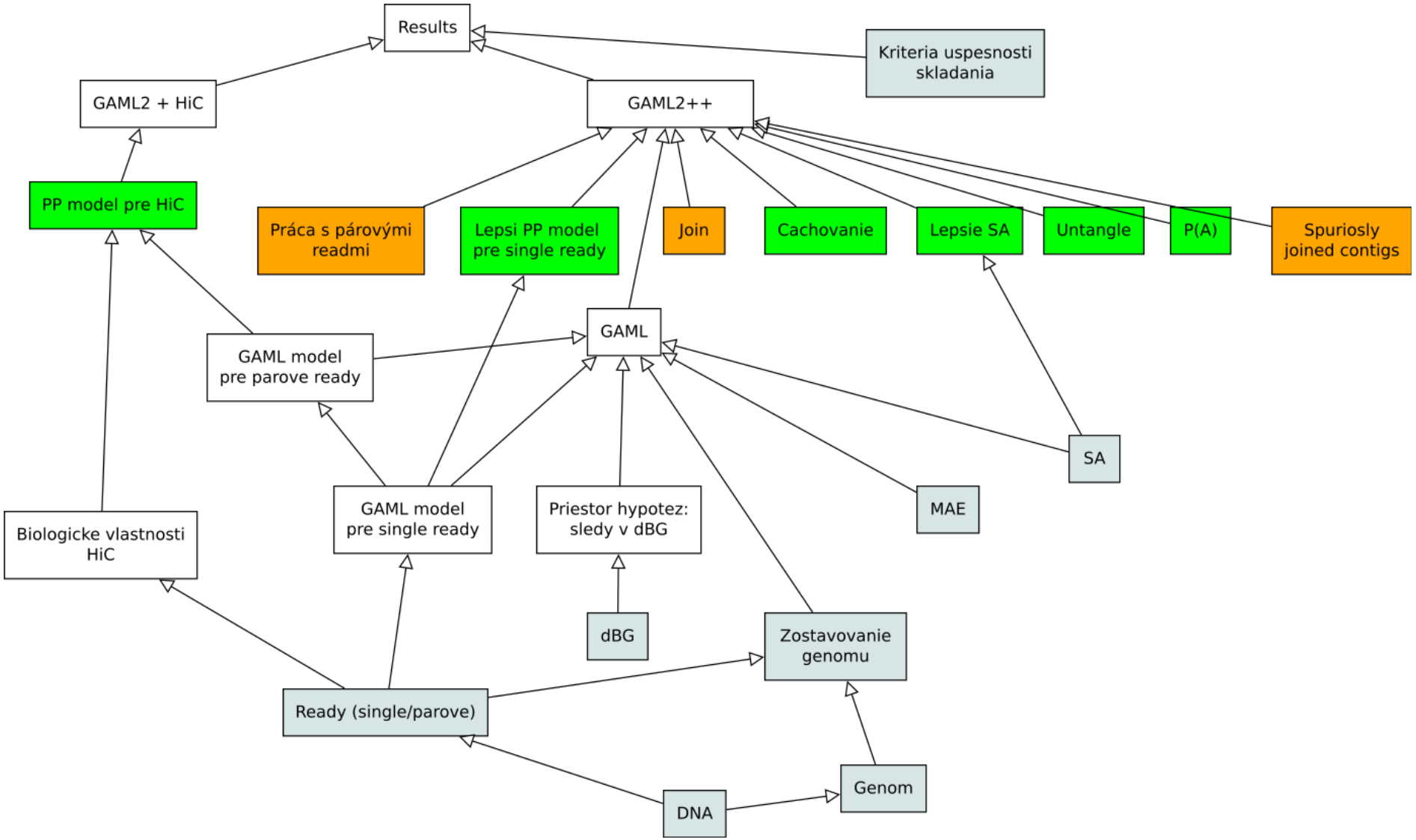
Pravdepodobnostné modely pre zostavovanie genómov s využitím chromatínových interakcií

Bc. Askar Gafurov
Vedúca práce: doc. Mgr. Bronislava Brejová, PhD.

