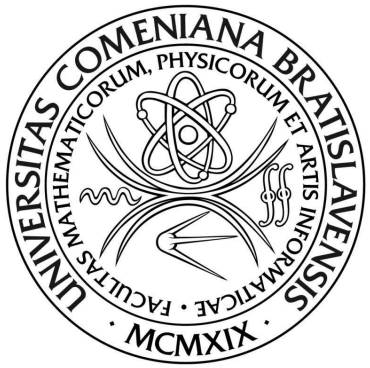
Metódy pre spracovanie sekvenačných dát z moderných sekvenátorov

Tomáš Vinař

Fakulta matematiky, fyziky a informatiky

Univerzita Komenského v Bratislave

http://compbio.fmph.uniba.sk/



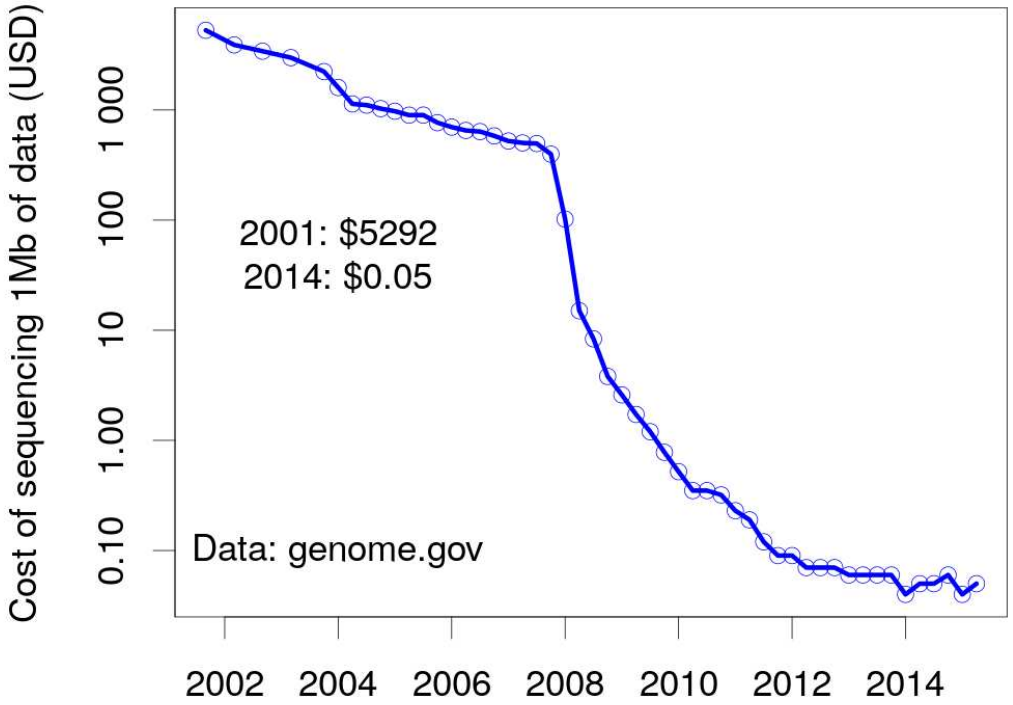
Rapídny vývoj sekvenovania DNA

ABI Sanger sequencing in 2001: 115kb/day

iba veľké sekvenovacie centrá, medzinárodné konzorciá

Illumina HiSeq 4000 in 2015: 107Gb/day

aj menšie laboratórium môže produkovať vlastné dáta



Čítania z moderných sekvenátorov

- Všetky súčasné technológie produkujú **čítania**: kratšie fragmenty DNA
- Čítania pochádzajú z náhodných a neznámych miest v genóme,
  pokrytie: priemerný počet čítaní prekrývajúce pozíciu v genóme

Technology	Read length	Errors
Sanger	up to 900bp	$< 2\%$
Illumina	150bp	$< 2\%$
454	400bp	$< 2\%$
PacBio	up to 14kbp	15%
Oxford Nanopore	up to 100kbp	30%

Rozmanitosť technológií: výzva pre vývoj nových metód

- **Hybridné dáta.**

  kombinácie rôznych vzdialeností párových čítaní

  kombinácie rôznych technológií

- **Fragmentovanosť dát.**

  veľmi ťažké zoskladať z krátky čítaní celé genómy.

  repetitívne regióny, veľké segmentálne duplikácie, . . .

