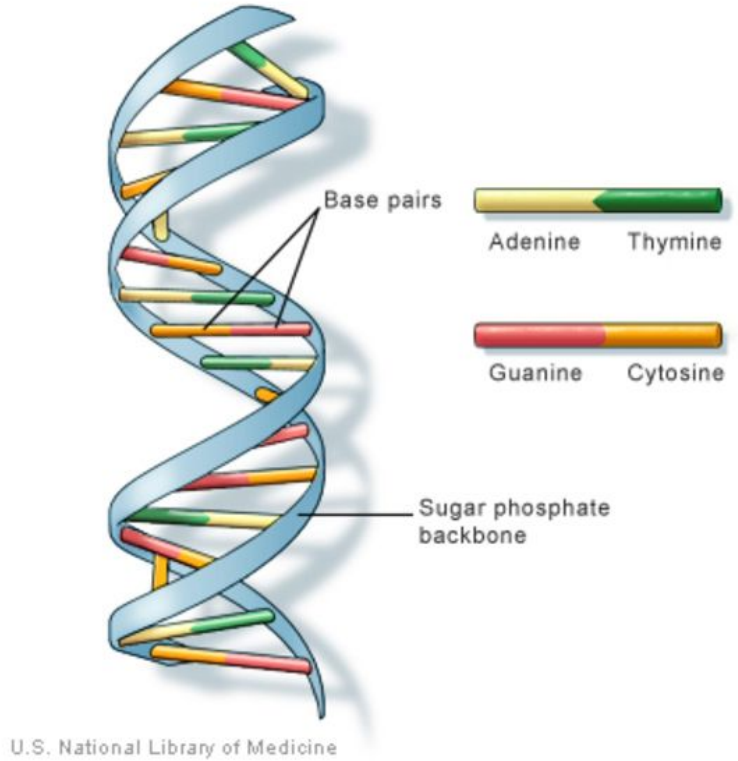
Algoritmy pre dynamické zostavovanie nanopórových čítaní

Jana Černíková
školiteľ: doc. Mgr. Tomáš Vinař, PhD.

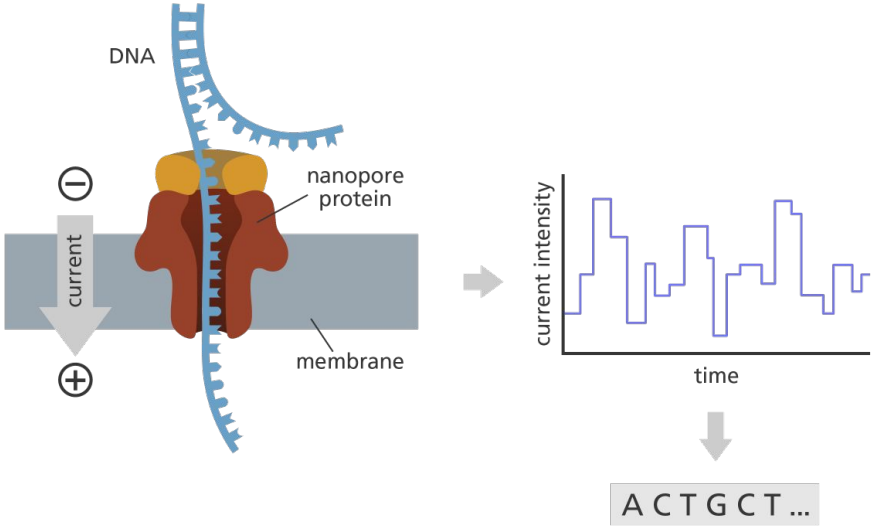
Biologické pozadie problému



TTGCCGCGCACTCGATATTGCGCTGCCGGAC
CGAGATTGCGGCCTGTCGCTGGGGTTACCGA
GGCAATGCCGACAGCGGCAATATCGGCCGGC
GCGCGAAAATCTCGCCGACAAAACCAGCGCA
CGTCGCCTTAATCAATGCGCCTGAATCTGGC
GGGATATGCGCAGTCGCCGACAGCGGCAATA
TCGGCCGGCGCGCGAAAATCTCGCCGACAAA
ACCAGCGCACGTCGCCTTAATCAAGTGGAAG
GAGATAGAGGATATACACACCACCACCACTGA
GATTTAATCAATGCGCCTGAATCTGGCGGGAT
ATGCGCAGTCGCCGACAGCGGCAATATCGGC
CGGCGCGCGAAAATCTCGCCGACAAAACCAG
CGCACGTCGCCTTAATCAAGTGGAAGGAGAT
AGAGGATATACG….

Oxford Nanopore Technologies - MinION

https://www.genengnews.com/insights/first-nanopore-sequencing-of-human-genome/

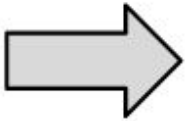


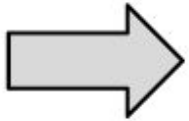
https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

<45g, 1000$
krátke (500bp) až veľmi dlhé (>4 Mb = 4,000,000 bp) čítania,
10-30 kb bežne

420 báz/s
50GB dát / 72h

Skladanie genómov

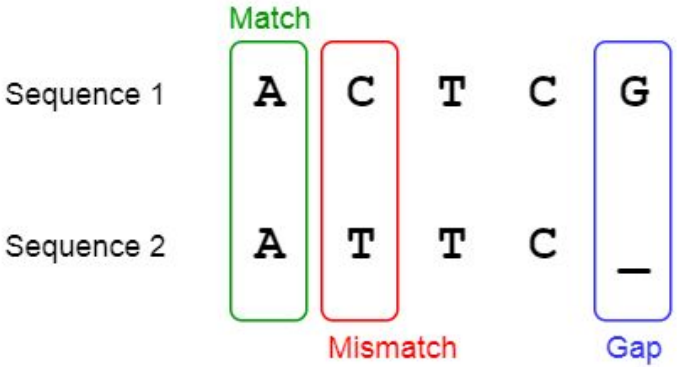


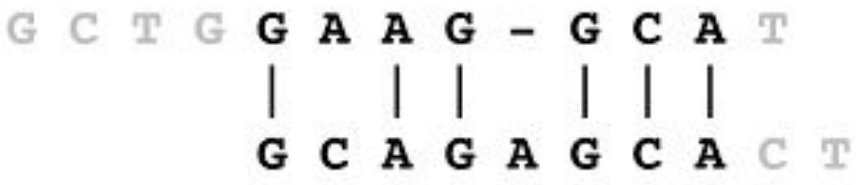


Čítania (reads)

@6bfee659-58a2-4870-ad8f-f37b1f37279f runid=37ed0c95ec68e8222ec9d629d5e1715ef411b250 read=101 ch=394 start_time=2018-10-10T09:41:10Z
CCCGTCCGTTCCATTCATTCCGATCATTCCGGTCCATTCCATTCCATTCCATCCATTCCATTCCAGGAGTCCATTCCATTCCAT
TCCATTCCTCGAGATCCGGTCAGTCCGGTCCATTCCATCCGTCCA
+
$&$&'(##$$#$"##$"#$%&&##$##$%%$#$%&"$$%&#$$%%#$$%%#$%%#%$$$"#$$%#####$%%"$$$%"$$$$"$$&%%$$$%#
%#####$%%$%$"#$%%$$$$$"%$$$##*%"'##"
@910d707b-a589-487b-9766-fe40fa349ce0 runid=37ed0c95ec68e8222ec9d629d5e1715ef411b250 read=101 ch=293 start_time=2018-10-10T09:39:04Z
TTGCCGCGCACTCGATATTGCGCTGCCGGACCGAGATTGCGGCCTGTCGCTGGGGTTACCGAGGCAATGCCGACAGCGGC
AATATCGGCCGGCGCGCGAAAATCTCGCCGACAAAACCAGCGCACGTCGCCTTAATCAATGCGCCTGAATCTGGCGGGATAT
GCGCAGTCGATTTGAGGTTCAAGAACCAGAAGAATACCGAGACATCTGGTCAAGGACGCCGGAATGTTTTCGGTCGAGAGAG
AACAGCGTCAGGAT….
+
###%%$%$%$&&(%$%&&%&&&&'%'&)($('&#'#+&%'$%%$%#$$$%%$).&*'$("&&$'(%##&%(&('&'%%&(&&&%&%$%')((&&%%$)
+,0&&#$#%$#$&)*)%&&#%'$&#&%'&$$$%$*'&$&$%#%$($$%$$)()'(%')%$$#&$'%&$'&"$+4*'&'("&$$'(',*%(*(&&$#$$$'$'$$%$$'+
--.-&,+$$$%$(,%(*%,+*)%%$%'&&&('"((+'%%$%&&'(&%$&&&&&&)#$')$%$$##&$)151&-&0%+3.**&&%$$&%&$('+$&)&$$%))*(&
&&)*$&)%'(&$$((,'*&&'%'&%'*+$%#%')3),)))*%')%$%%$#&$&&)))&(2)'$+)(+32,+,+((&$$$$%))"*(+,++$%$%&&($&&%&%%"*(&$#.
..