Fakulta matematiky, fyziky a informatiky UK

14. júna 2018


Results
Kriteria uspesnosti skladania
GAML2 + HiC
GAML2++
PP model pre HiC
Práca s párovými readmi
Lepsi PP model pre single ready
Join
Cachovanie
Lepsie SA
Untangle
P(A)
Spuriosly joined contigs
GAML
GAML model pre parove ready
SA
GAML model pre single ready
Priestor hypotez: sledy v dBG
MAE
Biologicke vlastnosti HiC
dBG
Zostavovanie genomu
Ready (single/parove)
Genom
DNA

Plán prezentácie

- Klasické zostavovanie genómu
- GAML: Pravdepodbnostné zostavovanie genómu
- HiC
- Vylepšenia GAML2
- Výsledky
- Budúca práca
- Pripomienky

- DNA := slovo nad abecedou $\{A, C, G, T\}$. Reverzný komplement DNA S značíme ako $rc(S)$
- Genóm := množina DNA, jednotlivé DNA sa volajú *chromozómy*

- DNA := slovo nad abecedou $\{A, C, G, T\}$. Reverzný komplement DNA S značíme ako $rc(S)$
- Genóm := množina DNA, jednotlivé DNA sa volajú *chromozómy*
- čítanie (*read*) := krátke slovo nad abecedou $\{A, C, G, T\}$, *skoro* podslovo niektorej z DNA (resp. komplementu)
- chyby čítania := substitucie, inzercie, delecie

- DNA := slovo nad abecedou $\{A, C, G, T\}$. Reverzný komplement DNA S značíme ako $rc(S)$
- Genóm := množina DNA, jednotlivé DNA sa volajú *chromozómy*
- čítanie (*read*) := krátke slovo nad abecedou $\{A, C, G, T\}$, *skoro* podslovo niektorej z DNA (resp. komplementu)
- chyby čítania := substitucie, inzercie, delecie
- zostavovanie genómu := nájdenie "najvierohodnejšieho"genómu na základe pozorovaných čítaní

