Jadrá zarovnaní

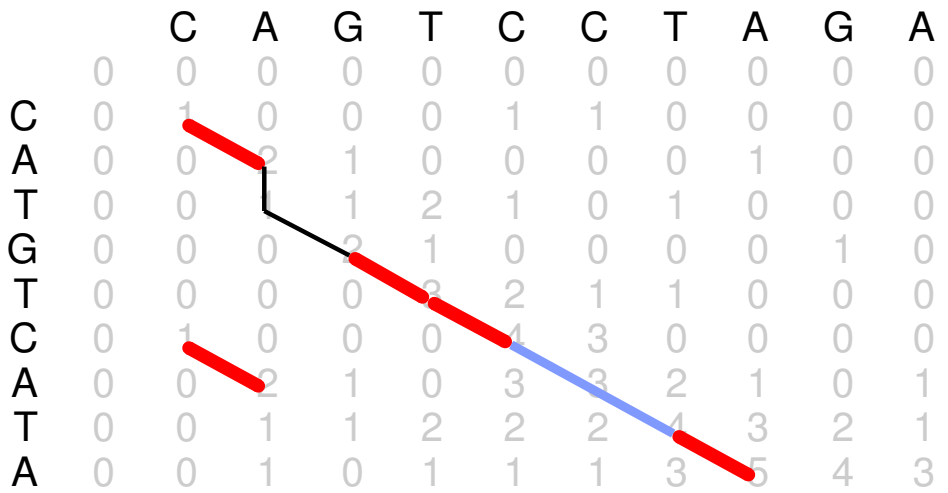
Tomáš Vinař

20.10.2016

Opakovanie: Heuristické lokálne zarovnávanie, BLAST

Príklad: $w = 2$ (začíname z jadier dĺžky 2).

(V praxi sa používa $w = 10$ a viac.)



1. nájdi zhodné úseky
2. rozšír bez medzier
3. spoj medzerami

Senzitivita heuristického algoritmu

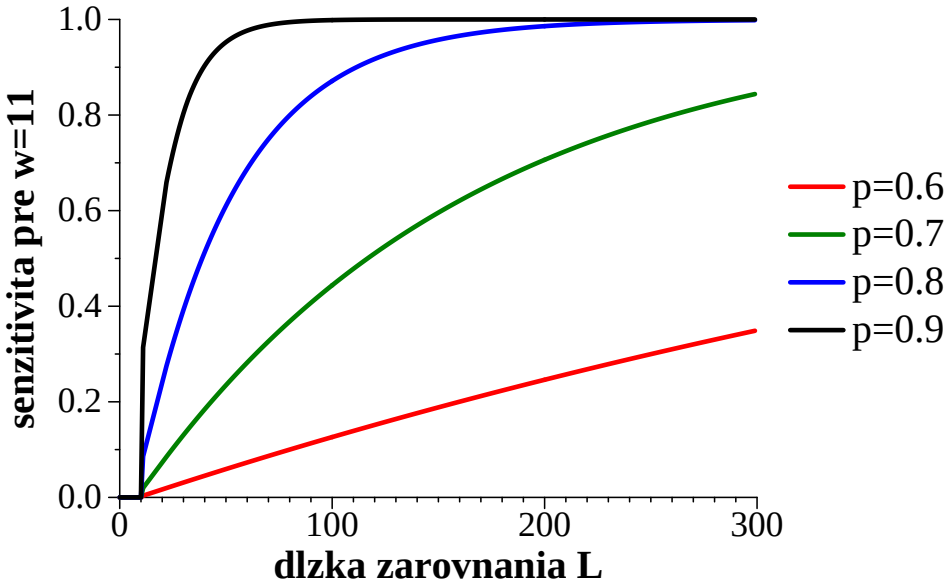
Odhad senzitivity:

Predpokladáme zarovnanie bez medzier, dĺžky L

Každá pozícia je zhoda s pravdepodobnosťou p

Senzitivita:

$$f(L,p) = \Pr(\text{zarovnanie obsahuje } w \text{ zhôd za sebou})$$



Jadrá z medzerami, spaced seeds

PatternHunter [Ma, Tromp, Li 2002]

Jadro s medzerami: vyžadovaná konfigurácia zhôd

Príklad:

"match—match—don't care—match" značíme ako `1101`

```
GTGGTGCTCTCTGACAAAGCC
 | || ||   ||| ||||
ATTGTTCTTAATGAGAAAGAA
   1101     1101
               1101
```