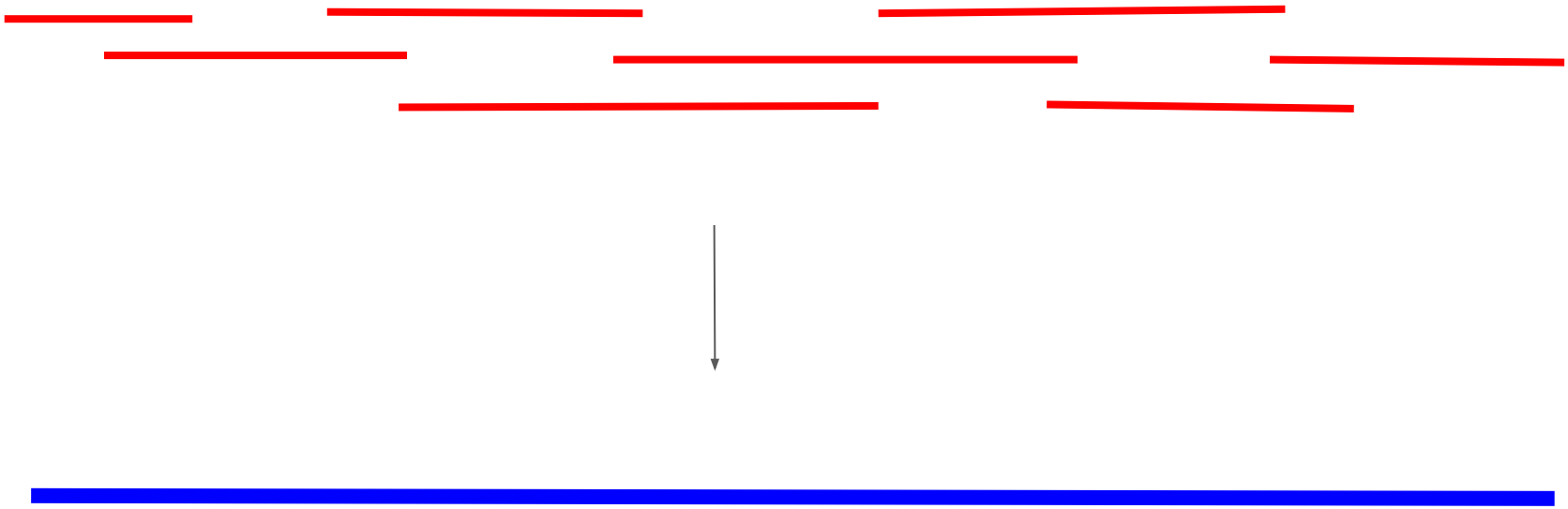
Zarovnanie (alignment)





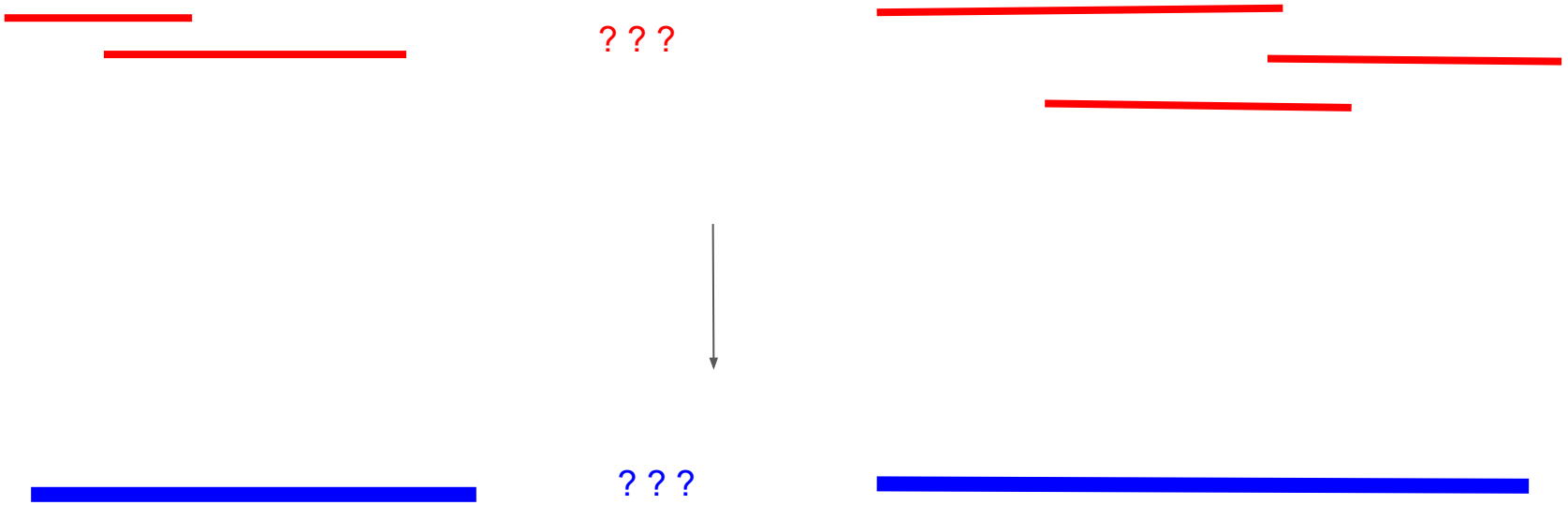
Zostavenie (assembly) genómu

(chceli by sme)



Zostavenie (assembly) genómu

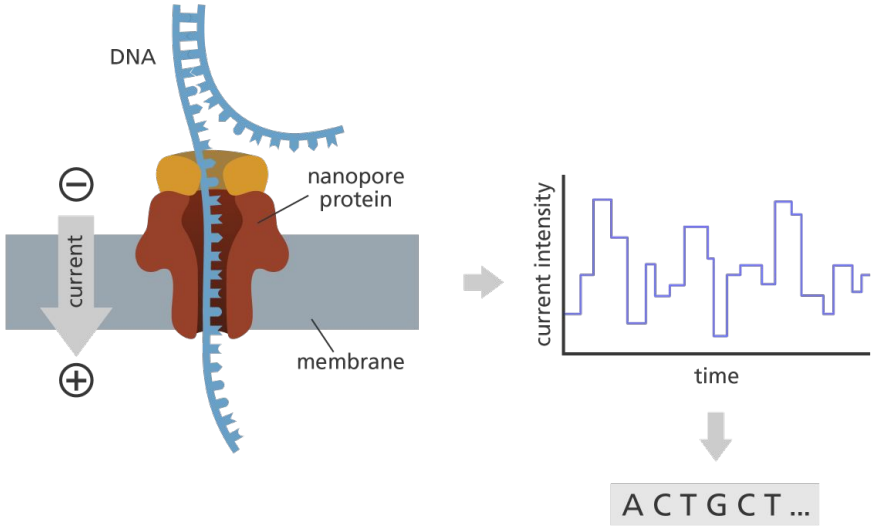
(máme)



poradie?
orientácia?