Klasické metódy zostavovania genómu

Overlap - Layout - Consensus

1. Nájdeme dva nadväzné useky v množine usekov
2. Zlepíme ich spolu a vložíme do množiny
3. Opakuj krok 1

Klasické metódy zostavovania genómu

Overlap - Layout - Consensus

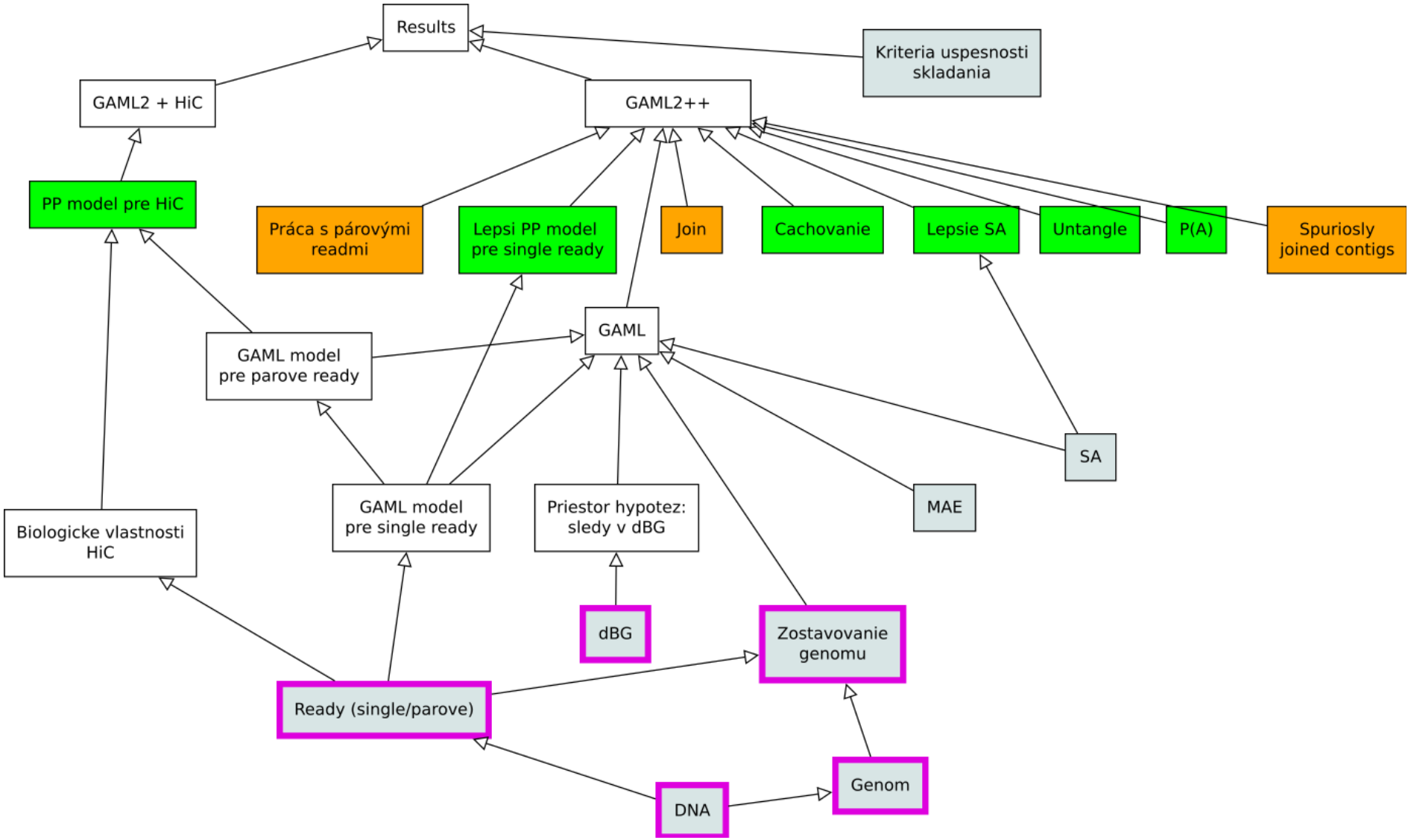
1. Nájdeme dva nadväzné useky v množine usekov
2. Zlepíme ich spolu a vložíme do množiny
3. Opakuj krok 1

de Bruijnove grafy (dBG)

1. na základe čítaní zostrojíme graf k-gramov, hrany sú medzi nadväznými
2. daný graf skomprimujeme (zlúčime vrcholy s jednoznačnou nadväznosťou)
3. v grafe nájdeme dlhé, "vierohodné"sledy

Problémy klasických metód zostavovania genómov

- Nie je možné určiť kvalitu výsledku
- Metódy sú šité na konkretné typy sekvenačných dát
- Algoritmy nevedia plnohodnotne využivať viacere typy sekvenačných dát naraz


Results
Kriteria uspesnosti skladania
GAML2 + HiC
GAML2++
PP model pre HiC
Práca s párovými readmi
Lepsi PP model pre single ready
Join
Cachovanie
Lepsie SA
Untangle
P(A)
Spuriosly joined contigs
GAML
GAML model pre parove ready
SA
GAML model pre single ready
Priestor hypotez: sledy v dBG
MAE
Biologicke vlastnosti HiC
dBG
Zostavovanie genomu
Ready (single/parove)
Genom
DNA