- **Chybovosť dát.**

  štandardné algoritmy vyvinuté pre 2. generáciu sekvenátorov

  nefungujú pre 3. generáciu, ktorá má oveľa väčšiu chybovosť

Príklad 1: Zostavovanie genómov z hybridných dát (GAML)

Zostavovanie z hybridných dát:

- ALLPATHS-LG: fragment + short jump Illumina
- Cerulean, PacbioToCA: fragment Illumina + PacBio

Náš cieľ: Transparentná kombinácia hybridných dát

Použijeme pravdepodobnostné modely / vierohodnosť

- Oprava zoskladania: REAPR (Hunt, Martin, et al. 2013)
- Vyhodnodcovanie/porovnávanie pomocou vierohodnosti:
  ALE (Clark et al. 2013), CGAL (Rahman et al. 2013),
  LAP (Ghodsi et al. 2013)

$\Rightarrow$ vierohodnosť je dobrou mierou pre porovnávanie zostavení

Vierohodnosť zoskladania - LAP model

- Dané: zostavenie genómu A, množina čítaní R
- Nezávislé čítania: $\Pr[R|A] = \prod_{r \in R} \Pr[r|A]$
- Log average vierohodnosť: $L(R, A) = \sum_{r \in R} \Pr[r|A]/|R|$
- Vierohodnosť zarovnania čítania ku zostaveniu genómu:
  $\varepsilon^e(1-\varepsilon)^{\ell-e}/2L$
  ε: chybovosť, L: dĺžka zostavenia, e: počet chýb
- Vierohodnosť čítania: súčet vierohodností všetkých zarovnaní
- Ľahko adaptovateľné na párové čítania
  (použijeme aj distribúciu vzdialeností)
- Pokrýva vlastnosti všetkých súčasných sekvenovacích technológií

Náš prístup k zostavovaniu genómov

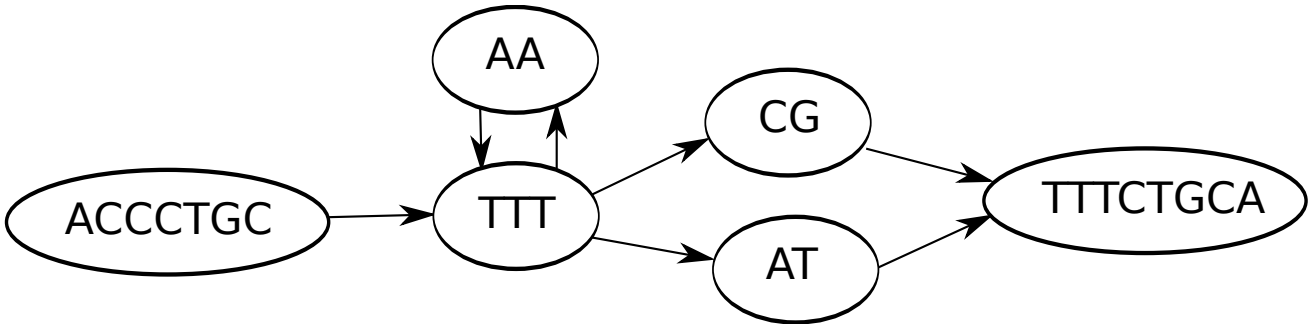
- Zostavovanie genómov $\Leftrightarrow$ hľadanie zostavenia s najvyššou vierohodnosťou (vzhľadom na dáta, ktoré máme k dispozícii)
- Prehľadávanie obrovského priestoru
- Predspracovanie pomocou Velvet-u $\Rightarrow$ graf reprezentujúci možné zostavenia
- Iteratívne vylepšovanie zostavenia pomocou simulovaného žíhania

Predspracovanie a formulácia úlohy

- Velvet s konzervatívnymi nastaveniami

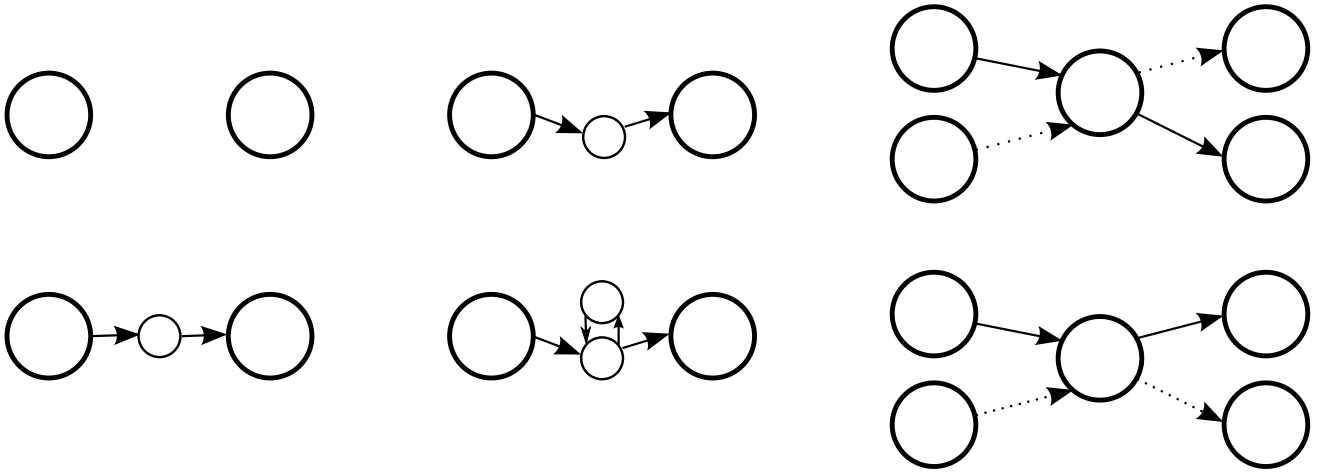
  (malé kontigy, ale bez chýb)