Statický vs. dynamický problém

https://www.genengnews.com/insights/first-nanopore-sequencing-of-human-genome/



https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

“Kedy máme dostatočné množstvo dát?”

cieľ práce: chceme odpoveď v reálnom čase

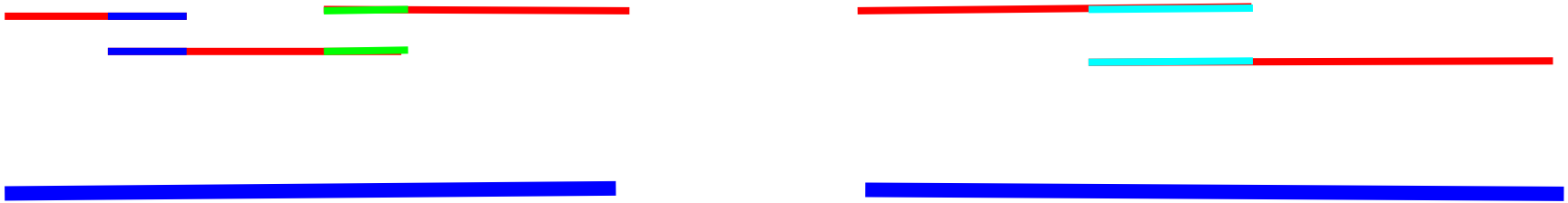
Minimap & Miniasm

Heng Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences , Bioinformatics, Volume 32, Issue 14, 15 July 2016, Pages 2103–2110, https://doi.org/10.1093/bioinformatics/btw152

- čiastočne implementuje overlap-layout(-consensus) prístup
- rieši “statický” problém - všetky dáta dostane naraz
- heuristiky + dynamické programovanie

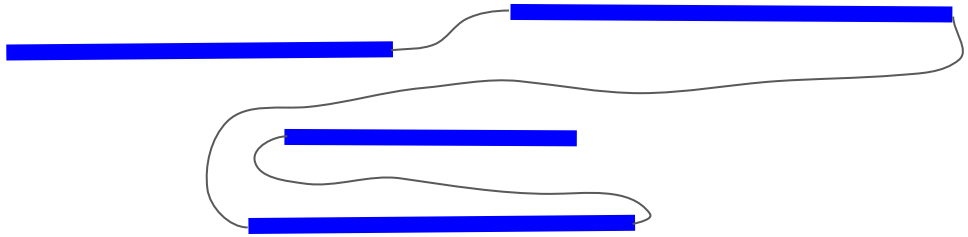
Overlap-layout-consensus [Minimap, Miniasm]

- **overlap:** hľadáme zarovnania (prekryvy), budujeme contigy



Overlap-**layout**-consensus [Miniasm]

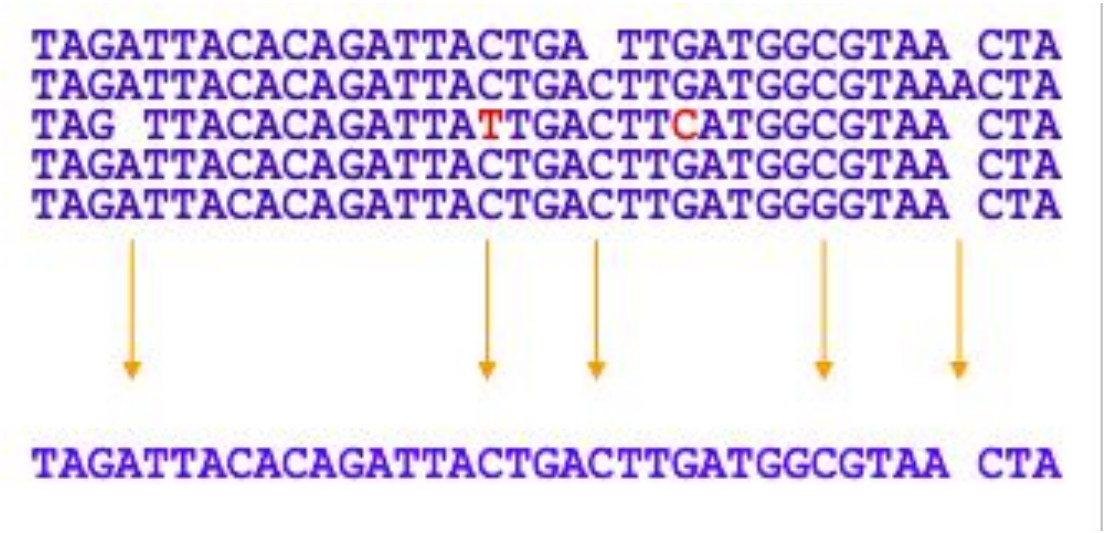
- **layout:** cieľ je zistiť relatívnu orientáciu contigov, vzdialenosti medzi nimi



supercontig

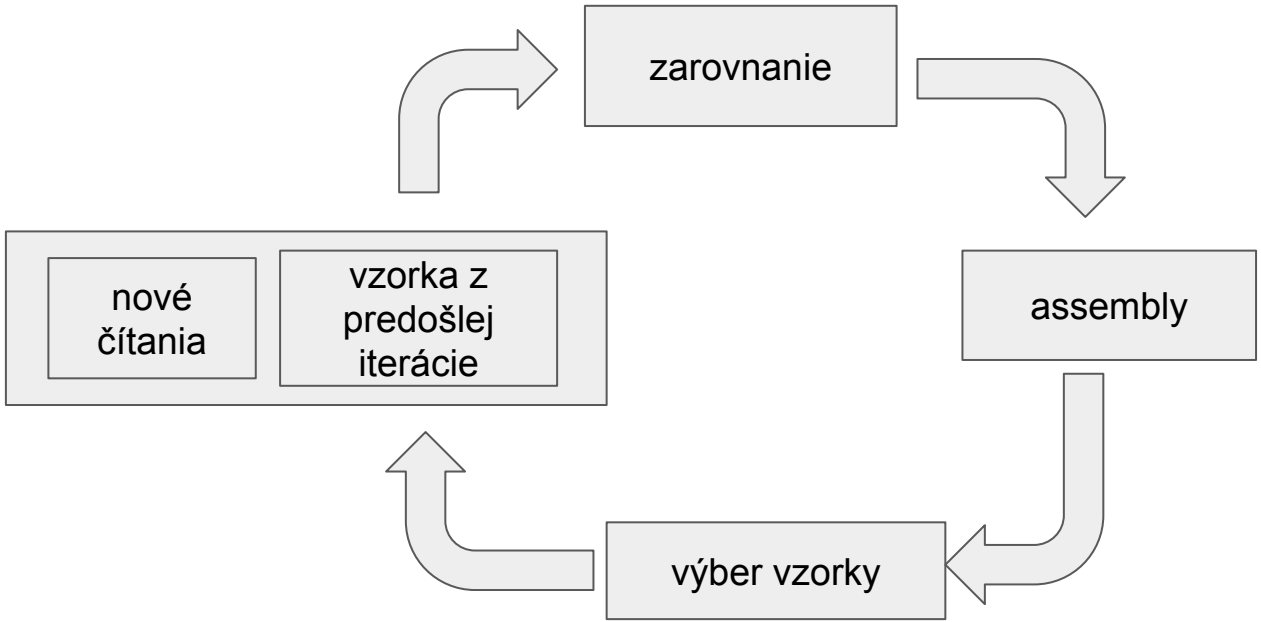
Overlap-layout-**consensus**

- **consensus:** presnejšie určenie báz na základe nejakého "*väčšinového pravidla*"



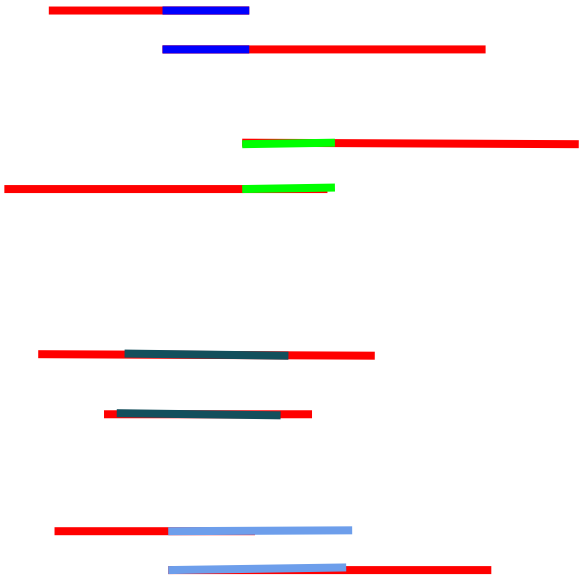
https://bio.libretexts.org/Bookshelves/Computational_Biology/

Idea

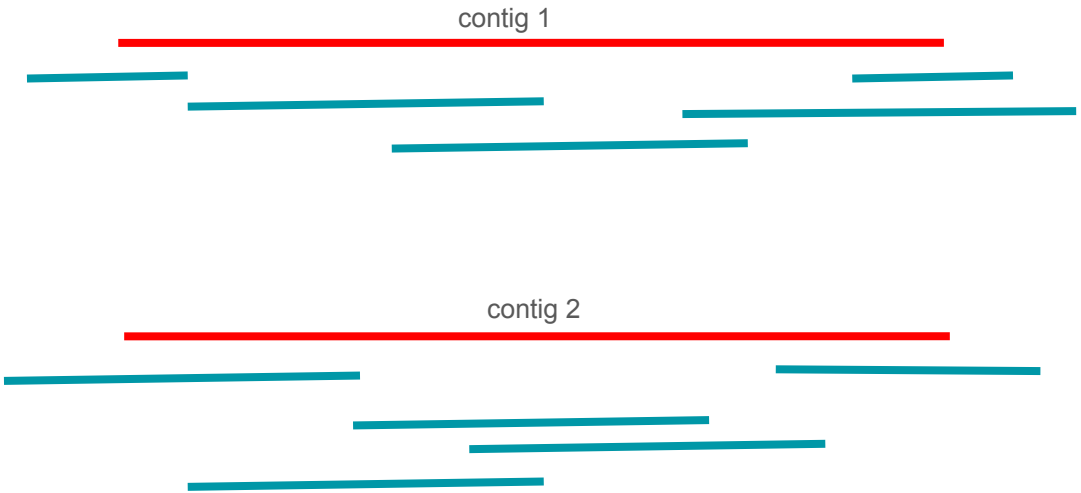

zarovnanie
assembly
výber vzorky
nové čítania
vzorka z predošlej iterácie

Zarovnanie (alignment) [Minimap]