Pravdepodobnostné zostavovanie genómu

Aposteriórna pravdepodobnosť $P(A|R)$ genómu A vzhľadom na pozorované čítania R.
Najlepší genóm je ten, čo maximalizuje aposteriórnu pravdepodbnosť:

$$A^*(R) := \arg\max_{A \in \Theta} P(A|R)$$

Pravdepodobnostné zostavovanie genómu

Aposteriórna pravdepodobnosť $P(A|R)$ genómu A vzhľadom na pozorované čítania R.
Najlepší genóm je ten, čo maximalizuje aposteriórnu pravdepodbnosť:

$$A^*(R) := \arg\max_{A \in \Theta} P(A|R)$$

1. Ako určiť $P(A|R)$?
2. Ako určiť priestor hypotéz Θ?
3. Ako maximalizovať danú funkciu?

Bayesova veta

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

Bayesova veta

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

$$A^*(R) := \arg\max_{A \in \Theta} \frac{P(R|A)P(A)}{P(R)} = \arg\max_{A \in \Theta} P(R|A)P(A)$$

Bayesova veta

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

$$A^*(R) := \arg\max_{A \in \Theta} \frac{P(R|A)P(A)}{P(R)} = \arg\max_{A \in \Theta} P(R|A)P(A)$$

Predpoklad 1: jednotlivé čítania sú nezávislé

Stačí definovať model pre jedno čítanie: $P(R|A) = \prod_{r \in R} P(r|A)$

Predpoklad 2: všetky genómy majú rovnakú apriórnu pravdepobnosť

$$P(A) = const \Longrightarrow A^*(R) = \arg\max_{A \in \Theta} \prod_{r \in R} P(r|A)$$

Pravdepodobnosť pre jedno čítanie a jeden chromozóm

Pravdepodobnosť pozorovania čítania r z pozície i chromozómu S

$$b_i(r, S) := \varepsilon^q (1 - \varepsilon)^m$$

$$P(r|S) := \sum_{i=1}^{|S|} \frac{b_i(r, S) + b_i(r, rc(S))}{2|S|}$$

Pravdepodobnosť pre jedno čítanie a jeden chromozóm

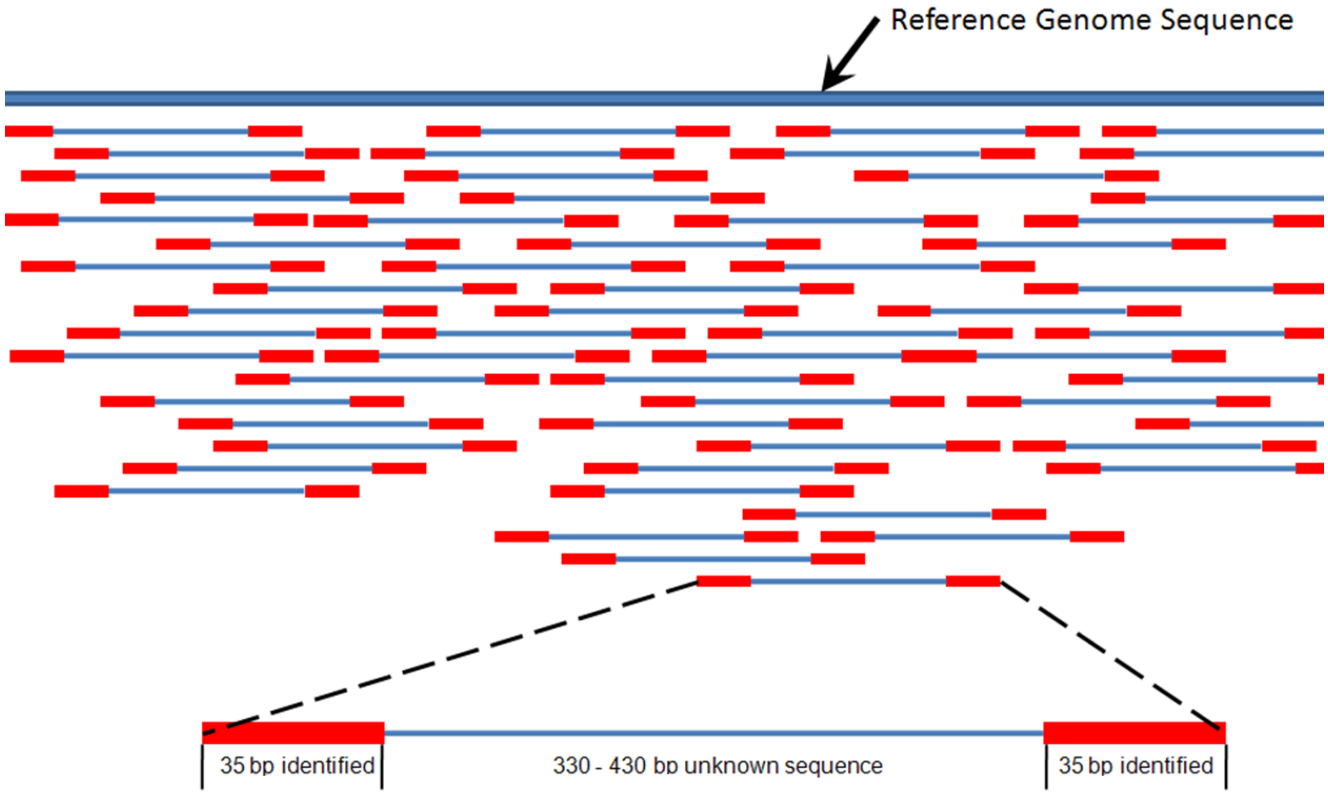
Pravdepodobnosť pozorovania čítania r z pozície i chromozómu S

