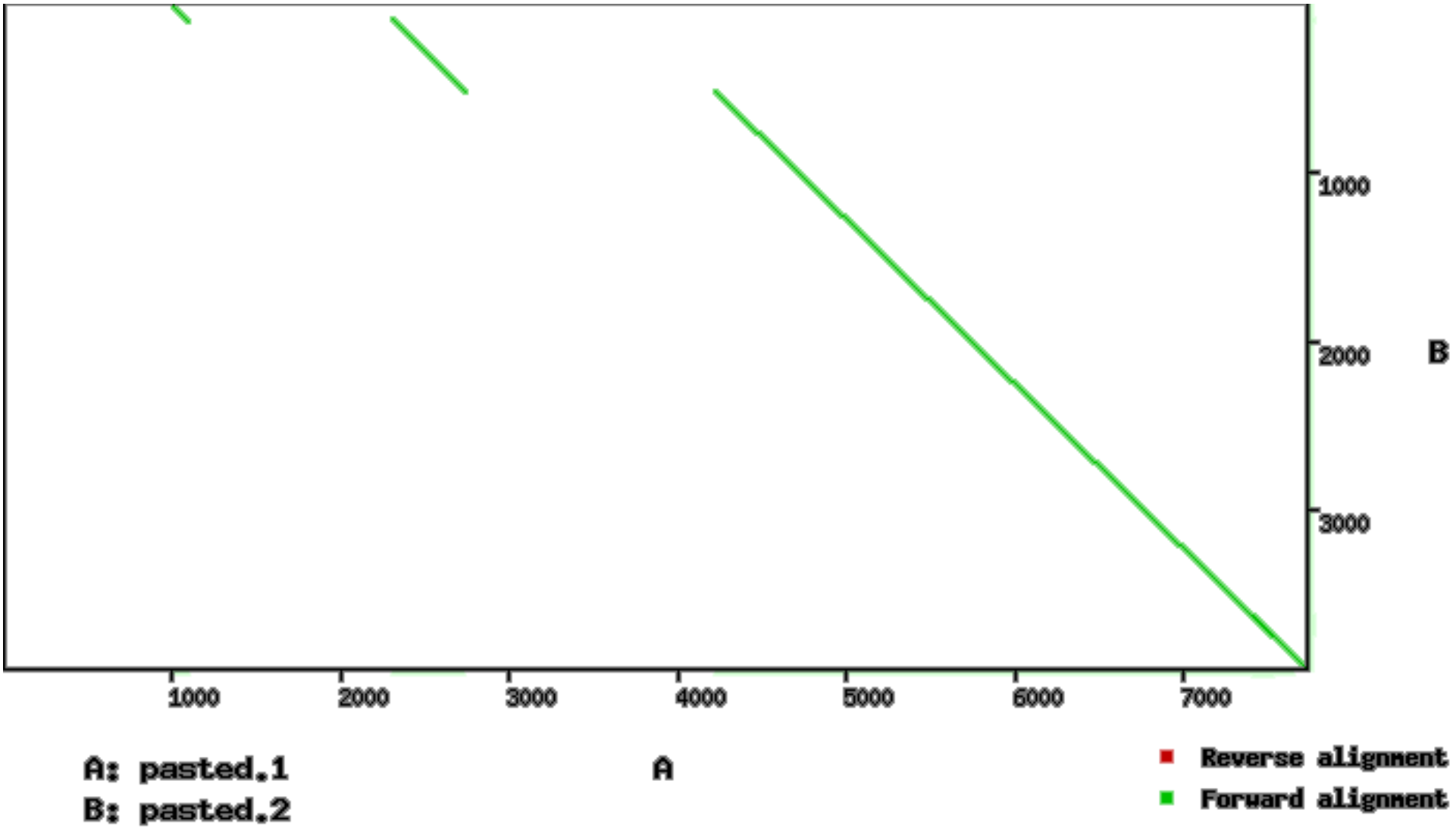
Mitochondriálny genóm človeka vs. Drosophila melanogaster