čítania k čítaniam



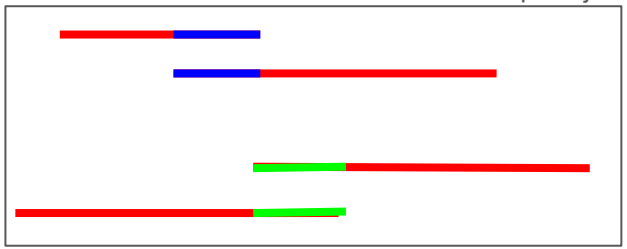
čítania k assembly



Zarovnanie (alignment) [Minimap]

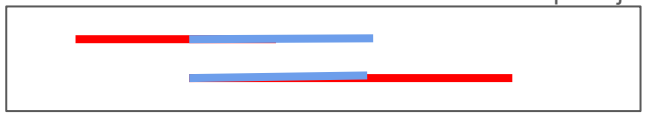
čítania k čítaniam

Miniasm použije



Miniasm zahodí

Miniasm použije

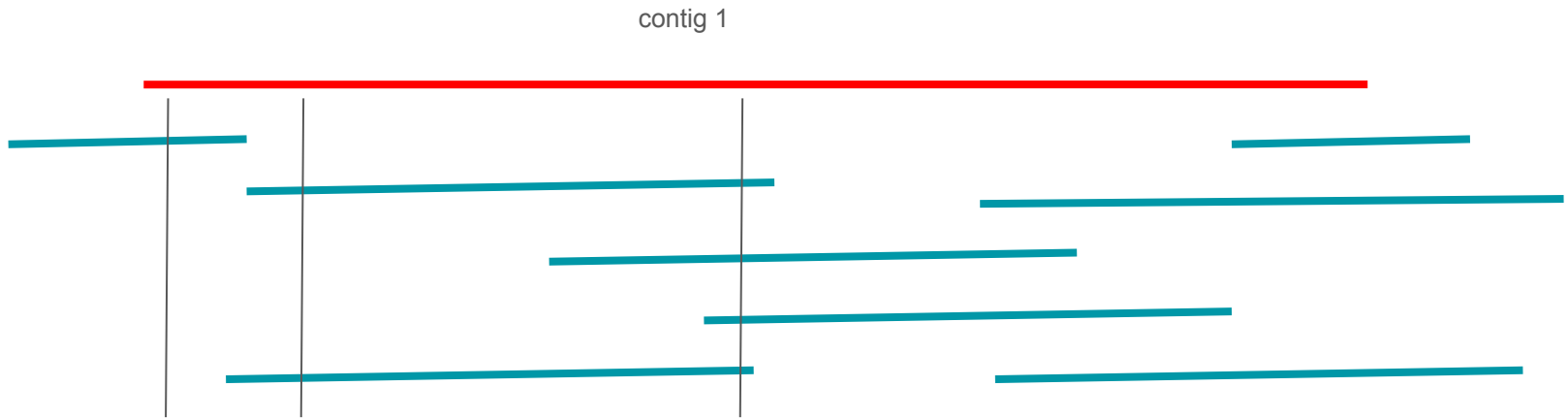


čítania k assembly

čítania, ktoré si chceme nechať (nová informácia)

zahadzujeme

Pokrytie (coverage)

