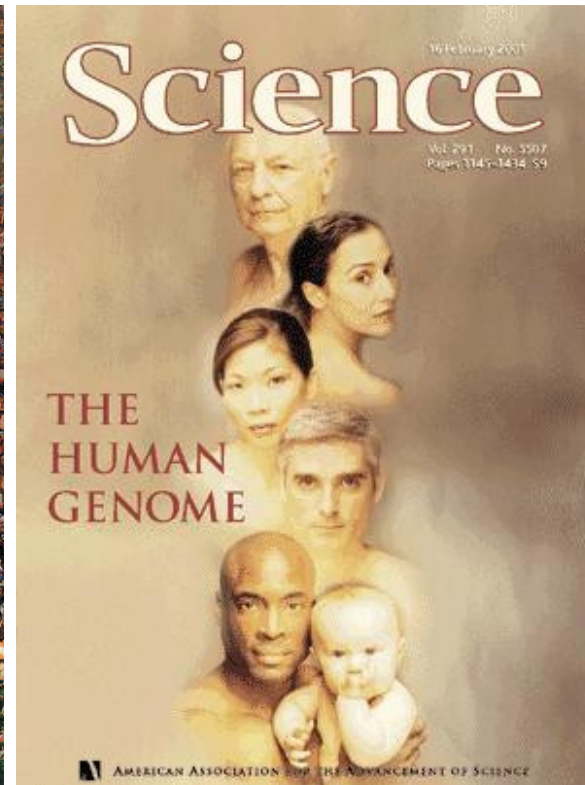


Sekvenovanie a zostavovanie genómov

(časť 2 - dlhé čítania)

Broňa Brejová

2.10.2025



Oznamy

- Nezabudnite na pravidelný kvíz v Moodli.
- Dnes kratšia prednáška, biológovia budú mať dlhšie cvičenia v F1 109.

Prehľad sekvenovacích technológií

Technológia	Dĺžka čítania	Chybovosť	Za deň	Cena za GB
1. generácia				
Sanger	do 1000 bp	$< 1\%$	3 MB	\$4 mil.
2. (next) generácia (cca od 2004)				
Illumina	300bp	$< 0.1\%$	2 TB	\$3
3. generácia (cca od 2018)				
PacBio HiFi	cca 15 Kbp	$< 1\%$	360 GB	\$15
Oxford Nanopore	5-100+kbp	$< 5\%$	50 GB	\$10

Na minulej prednáške

- Genóm je potrebné zostaviť zo sekvenačných čítaní
- Zostavovanie genómov pomocou de Bruijnových grafov
- Nie je vhodné pre najnovšie technológie s dlhými a chybovými čítaniami
 - Rozklad na k -mery zahadzuje príliš veľa informácie
(dĺžka čítania 10000+, k obvykle medzi 30 a 70)
 - Chybovosť okolo 5% robí de Bruijnov graf neprehľadným
(pre $k = 31$, **každý** k -mer v priemere 1-2 chyby)

Prístup Overlap–Layout–Consensus

- **Overlap:** Nájdí prekryvy medzi čítaniami a zostav tzv. **graf prekryvov**
- **Layout:** Zjednoduš graf prekryvov a nájdí v ňom cesty, ktoré budú zodpovedať **kontigom**
- **Consensus:** Ku každému kontigu zostav sekvenciu, ktorá je konsenzom sekvencií čítaní, ktoré kontig tvoria (opravovanie lokálnych chýb)

Overlap: hľadanie prekryvov

CATCTCTAGGCCAGC

| | | | | | |

TAGGCCTGCTTCTTG

- špeciálny prípad zarovnávanie sekvencií (nasledujúca prednáška)
- prekryvy **budú obsahovať chyby**
(v našom prípade cca 1 chyba na 20 báz prekryvu)
- **čítaní je veľa:** $30\times$ pokrytie ľudského genómu
 \Rightarrow cca 9 mil. čítaní dĺžky 10000
nemôžeme porovnávať každé čítanie s každým
- praktický prístup:
 - rýchle predfiltrovanie **vhodných kandidátskych párov čítaní**
(napríklad musia obsahovať dosť dlhý spoločný k -mer)
 - pomalšie zarovnávanie len pre kandidátske páry

Zostavenie grafu prekryvov

- Výsledok predchádzajúcej fázy:
CATCTCTAGGCCAGC / TAGGCCTGCTTCTTG, prekryv 9 báz
...
- Zostavíme **graf prekryvov**:
vrcholy: čítania ohodnotené hrany: prekryvy s dĺžkami

Príklad:

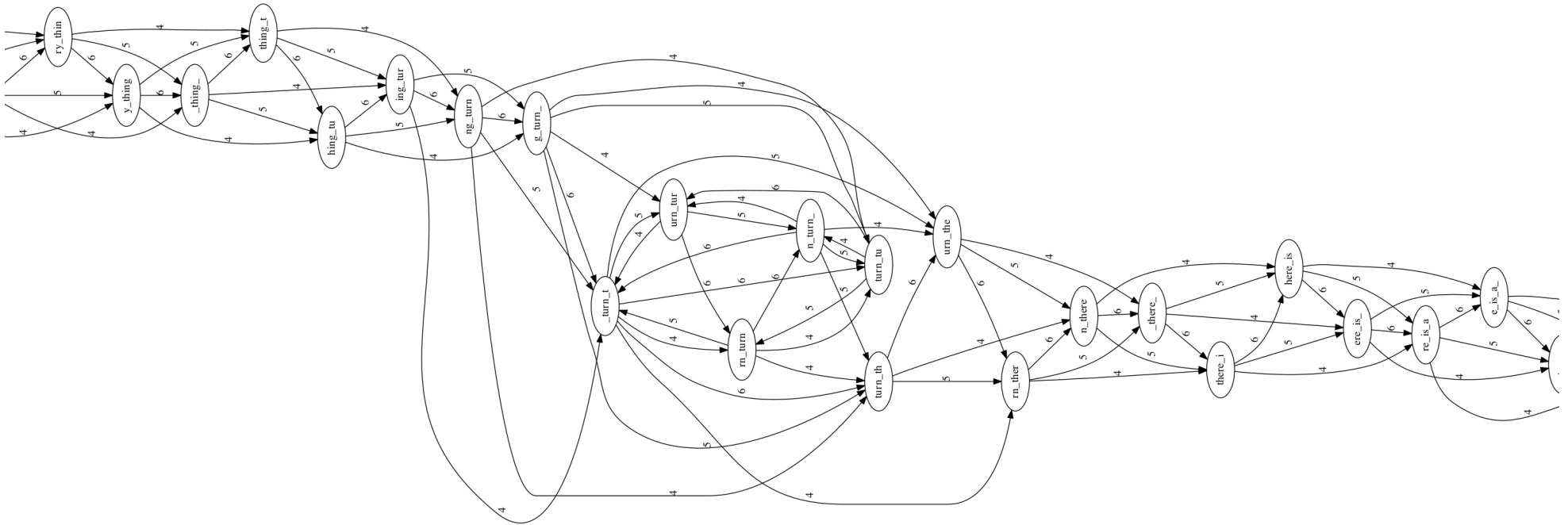
`to_every_thing_turn_turn_turn_there_is_a_season`

čítania dĺžky 7 písmen, minimálny prekryv 4

Príklad:

to_everything_turn_turn_turn_there_is_a_season

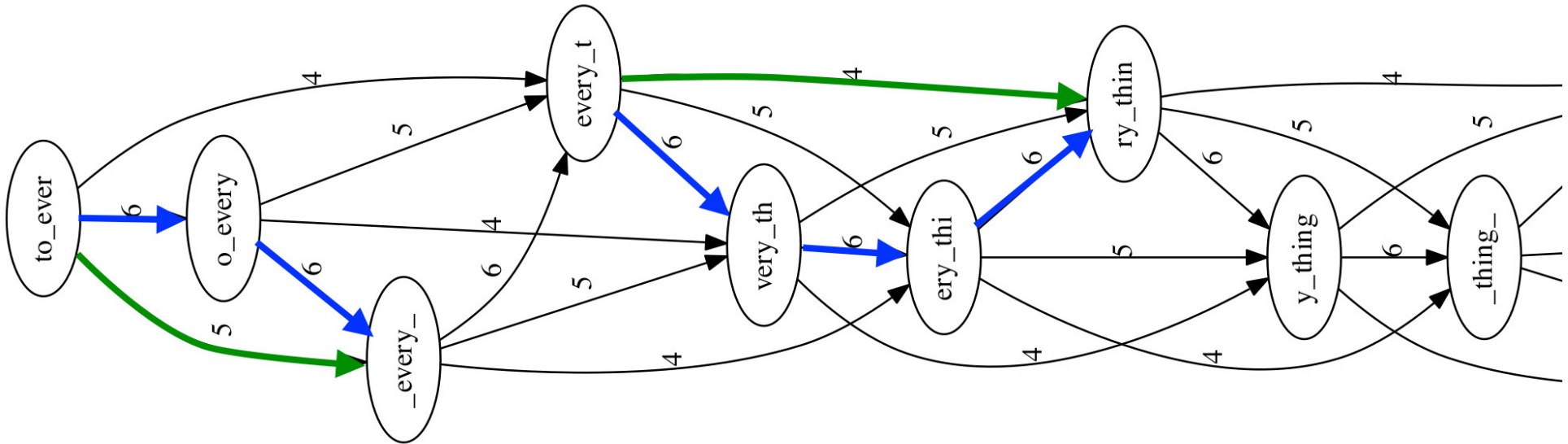
čítania dĺžky 7, minimálny prekryv 4



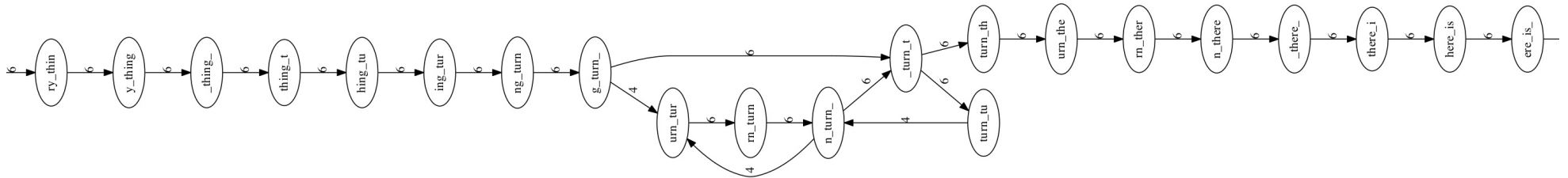
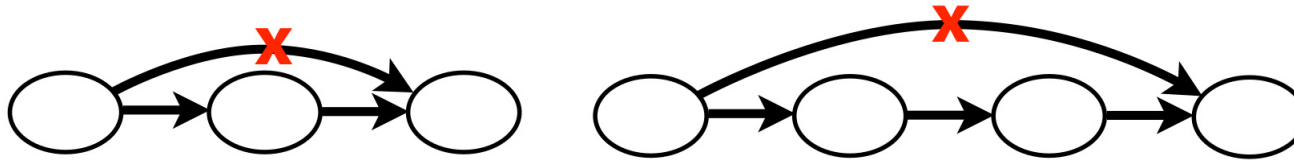
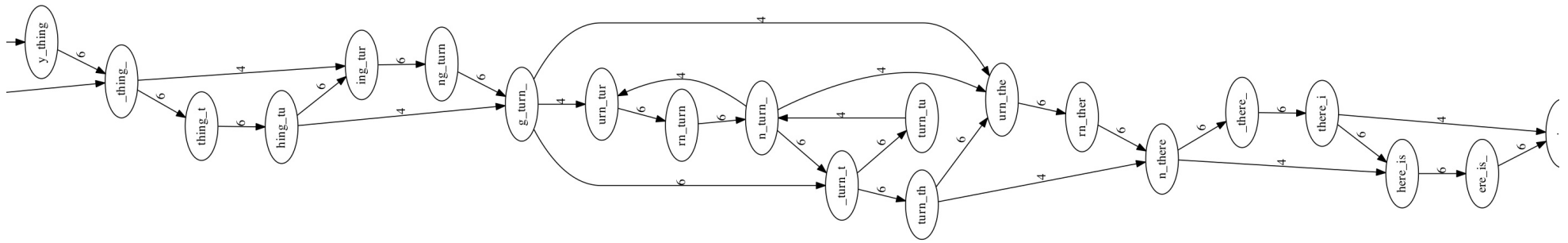
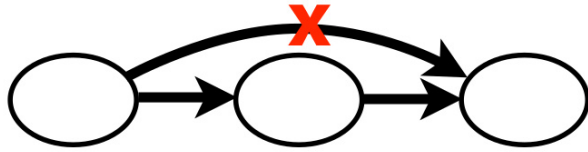
Príklad a obrázky Ben Langmead

Layout: Tranzitívne hrany

- Niektoré hrany sú nadbytočné, lebo hovoria to isté ako cesty z iných hrán



Layout: Odstránenie tranzitívnych hrán

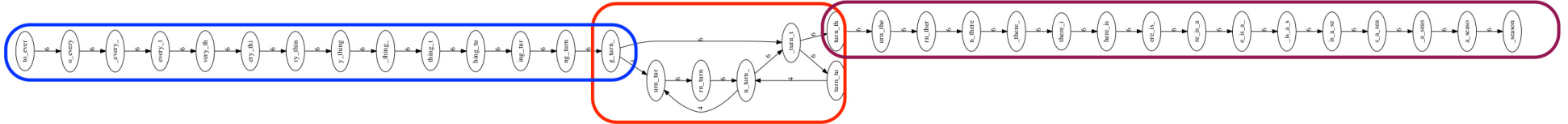


Layout: Rozdelenie na kontigy

Pôvodná sekvencia:

to_every_turn_turn_turn_there_is_a_season

Nerozvetvujúce sa cesty reprezentujú kontigy



Výsledok:

Contig 1

to_every_turn_

Contig 2

turn_there_is_a_season

┌──────────┐
Unresolvable repeat

Consensus: Získanie finálnej sekvencie

TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

↓ ↓ ↓ ↓ ↓
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

Take reads that make
up a contig and line
them up

Take *consensus*, i.e.
majority vote

Ako sa líši/podobá de Bruijnov graf od grafu prekryvov?

de Bruijnov graf

- fixná dĺžka prekryvov
- zahadzujeme informáciu o kontinuite presahujúcej k znakov
- **cesty reprezentujú genóm**
- chyby \Rightarrow bubliny a výbežky
- riešia sa v predspracovaní
- kontigy pokrývajú takmer všetky hrany

Graf prekryvov

- variabilná dĺžka prekryvov
- maximálne využitie informácie o prekryvoch
- **cesty reprezentujú genóm**
- chyby sú zväčša “schované”
- riešia sa dodatočne (consensus)
- treba odstraňovať tranzitívne hrany

Príklad: Skladanie genómu *Magnusiomyces capitatus*

(dĺžka genómu 19.6 Mbp, 4 chromozómy + mtDNA)

Technológia	Pokrytie	# kontigov	najväčší	N50
Illumina / Spades	250x	1102	172.6 Kbp	62.0 Kbp
PacBio / Canu	37x	17	4.7 Mbp	1.7 Mbp
PacBio + nanopore	65x	11	4.4 Mbp	2.0 Mbp

Zhrnutie

- Dlhé čítania nám umožňujú poskladať genómy do podstatne menej fragmentovanej podoby ako krátke čítania
- Na hľadanie prekryvov medzi čítaniami sú potrebné rýchle algoritmy (niektoré si ukážeme o dve prednášky)
- Grafy prekryvov a de Bruijnové grafy sa podobajú, existujú snahy o zjednotenie týchto dvoch konceptov

História sekvenovania genómov

1976	MS2 (RNA vírus) 40 kB
1988	projekt sekvenovania ľudského genómu (15 rokov)
1995	baktéria <i>H. influenzae</i> 2 MB, shotgun (TIGR)
1996	<i>S. cerevisiae</i> 10 MB, BAC-by-BAC (Belgicko, Británia)
1998	<i>C. elegans</i> 100 MB, BAC-by-BAC (Wellcome Trust)
1998	Celera: ľudský genóm do troch rokov!
2000	<i>D. melanogaster</i> 180 MB, shotgun (Celera, Berkeley)
2001	2x ľudský genóm 3 GB (NIH, Celera)
po 2001	Myš, potkan, kura, šimpanz, pes, . . .
2007	Watsonov a Venterov genóm (454)
2012	1000 ľudských genómov
2021	3,5 milióna genómov SARS-CoV-2
2021	UK Biobank 200,000 ľudských genómov + veľa ďalších dát
2022	Naozaj dokončený ľudský genóm (telomere to telomere)
2024	All of US 246,000 ľudských genómov + zdravotné záznamy

Použitie NGS: Populačná genetika

- Sekvenujeme čítania z genómu určitého človeka
- Ako sa môj vlastný genóm líši od referenčného ľudského genómu?
- Ako jednotlivé genetické rozdiely ovplyvňujú fenotyp?
- Personalizovaná medicína
- Populačná štruktúra, história ľudstva
- Etické otázky

Problémy:

- Mapovanie čítaní na referenčný genóm
- Identifikácia rozdielov (malých a väčších)

Použitie NGS: Environmentálne sekvenovanie – Metagenomika

- Mikrobióm: komunita mikroorganizmov žijúcich v určitom prostredí
- V medicíne sú dôležité mikrobiómy ľudskeho čreva, ústnej dutiny atď
- Zaujímavé sú ale aj mikrobiómy z iných prostredí
- Klasický postup: izolujeme jednotlivé druhy, ktoré pestujeme osobitne v laboratóriu
- Metagenomika: sekvenujeme zmes čítaní z rôznych genómov bez izolácie
- Pribúdajú MAG: metagenome-assembled genome

Problémy:

- Porovnanie veľkého množstva čítaní s veľkou databázou známych genómov, kvantifikácia zastúpenia známych druhov
- Skladanie čítaní do kontigov

Použitie NGS: Hľadanie génov, väzobných miest,...

- RNA-seq: Sekvenovať môžeme aj RNA, dostávame gény v genóme
- ChIP-seq: vyfiltrujeme kúsky DNA, na ktoré je naviazaný určitý proteín, sekvenujeme, mapujeme na genóm
- Veľa ďalších technológií mapujúcich pomocou sekvenovania modifikácie DNA, stav chromatínu, 3D rozmiestnenie a pod. (viď predmet Genomika)

Problémy:

- Opäť mapovanie čítaní na referenčný genóm
- Identifikácia miest zostrihu
- Identifikácia väzobných miest podľa hĺbky pokrytia