- Úloha: hľadáme **množinu sledov vo výslednom grafe** tak, aby vierohodnosť bola najväčšia možná



Simulované žíhanie

- Začnime s nejakou rozumnou množinou sledov
- Iteratívne navrhujeme jej modifikácie
- Ak dostaneme vyššiu vierohodnosť, akceptujeme modifikáciu
- Ak dostaneme nižšiu vierohodnosť, občas akceptujeme modifikácie (podľa žíhacieho rozvrhu)



Kľúč k úspechu: Rýchly výpočet vierohodnosti

- Vierohodnosť musíme prepočítať v každom kroku
- Rozdiely medzi zostaveniame v nasledujúcich krokoch sú veľmi malé
- Rozdelíme zostavenia na prekrývajúce okná
  v každom okne zarovnáme čítania zvlášť a zapamätáme si výsledok
  každý krok ovplyvní len malý počet okien
  musíme sa uistiť, aby sme žiadne čítanie nezapočítali viackrát
- Implementácia GAML: funguje dobre pre malé genómy (desiatky MB),
  treba použiť ďalšie triky pre veľké genómy

Experimental evaluation

GAML (genome assembly by maximum likelihood)

Compare to: GAGE (Salzberg et al. 2012), Cerulean (Deshpande et al. 2013)

ID	Technology	Insert len. (bp)	Read len. (bp)	Coverage	Error rate
Staphylococcus aureus (2.87Mbp)					
SA1	Illumina (fragment)	180bp	101bp	90	3%
SA2	Illumina (short jump)	3500bp	37bp	90	3%
Escherichia coli (4.64Mbp)					
EC1	Illumina (fragment)	300bp	151bp	400	0.75%
EC2	PacBio (long read)		4000bp	30	13%
EC3	Illumina (simulated long jump)	37,000bp	75bp	0.5	4%

Tests: SA1 + SA2, EC1 + EC2, EC1 + EC2 + EC3

Fragment and short jump Illumina libraries *S. aureus*

Assembler	Longest scaffold (kb)	N50 (kb)	Err.	Longest scaffold corr. (kb)	N50 corr. (kb)	LAP
GAML	1191	514	**0**	1191	514	−**23.45**
Allpaths-LG	1435	1092	**0**	**1435**	**1092**	−25.02
SOAPdenovo	518	332	**0**	518	332	−25.03
Velvet	958	762	17	532	126	−25.34
MSR-CA	**2412**	**2412**	3	1343	1022	−26.26
ABySS	125	34	1	125	28	−29.43
Cons. Velvet	95	31	**0**	95	31	−30.82
SGA	286	208	1	286	208	−31.80

Fragment Illumina library and long-read Pacbio library *E. coli*

Assembler	Longest scaffold (kb)	N50 (kb)	Err.	Longest scaffold corr. (kb)	N50 corr. (kb)	LAP
PacbioToCA	1533	**957**	**0**	1533	**957**	**−33.86**
GAML	1283	653	**0**	1283	653	−33.91
Cerulean	**1991**	694	**0**	**1991**	694	−34.18
AHA	477	213	5	477	194	−34.52
Cons. Velvet	80	21	**0**	80	21	−36.02

Fragment Illumina library, long-read Pacbio library and a long jump Illumina library *E. coli*

Assembler	Longest scaffold (kb)	N50 (kb)	Err.	Longest scaffold corr. (kb)	N50 corr. (kb)	LAP
GAML	**4662**	**4662**	**3**	**4661**	**4661**	**−60.38**
Celera	4635	4635	19	2085	2085	−61.47