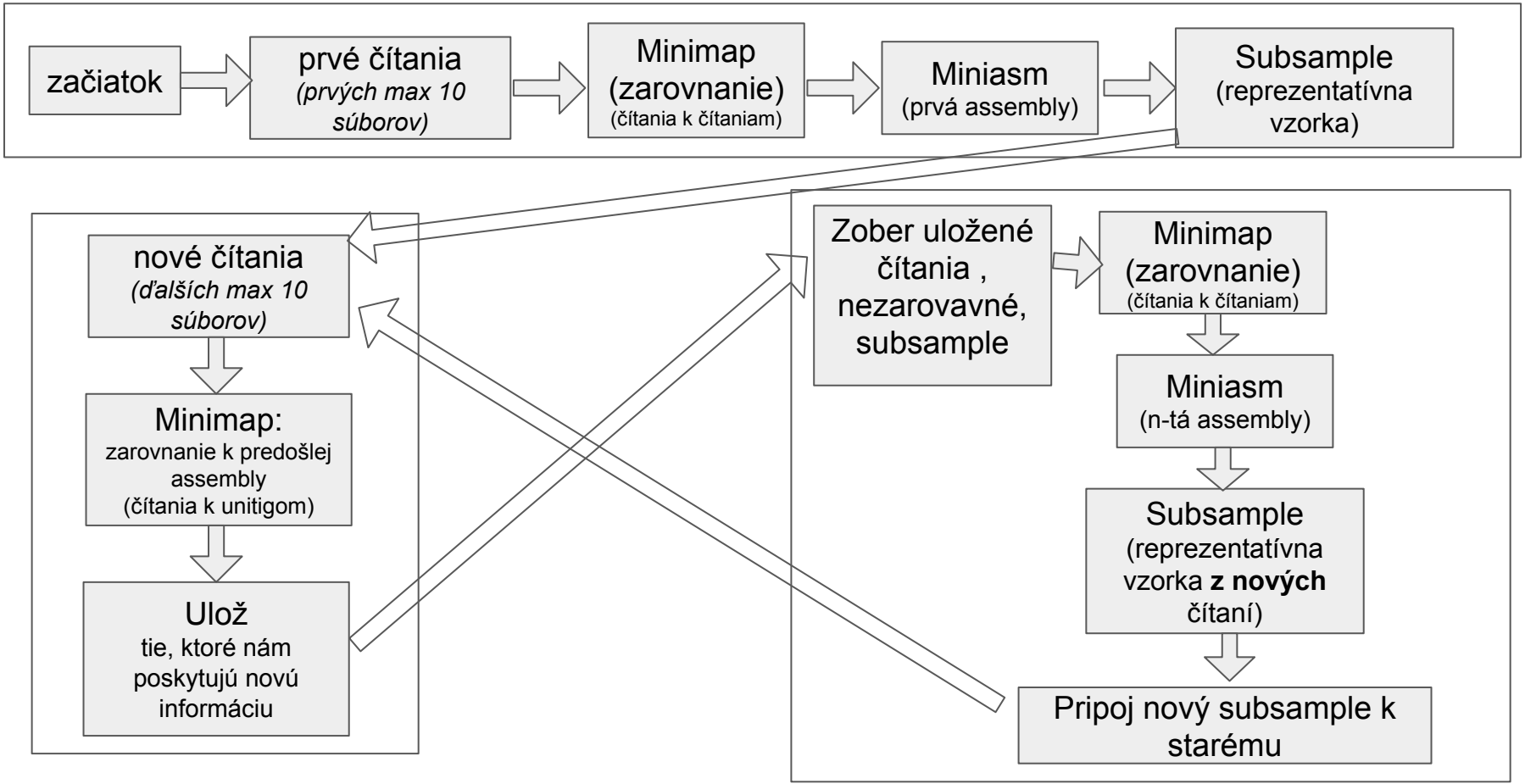


pozície p_ j=1 p_k =2 p_m=4

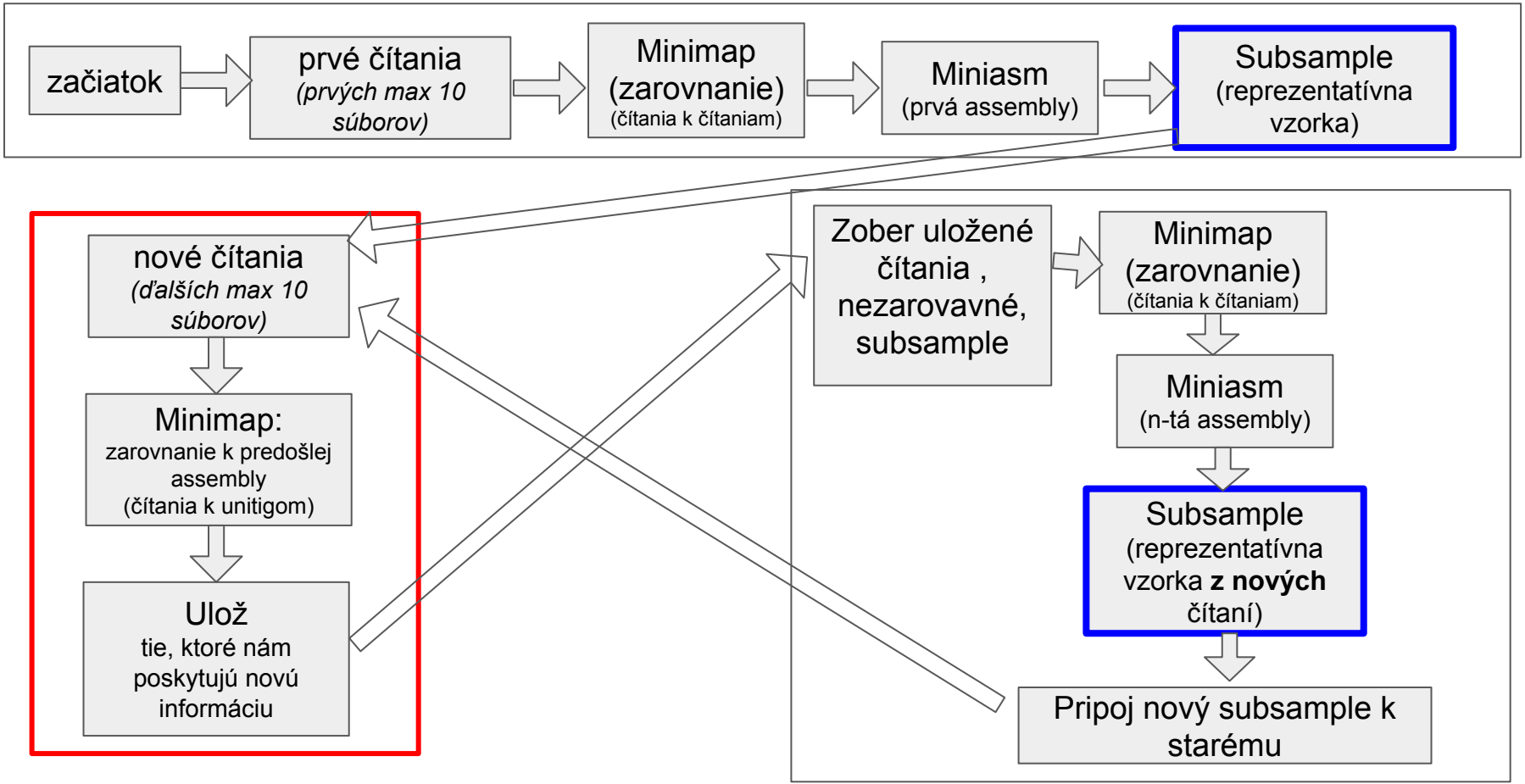
priemerné pokrytie contigu $\frac{\sum p_i}{\text{dĺžka}}$

reprezentatívna vzorka

Dynamický prístup s využitím minimap & miniasm



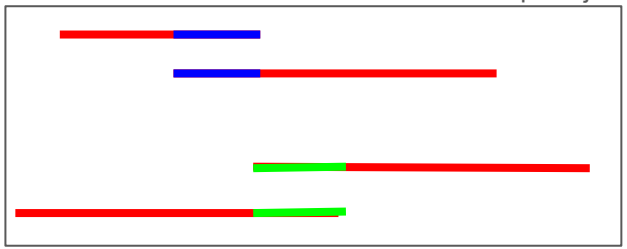
Dynamický prístup s využitím minimap & miniasm



Iný prístup: Zarovnanie (alignment) [Minimap]

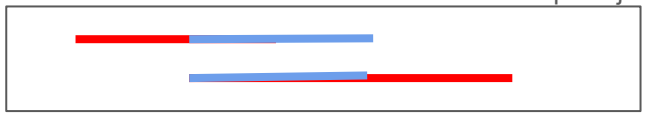
čítania k čítaniam

Miniasm použije



Miniasm zahodí

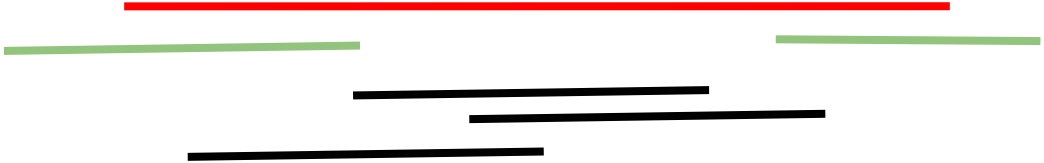
Miniasm použije



čítania k assembly

contig 1

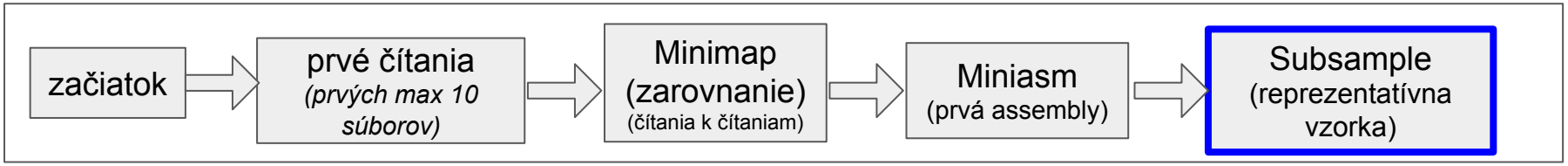
contig 2

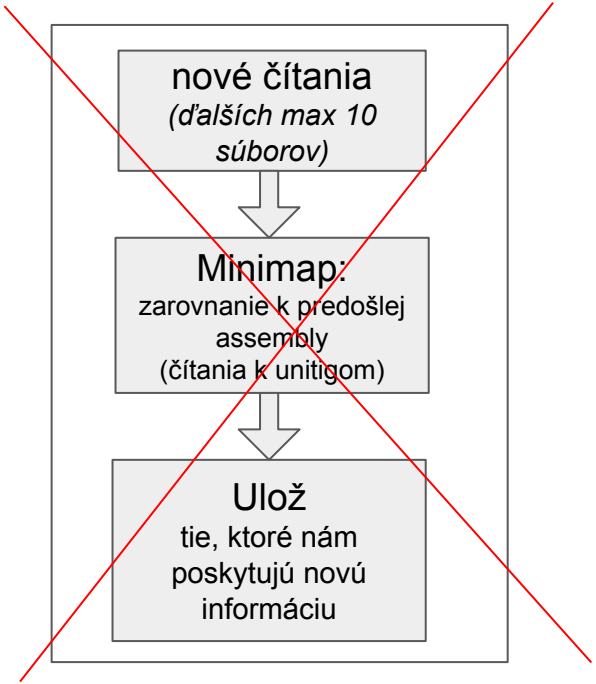


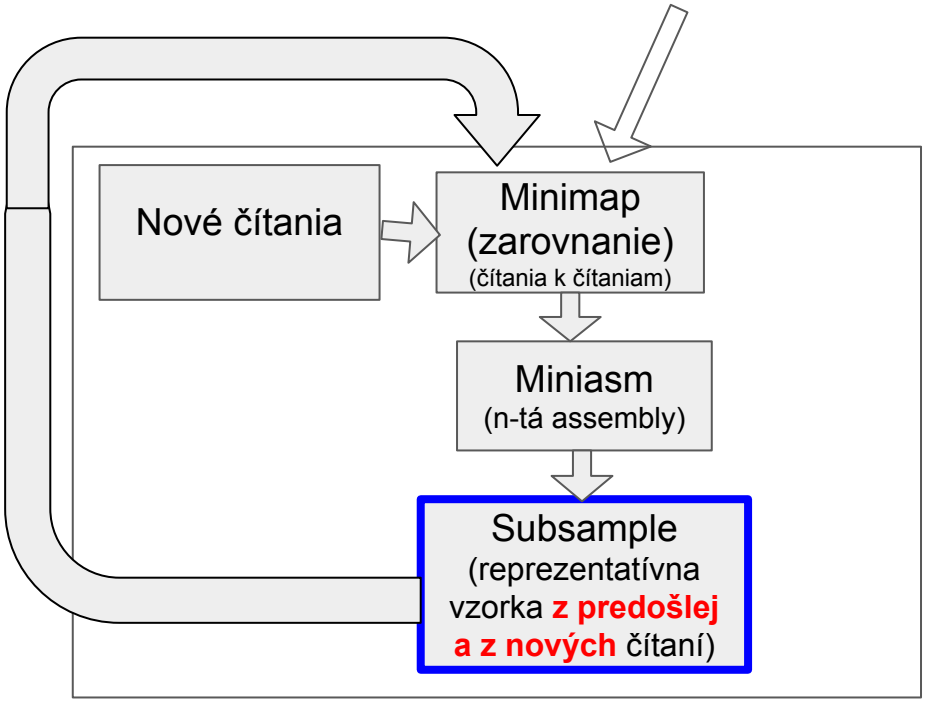
čítania, ktoré si chceme nechať (nová informácia)

nie je dobrý nápad zahodiť..