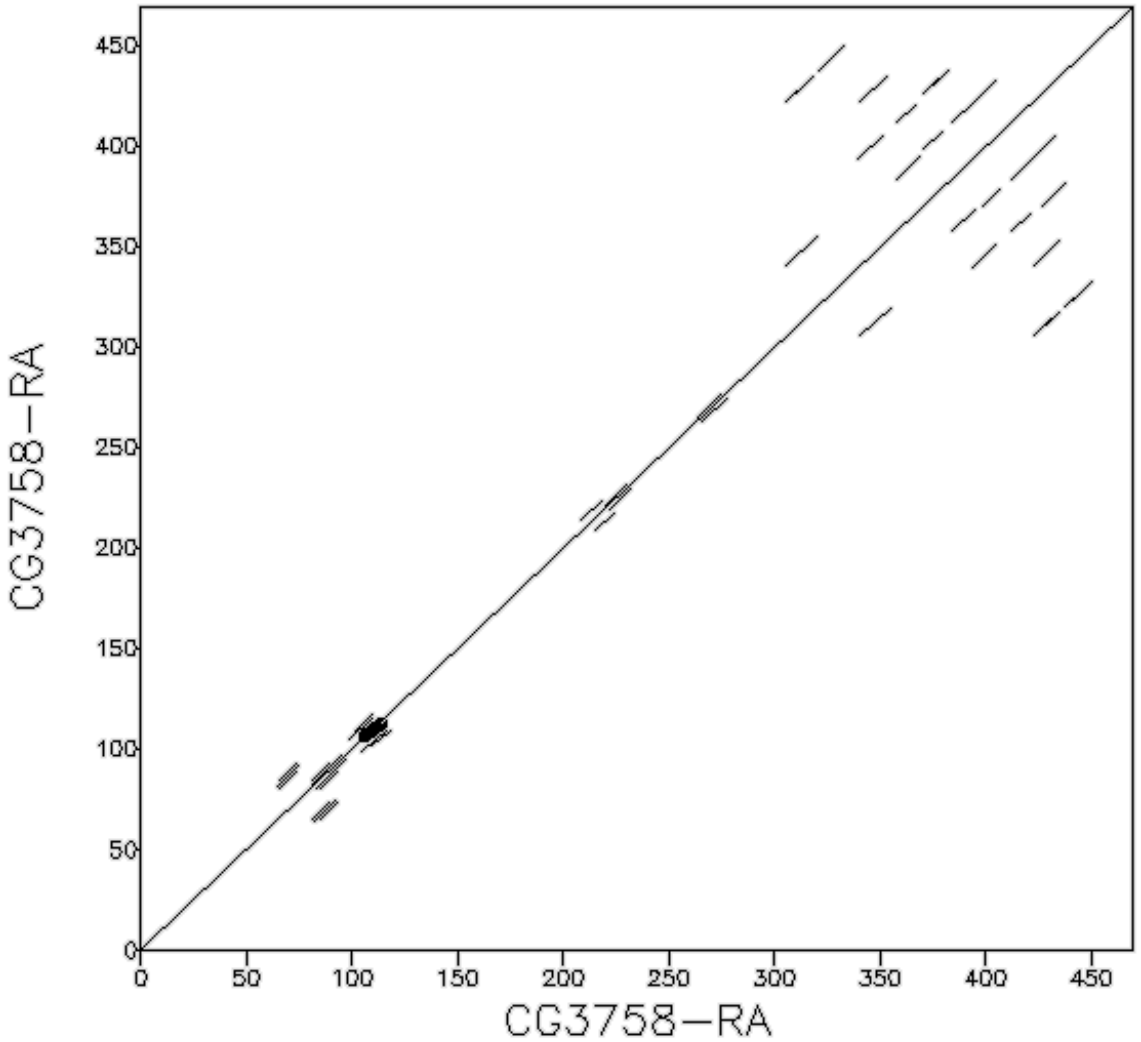
Mitochondriálny genóm človeka vs. to isté



Drosophila mRNA Oaz zinc finger vs. genomický usek (časť chr2R)



Drosophila proteín Escargot zinc finger vs. to isté



Zhluk génov PRAME v človeku vs. to isté