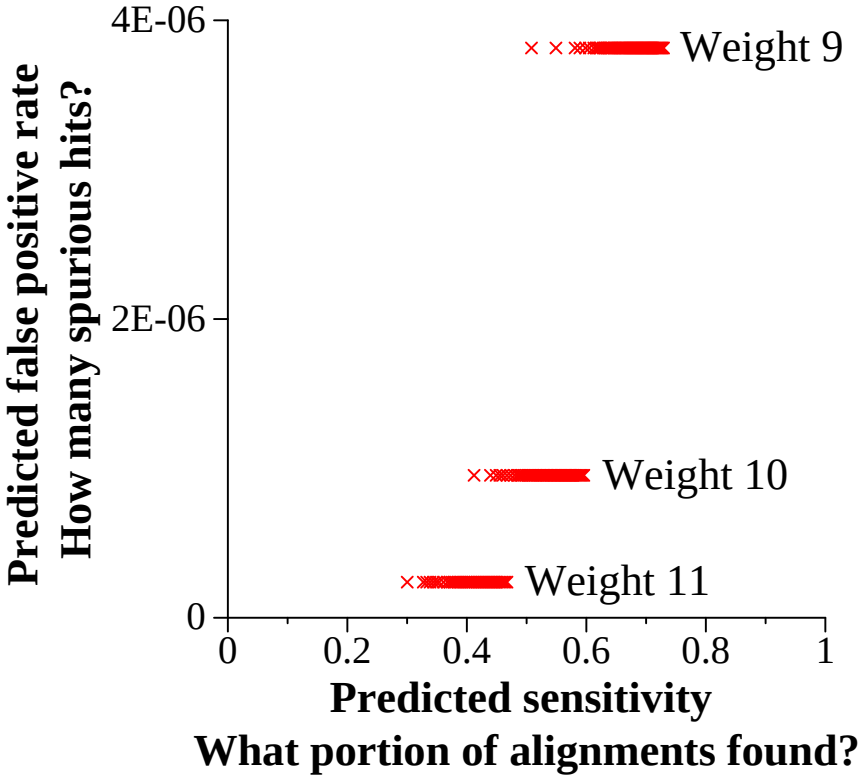
BLASTN jadro (11 za sebou idúcich zhôd)

ekvivalentné jadru `11111111111`

Nie všetky jadrá sú rovnaké

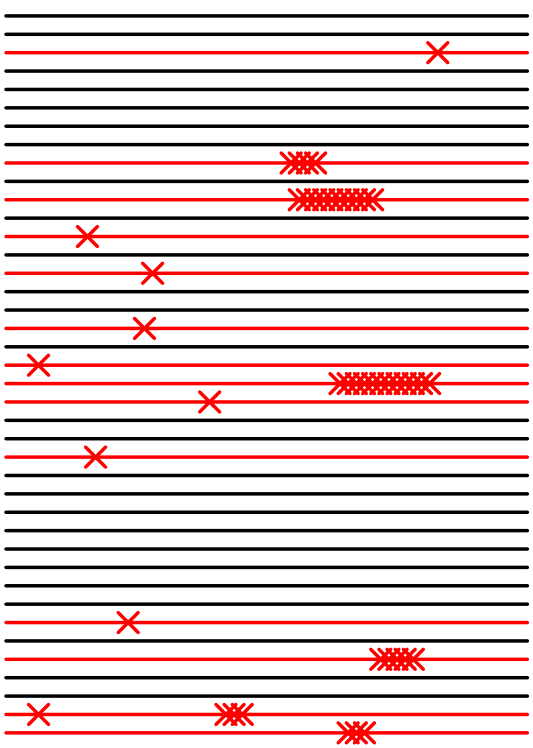
Váha jadra: počet vyžadovaných zhôd

Každý krížik: senzitivita vs. čas pre jedno jadro v pravdepodobnostnom modeli



Prečo sú jadrá s medzerami senzitívnejšie?

Príklad: dĺžka zarovnania $L = 64$,
pravdepodobnosť zhody $p = 0.7$ na každej pozícii
40 náhodných zarovnaní, výskyty jadra



11111111111
Sn.: 14/40, hits: 46

11101001010011 0111
Sn.: 18/40, hits: 35

Prečo sú jadrá s medzerami senzitívnejšie?