Dynamický prístup s využitím minimap & miniasm



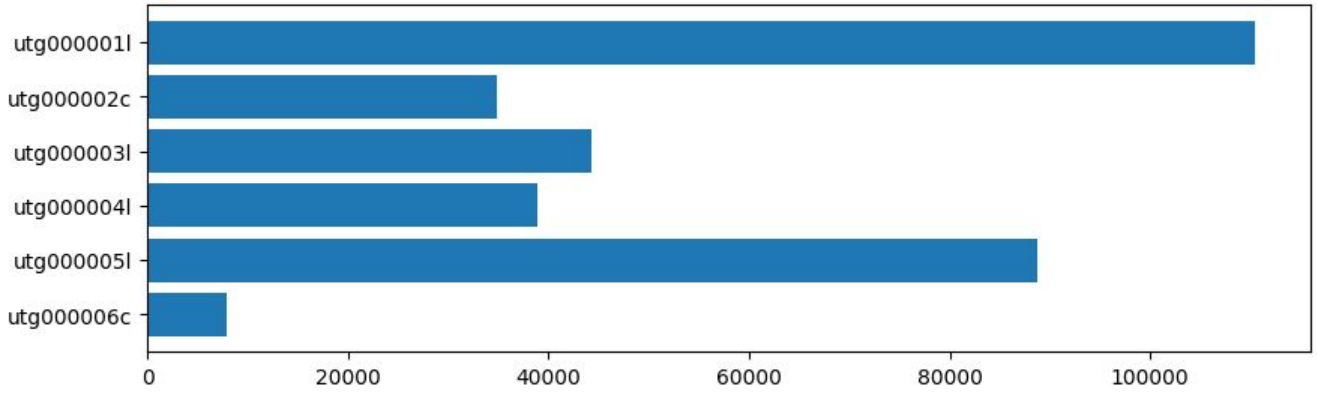




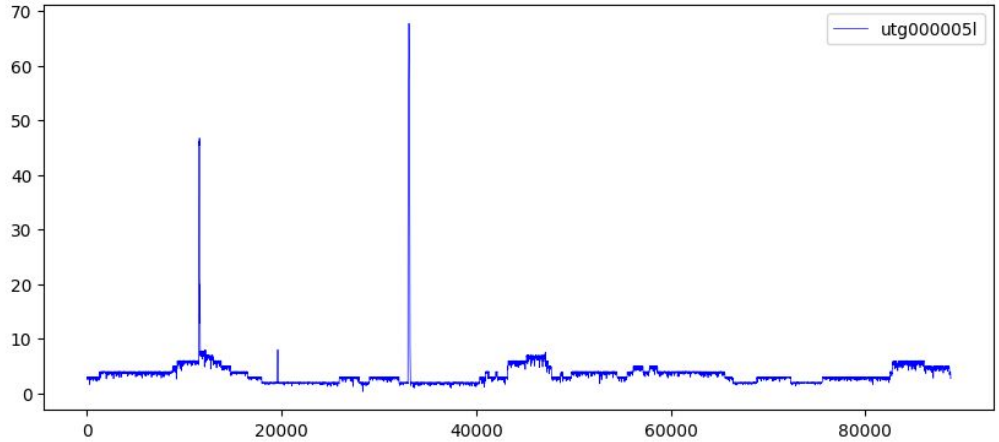
Výsledky

- implementovaná pipeline
- jednoduchý pravdepodobnostný prístup na výber reprezentatívnej vzorky, hľadáme lepší, porovnávame prístupy
- vizualizácie
- pipeline je časovo pozadu oproti (bežnej) rýchlosti pribúdania dát
    - nemáme odpoveď (úplne) v reálnom čase

grafy pre každú iteráciu



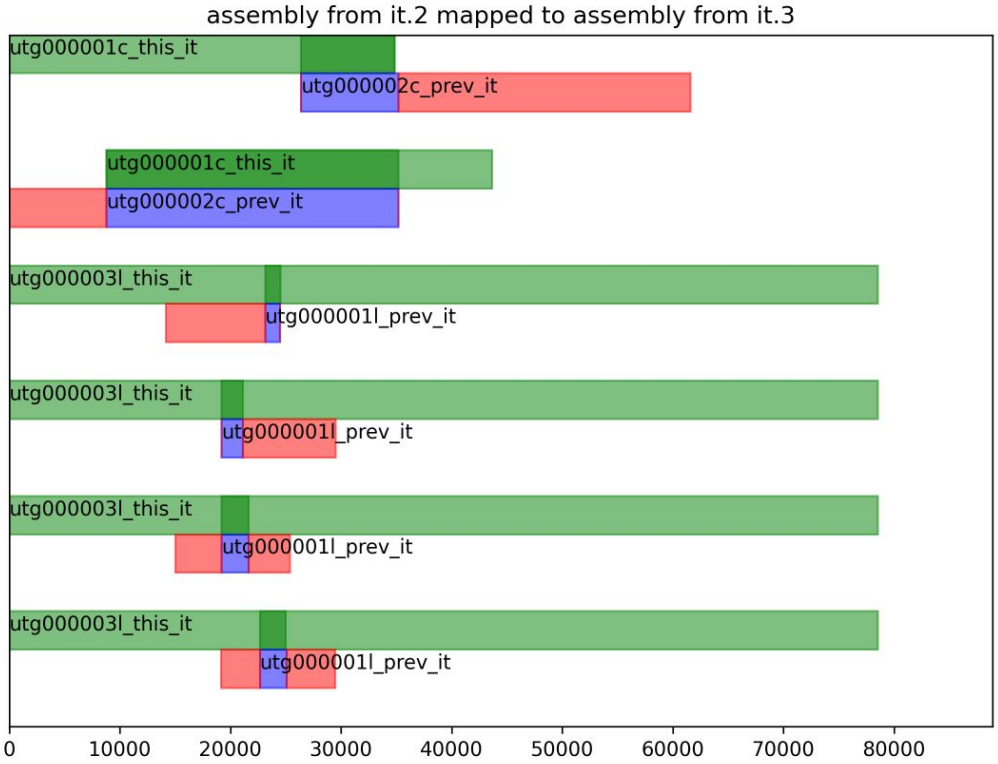
dĺžky contigov



pokrytie (pre každý contig)

grafy pre každú iteráciu

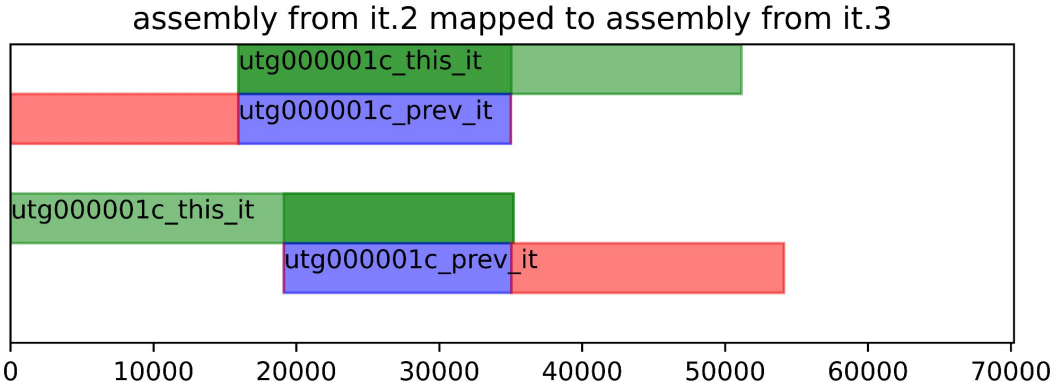
zarovnanie assembly z predošlej iterácie k aktuálnej assembly