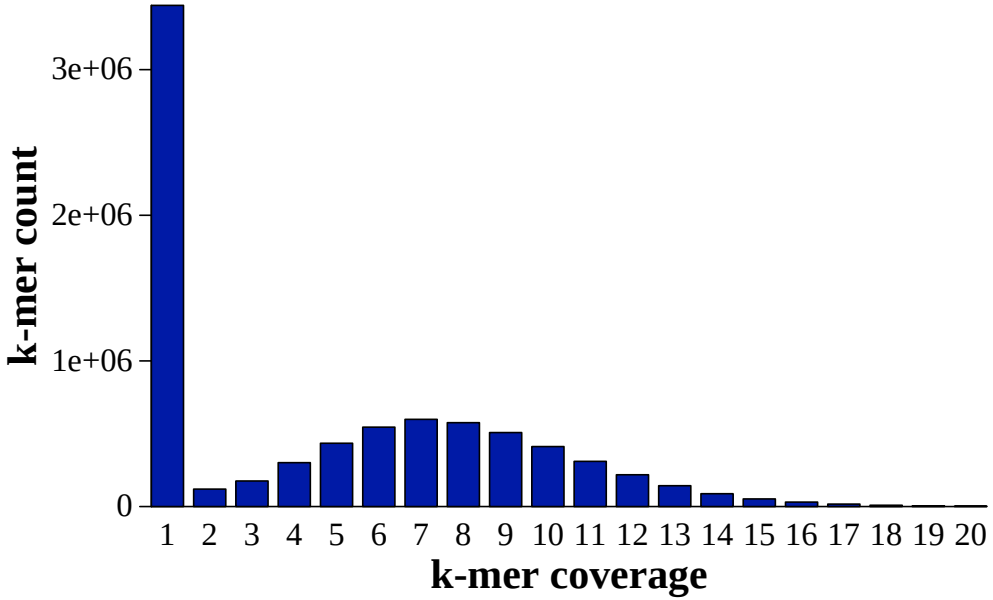
Príklad 2: Odhadovanie veľkosti genómov (CovEst)

- Koľko sekvenovacích behov potrebujeme?
- Obvyklý postup: poskladať genóm, potom odhad veľkosti
- Repetitívne sekvencie $\Rightarrow$ menší genóm
- Polymorfizmy $\Rightarrow$ väčší genóm
- Časti genómy nepokryté $\Rightarrow$ menší genóm
- ...

Prístup bez skladania genómov:

- štatistika priamo z čítaní (výskyt k-merov)
- odhad pokrytia na základe pravdepodobnostného modelu

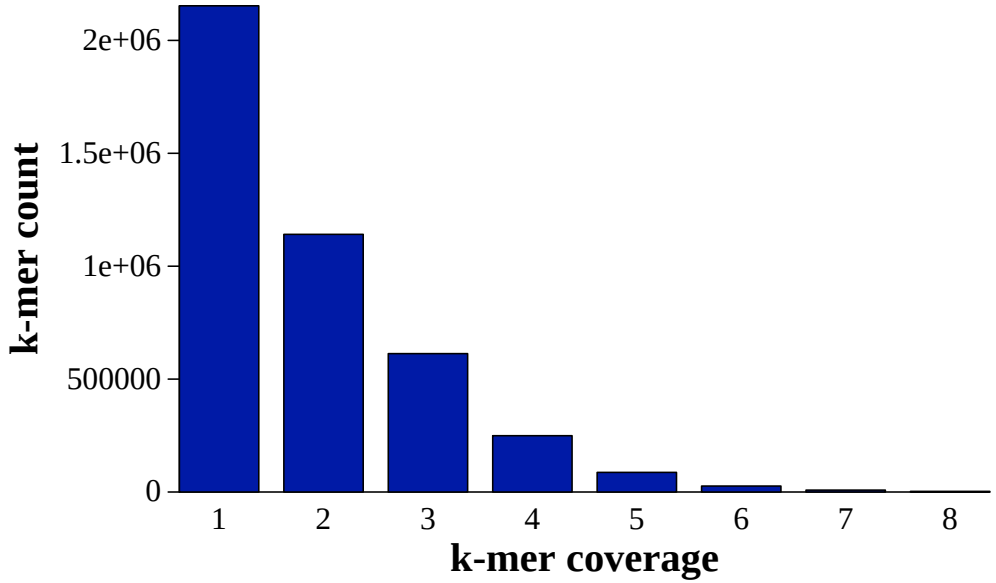
Spektrum k-merov pre Illumina čítania z *E. coli* ($k = 21$, pokrytie 10)