Príklad: dĺžka zarovnania $L = 64$,
pravdepodobnosť zhody $p = 0.7$ na každej pozícii

Bez medzier	**S medzerami**
11111111111	111010010100110111
Stredná hodnota počtu výskytov v zarovnaní:	
$54 \cdot 0.7^{11} = 1.1$	$47 \cdot 0.7^{11} = 0.9$
Pravdepodobnosť výskytu na poz. $i+1$ ak výskyt na i:	
0.7	$0.7^6 = 0.12$
1111111111 1111111111	111010010100110111 111010010100110111
Výskyty často vedľa seba	Výskyty "nezávislejšie"
Senzitivita (pravdepodobnosť aspoň jedného výskytu):	
0.30	0.47

Ďalšie hašovacie stratégie

Nukleotidový BLAST: 10 zhôd za sebou

Jadro s medzerami: povoľuje nezhody na 8 z 18 pozícií

BLAT [Kent 2002]: povoľuje 1 nezhodu na ľub. z 11 pozícií

BLASTP: 3 amino kyseliny so skóre aspoň 13 v matici BLOSUM62

Výskyt:

```
N I R
N L R

6+2+5=13
```

Nie výskyt:

```
A I L
A I L

4+4+4=12
```

Vektorové jadrá: kombinácia jadier s medzerami a BLAT/BLASTP

Viaceré výskyty: začni rozširovať iba ak viac výskytov blízko seba na tej istej uhlopriečke

Viaceré jadrá: zober zjednotenie výskytov

Záleží na modeli zarovnania

Pravdepodobnosť zhody kolíše v rámci kodónu:

Poloha v kodóne:	prvá	druhá	tretia
Pravdepodobnosť zhody:	0.67	0.77	0.40

Senzitivita na testovacej vzorke exónov kódujúcich proteíny:

Jadro		Človek vs. Drosophila	myš
Optimálne pre dáta	**110 110 000 110 110 11**	86%	92%
Optimalne pre kodónový model	**110 110 010 110 010 11**	86%	91%
WABA [Kent, Zahler 2000]	**110 110 110 110 11**	80%	90%
Optimálne pre i.i.d. model	**1110010010010101111**	60%	86%
BLAST	**1111111111**	43%	81%
Najhoršie	**1010101010101011**	39%	79%

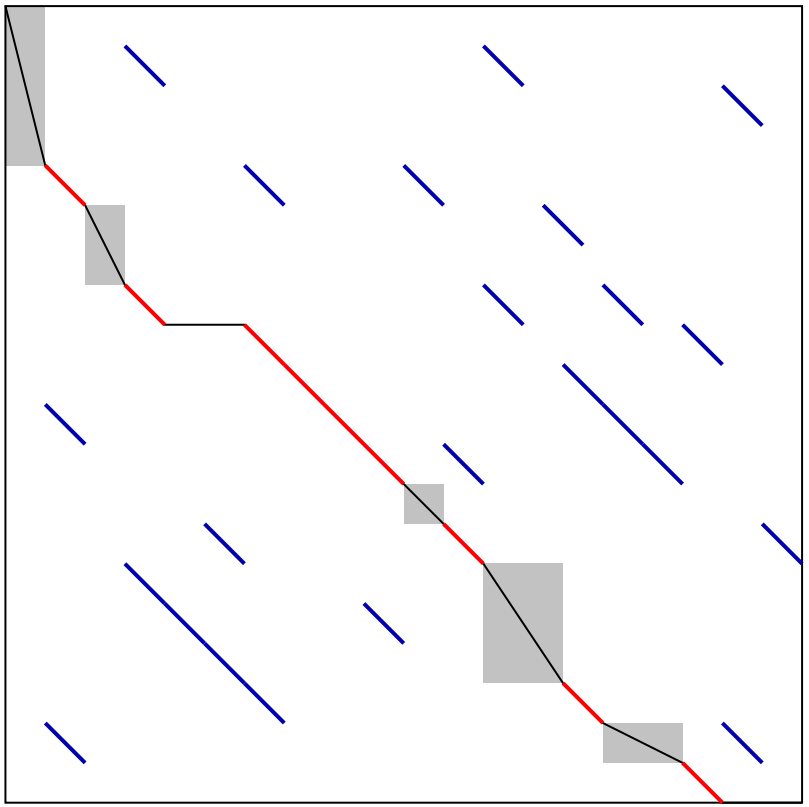
A čo globálne zarovnania?

Ukotvené zarovnanie (Anchored alignment)

- Nájdime lokálne zarovnania (alebo výskyty nejakého jadra)
  – možné **ukotvenia**
- Zvoľ konzistentnú množinu ukotvení
  (monotónna postupnosť)
- Zarovnaj časti sekvencií medzi ukotveniami
  (pomocou dyn. prog. alebo rekurzívne ďalším kotvením)

MUMMER [Delcher 1999]
GLASS [Batzoglou et al 2000]
AVID [Bray et al 2003]

Ukotvené zarovnanie



Modré: nezvolené ukotvenia

Červené: zvolené ukotvenia

Sivé: riešime dyn. prog.

Čierne: globálne zarovnanie

Znova protichodné vplyvy:

málo spoľahlivých ukotvení – dobrá kvalita, pomalé

veľa slabších ukotvení – rýchle (malá sivá plocha), viac chýb v ukotvení