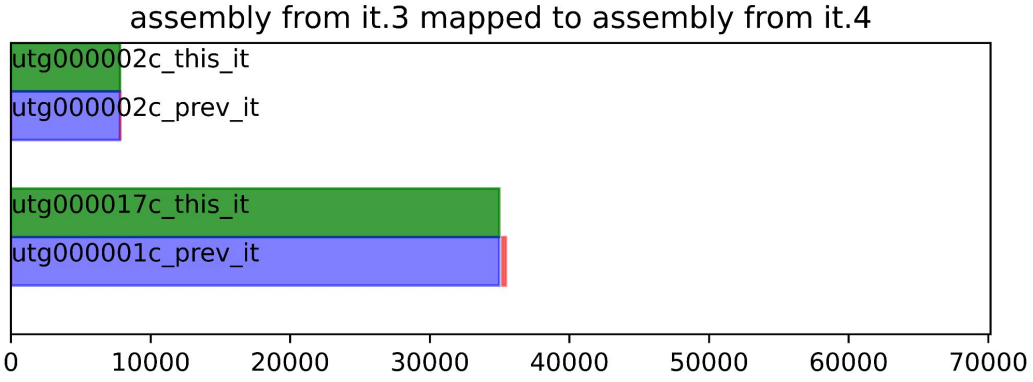


Ďakujem za pozornosť

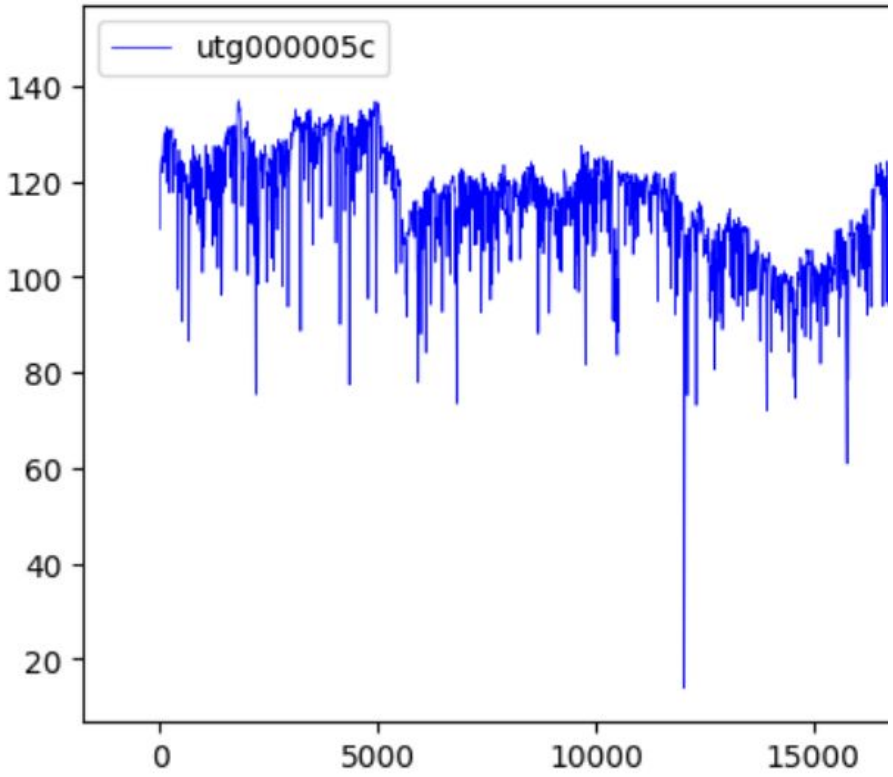
(nepoužité/doplnkové slajdy)

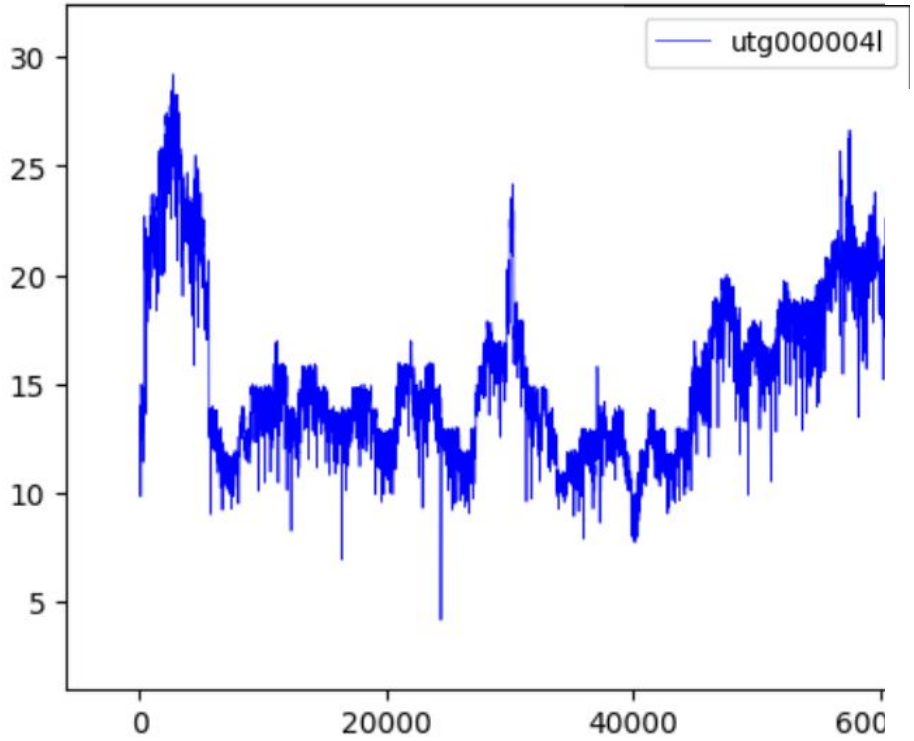
zarovnanie assembly z predošlej iterácie k aktuálnej assembly