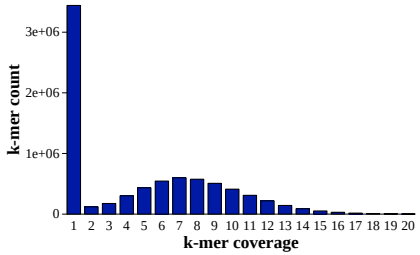
- Sekvenovacie chyby vytvárajú veľké množstvo jedinečných k-merov
- Jedna sekvenovacia chyba prekrýva do 21 k-merov
- Pokrytie genómu je 10, no mód je posunutý na 7
- Chybovosť a pokrytie na hraniciach čítaní

Spektrum k-merov pre Illumina čítania z *E. coli*

($k = 21$, pokrytie 2)



Modelovanie k-merového spektra



- Vstupom je množina čítaní, každé o dĺžke r
- Spektrum: $W = w_1, \ldots, w_m$
- Parametre modelu: θ (vrátane **pokrytia**)
- Nech p_j je pravdepodobnosť pozorovania j výskytov daného k-meru
- Logaritmus pravdepodobnosti pozorovania spektra W

  $\log L(W|\theta) = \log \left(\prod_{j=1}^{m} p_j^{w_j} \right) = \sum_{j=1}^{m} w_j \log p_j.$
- Hľadáme parametre θ maximalizujúce $L(W|\theta)$.

  Veľkosť genómu vypočítame z pokrytia

Jednoduhý model bez uvažovania chýb

- Parameter: pokrytie c
- Efektívne pokrytie k-meru $c_k = c(r - k + 1)/r$ (okraje čítania)
- Pokrytie daného k-meru:

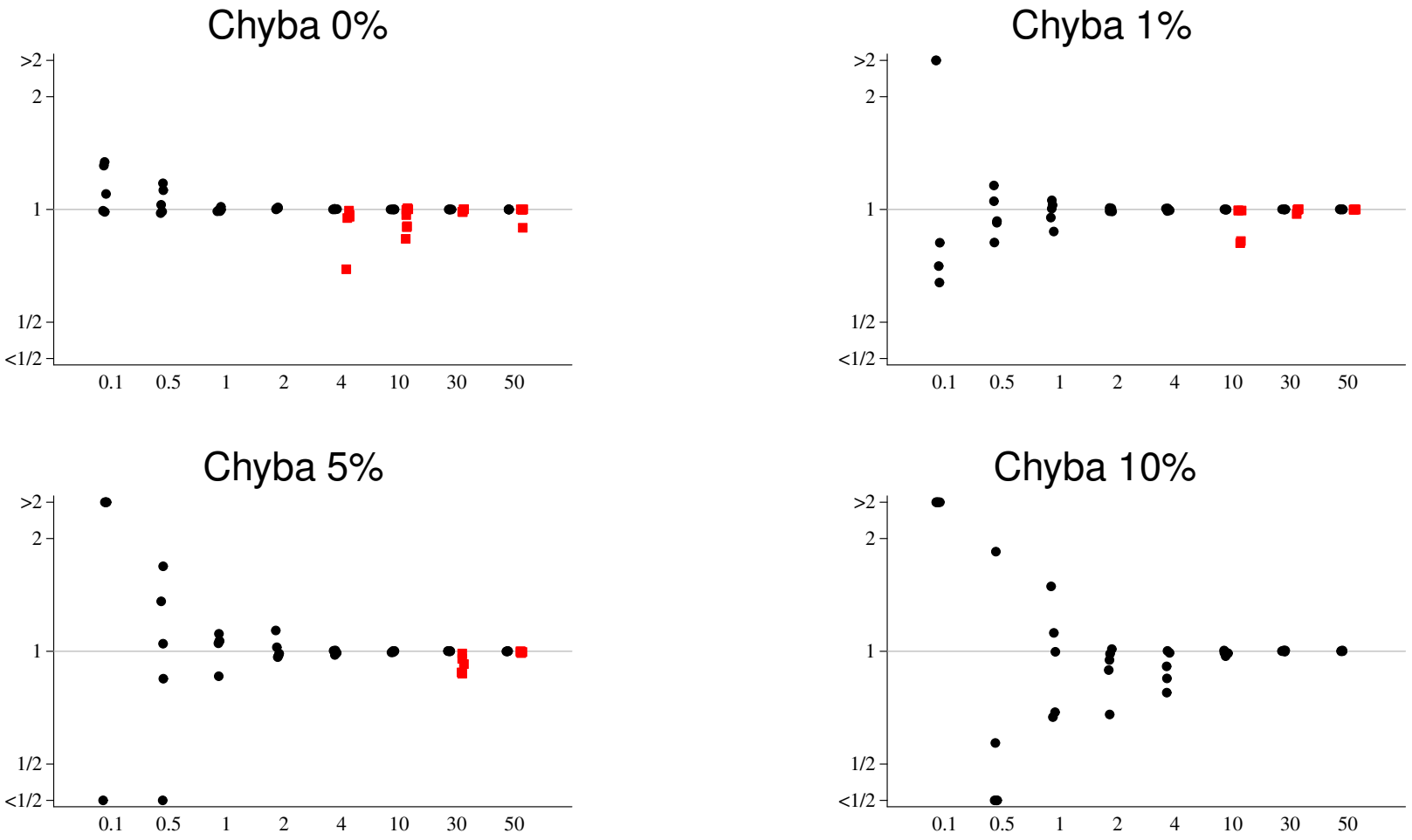
  Poissonova distribúcia so strednou hodnotou c_k
- S pravdepodobnosťou e^{-c_k} pokrytie 0, nevidíme v spektre
- Pravdepodobnosť pozorovania j výskytov