rôzne pokrytie





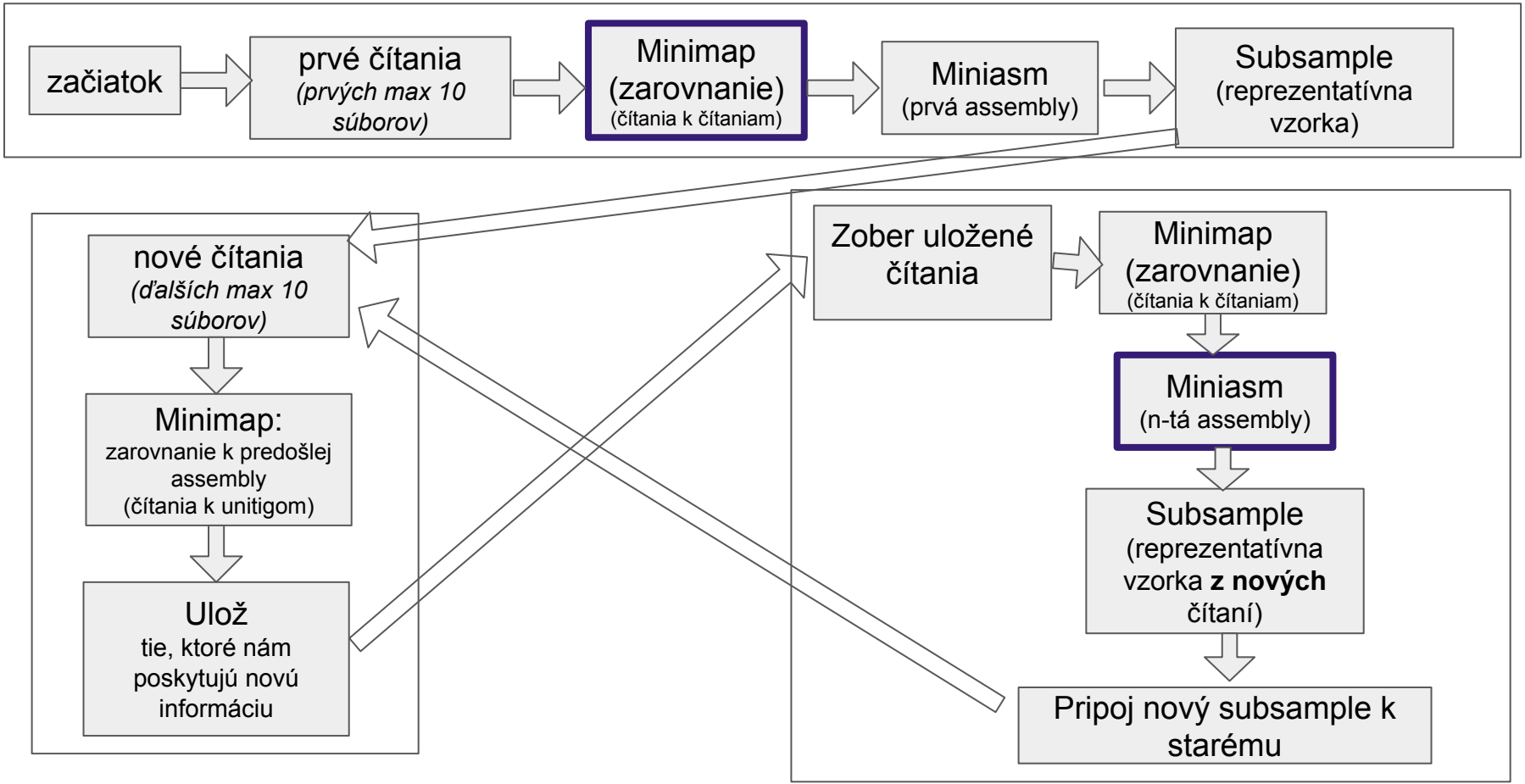
Saprochaete ingens

nepatogénna kvasinka

21,2 Mb

5 chrmozómov

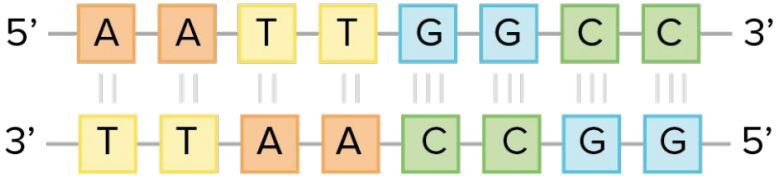
Dynamický prístup s využitím minimap & miniasm



Zarovnanie (čítania k čítaniam)

- čítania k assembly - pomerne rýchle
- čítania k čítaniam - môže trvať *dlho*
  - porovnáva sa každé čítanie s každým pomocou heuristiky ktorá hľadá zhody + dynamické programovanie
  - ku koncu behu môže trvať dlho (pri genóme kvasinky 30+ minút)
- možno sa dá využiť informácia z predošlého behu - už zarovnaná podmnožina čítaní
  - (zatiaľ neimplementované)

Assembly graph [Miniasm]



- $G = (V,E,\ell)$
- V …. množina DNA sekvencií
- E …. množina prekryvov (overlaps) medzi sekvenciami z V
  - bez násobných hrán
- ℓ: $E \rightarrow R^+$
- Watson-Crick complete $(i)\ \forall v \in V,\ \bar{v} \in V \quad (ii)\ \forall v \rightarrow w \in E,\ \overline{w} \rightarrow \bar{v} \in E$

- Containment-free - žiadna sekvencia nie je obsiahnutá v inej

overlap

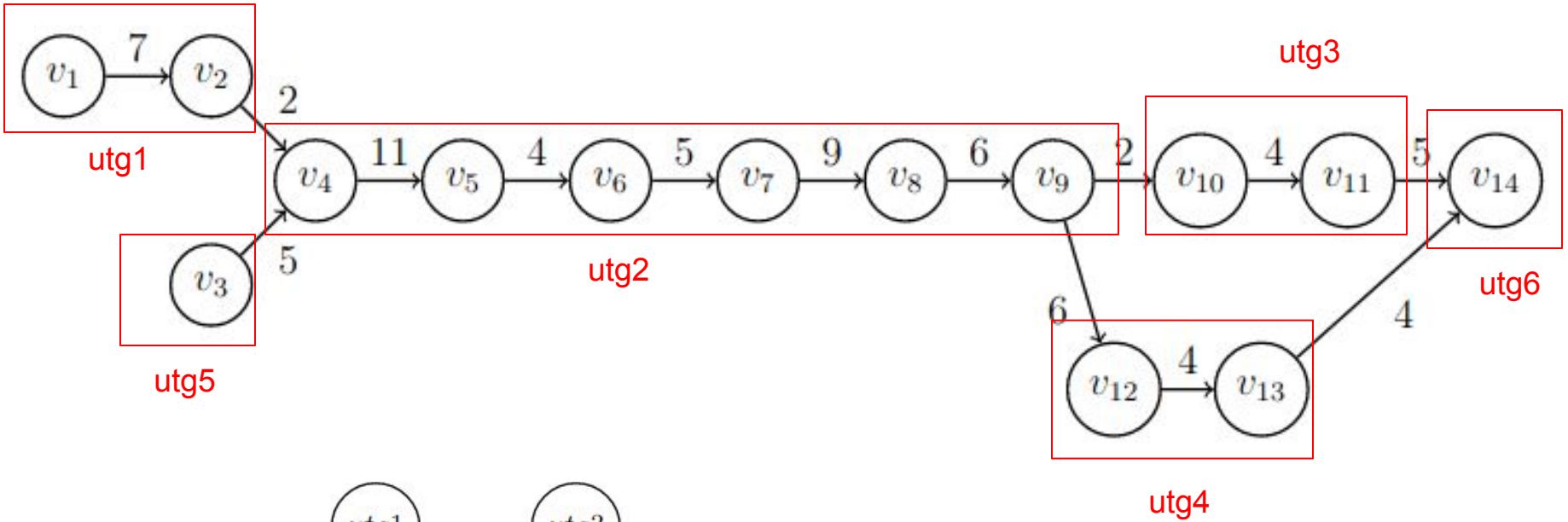
contained

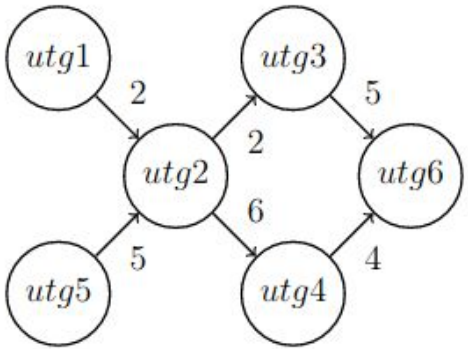
- (special case:) De Bruijn - všetky možné substringy z daných sekvencií dĺžky k

Hľadanie unitigov [Miniasm]

- assembly graph na základe zarovnaní
- tranzitívna redukcia,
- detekcia a odstránenie (malých) bublín (SNPy)
- unitig = cesta, v grafe, ktorú je možné “spojiť” bez toho aby sme stratili nejakú informáciu, resp. sekvencia ktorú vieme “prečítať” z takejto cesty

Definition (unitig): Let $G = (V, E, l)$ be a transitively reduced assembly graph. Then a unitig is a path $v_1 \rightarrow v_2 \rightarrow ... \rightarrow v_k$ such that $deg^+(v_i) = deg^- v_{i+1}$ and at least one of the following conditions is satisfied: (i) $v_1 = v_k$ or (ii) $deg^-(v_1) \neq 1$ and $deg^+(v_k) \neq 1$.


v_1
7
v_2
2
utg1
v_4
11
v_5
4
v_6
5
v_7
9
v_8
6
v_9
2
utg3
v_{10}
4
v_{11}
5
v_{14}
utg6
v_3
5
utg5
utg2
6
v_{12}
4
v_{13}
4
utg4


$utg1$
2
$utg3$
5
2
$utg2$
$utg6$
6
5
4
$utg5$
$utg4$

čiastkové zostavenia genómu (assembly) zarovnané k referenčnému genómu

výsledný subsample je ~25% všetkých čítaní

Reprezentatívna vzorka

- kontig dĺžky **c**, chceme pokrytie **t**, máme **r** čítaní, l_i je dĺžka zarovnaní i-teho čítania
  - chceme aby stredná hodnota súčtu dĺžok zarovnaných úsekov z čítaní bola $c \cdot t$, t.j.:

$$\sum_i p_i \cdot l_i = c \cdot t$$

- pravdepodobnosť výberu i-teho čítania $p_i = \frac{c \cdot t}{r \cdot l_i}$
  - zvolená tak, aby čítanie ktoré má dlhšie zarovnanie malo vyššiu pravdepodobnosť výberu

$$\sum_i p_i \cdot l_i = \sum_i \frac{c \cdot t}{r \cdot l_i} \cdot l_i = \frac{c \cdot t}{r} \sum_i \frac{l_i}{l_i} = \frac{c \cdot t}{r} \cdot r = c \cdot t$$

- problémy:
  - predpoklad rovnomernosti pokrytia (nerovnomerné pokrytie -> rozbitie contigu)
  - čítanie sa môže zarovnať na viacero miest vo viacerých contigoch