  $p_j = \frac{c_k^j e^{-c_k}}{j!} \cdot \frac{1}{1-e^{-c_k}} = f(j; c_k)$

Skutočný model zahŕňa

- chyby v čítaniach
- repetitívne sekvencie

Výsledky na simulovaných dátach (1 Mbp genóm)



CovEst čierne, KSA [Williams et al. 2013] červené;

y: pomer predpovedaného a skutočného c

Illumina čítania z *E. coli*

Veľkosť genómu 4.64 Mbp, 151bp Illumina

Odhad veľkosti genómu v Mbp

Method	$c = 0.5$	$c = 1$	$c = 2$	$c = 4$	$c = 10$	$c = 30$	$c = 50$
RE	4.16	4.70	4.58	4.63	4.71	4.69	4.68
KSA	N/A	N/A	N/A	6.03	4.61	4.59	4.58

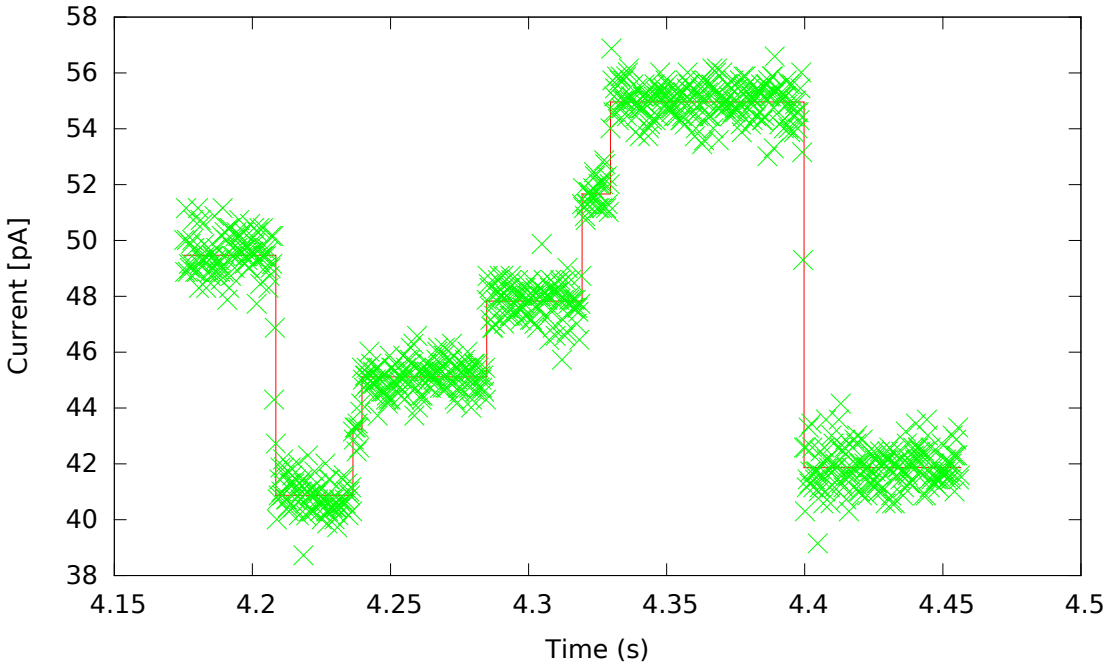
Príklad 3: Base calling pre MinION (DeepNano)

MinION early access program

v spolupráci s J. Nosekom (Prírodovedecká fakulta UK)

Nanopórové sekvenovanie

- DNA prechádzajúca cez nanopór spôsobuje zmenu elektrického napätia podľa aktuálneho kontextu k báz
- Priebeh elektrického signálu sa segmentuje na jednotlivé **udalosti**; každá udalosť ideálne zodpovedá posunu o 1 bázu

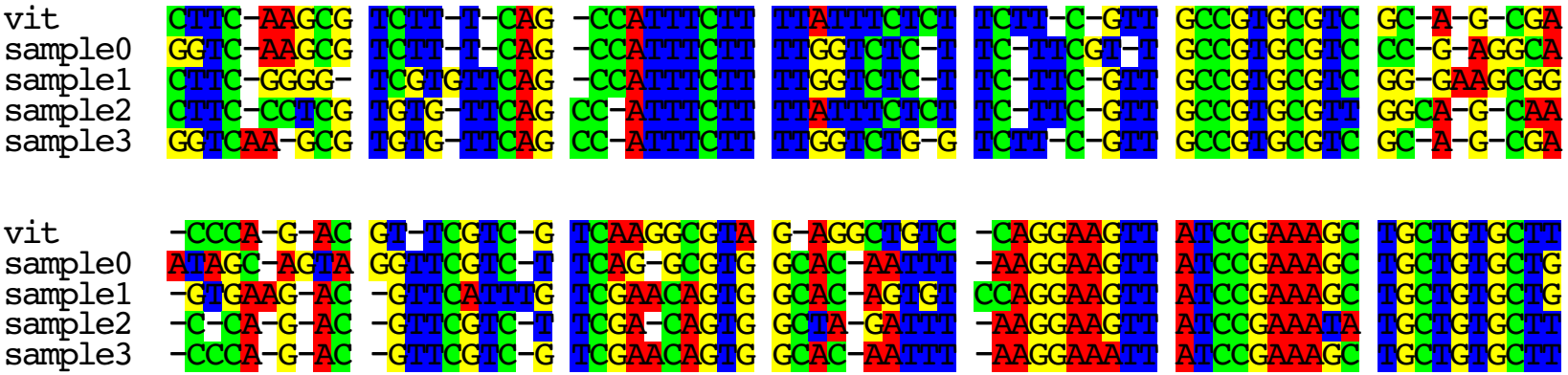


Basecalling pre nanopórové sekvenovanie

- Basecaller prekladá postupnosť udalostí do DNA sekvencie
- Udalosti sa “prekrývajú” obvykle $k - 1$ bázami
- Tradične reprezentované pomocou HMM (Metrichor, NanoCall)

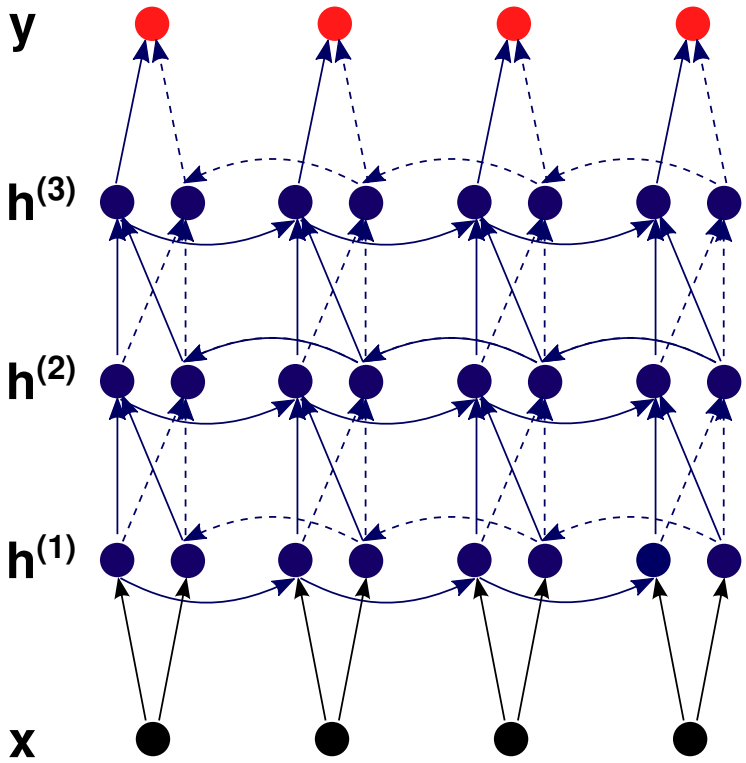
  štruktúra HMM á la de Bruijnov graf; skryté stavy = k-tice DNA

  emisné pravdepodobnosti Gaussiány (dané výrobcom zariadenia)
- Veľká neistota $\Rightarrow$ veľká chybovosť



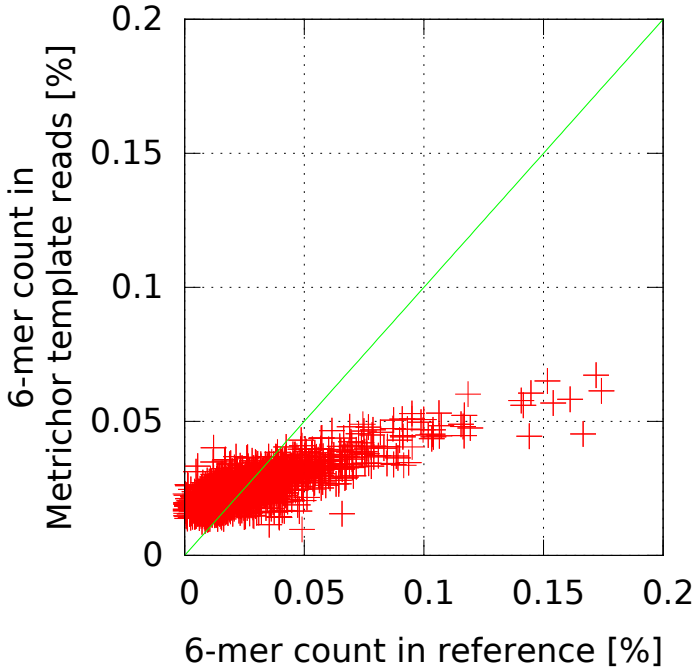
Diskriminatívne modelovanie namiesto generatívneho

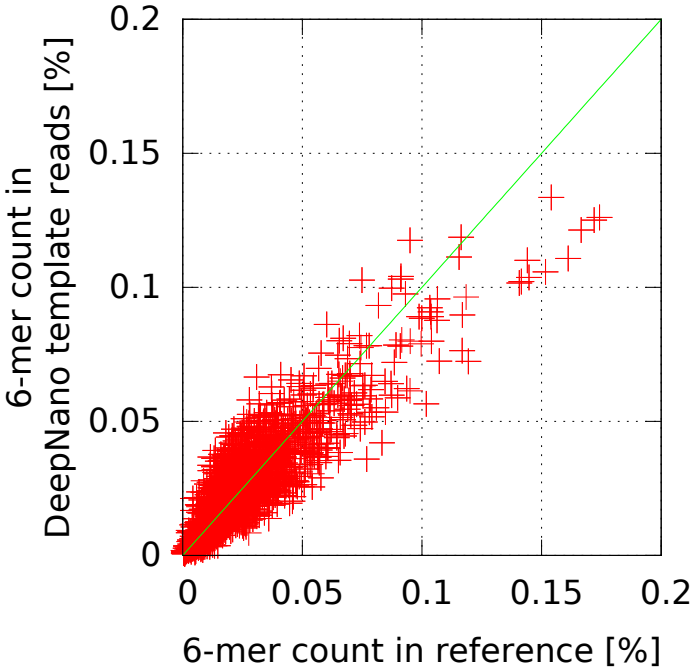
- Rekurentné neurónové siete
- Vstupné vektory: pre každú udalosť stredná hodnota, štd. odchýlka, dĺžka
- Výstupná vrstva: pre každú udalosť 0, 1 alebo 2 bázy
- Trénovanie cca 2 týždne, predikcia veľmi rýchla



Porovnanie s Oxford Nanopore base callerom

		E. coli	*K. pneumoniae*
Template reads	Metrichor	71.3%	68.1%
	DeepNano	77.9%	76.3%
Complement reads	Metrichor	71.4%	69.5%
	DeepNano	76.4%	75.7%
2D reads	Metrichor	86.8%	84.8%
	DeepNano	88.5%	86.7%


0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
6-mer count in
Metrichor template reads [%]
6-mer count in reference [%]


0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
6-mer count in
DeepNano template reads [%]
6-mer count in reference [%]

GAML: genome assembly by maximum likelihood

(Algorithms for Molecular Biology 2015)

Vlado Boža, Broňa Brejová

How Big is That Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra (SPIRE 2015)

Michal Hozza, Werner Krampl, Broňa Brejová

DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads (arXiv 2016)

Using Sequence Ensembles for Seeding Alignments of MinION Sequencing Data (submitted 2016)

Vlado Boža, Rasťo Rabatin, Broňa Brejová

Funding acknowledgments: VEGA 1/0719/14 a 1/0684/16, APVV-14-0253.

Thanks: Jozef Nosek, Prírodovedecká fakulta UK

Výpočtová biológia na FMFI UK

http://compbio.fmph.uniba.sk/ **hľadáme doktorandov**

bakalársky študijný program: http://compbio.fmph.uniba.sk/program

letná škola NGSchool 2016: http://ngschool.eu/ (prihlášky do 1.7.2016)