$$b_i(r, S) := \varepsilon^q (1 - \varepsilon)^m$$

$$P(r|S) := \sum_{i=1}^{|S|} \frac{b_i(r, S) + b_i(r, rc(S))}{2|S|}$$

$$P(r, S) \approx \frac{1}{2|S|} \sum_{a \in AL(r,S) \cup AL(r,rc(S))} \varepsilon^{a_q} (1 - \varepsilon)^{a_m}$$

Párové čítania



Pravdepodobnosť pre jedno párové čítanie a jeden chromozóm

$$P(r,S) \approx \frac{1}{2|S|} \sum_{\substack{a_1 \in BAL(r_1,S) \\ a_2 \in BAL(r_2,S)}} p_{single}(a_1) p_{single}(a_2) p_{orient}(a_1,a_2)$$

$$p_{orient}(x,y) := \begin{cases} 0 & \text{if } x_o + y_o \neq 1 \vee y_p - x_p < l \\ p_{insert}(y_p - x_p) & \text{otherwise} \end{cases}$$

Heuristika na maximalizáciu

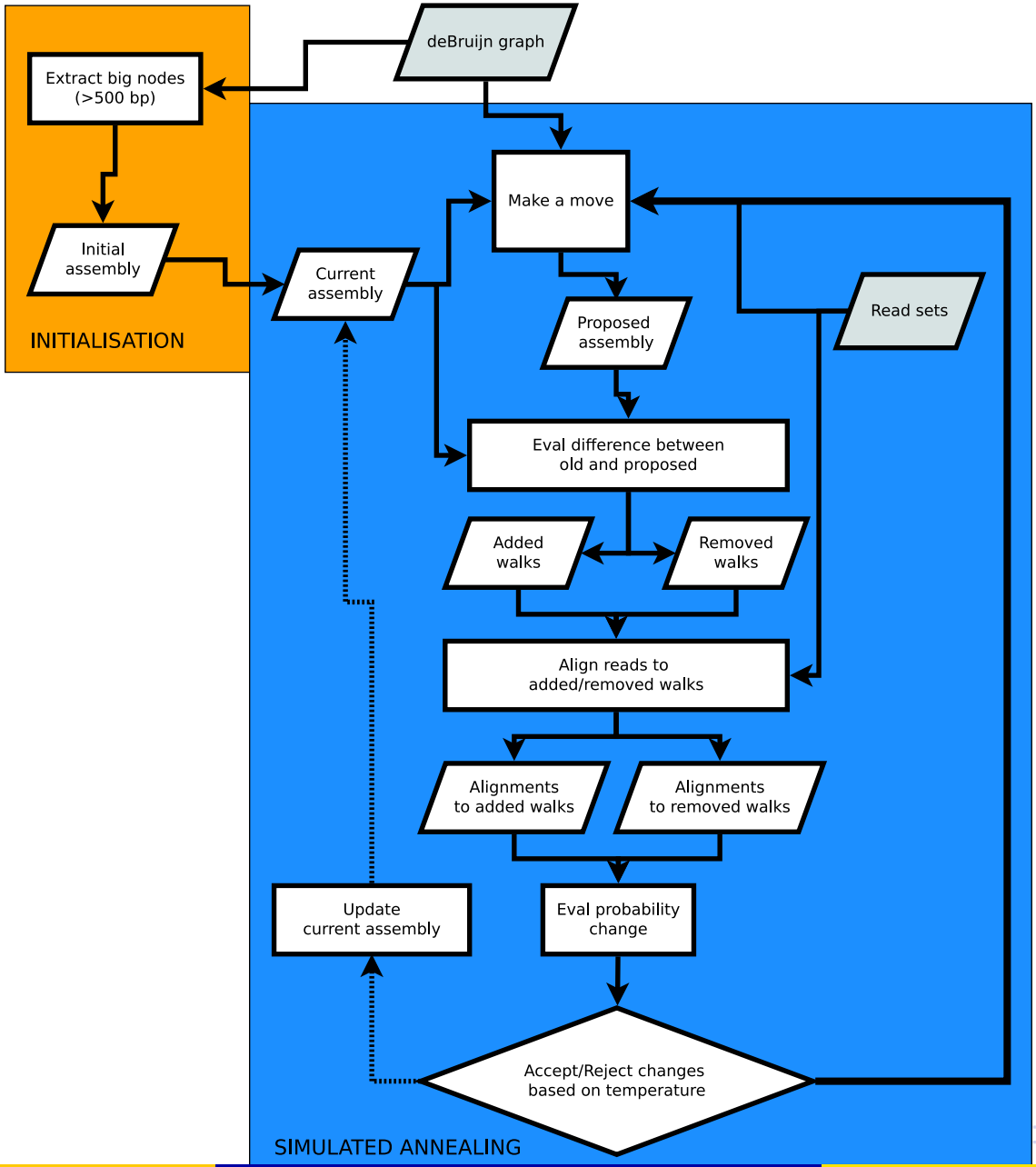
Simulované žíhanie — hill climbing + dovolené zhoršenie

$$P_{transit}(x, x') := \begin{cases} 1 & \text{if } f(x') > f(x) \\ \exp\left(\frac{f(x')-f(x)}{T}\right) & \text{otherwise} \end{cases}$$

$$T(n) := \frac{T_0}{\log\left(1 + \frac{n}{d}\right)}$$

Priestor hypotéz

- Hypotézy — konečná množina sledov v komprimovanom de Bruijnovom grafe, postavenom nad vstupnými dátami
- Počiatočny stav — vrcholy (sledy dĺžky 1), ktorým zodpovedajú dostatočne dlhé sekvencie (≥ 500 bp)
- Prechody (*moves*) — spájanie dvoch sledov, náhodné predĺženie sledu, rozstrihnutie sledu, etc.


deBruijn graph
Extract big nodes (>500 bp)
Initial assembly
INITIALISATION
Make a move
Current assembly
Proposed assembly
Read sets
Eval difference between old and proposed
Added walks
Removed walks
Align reads to added/removed walks
Alignments to added walks
Alignments to removed walks
Update current assembly
Eval probability change
Accept/Reject changes based on temperature
SIMULATED ANNEALING

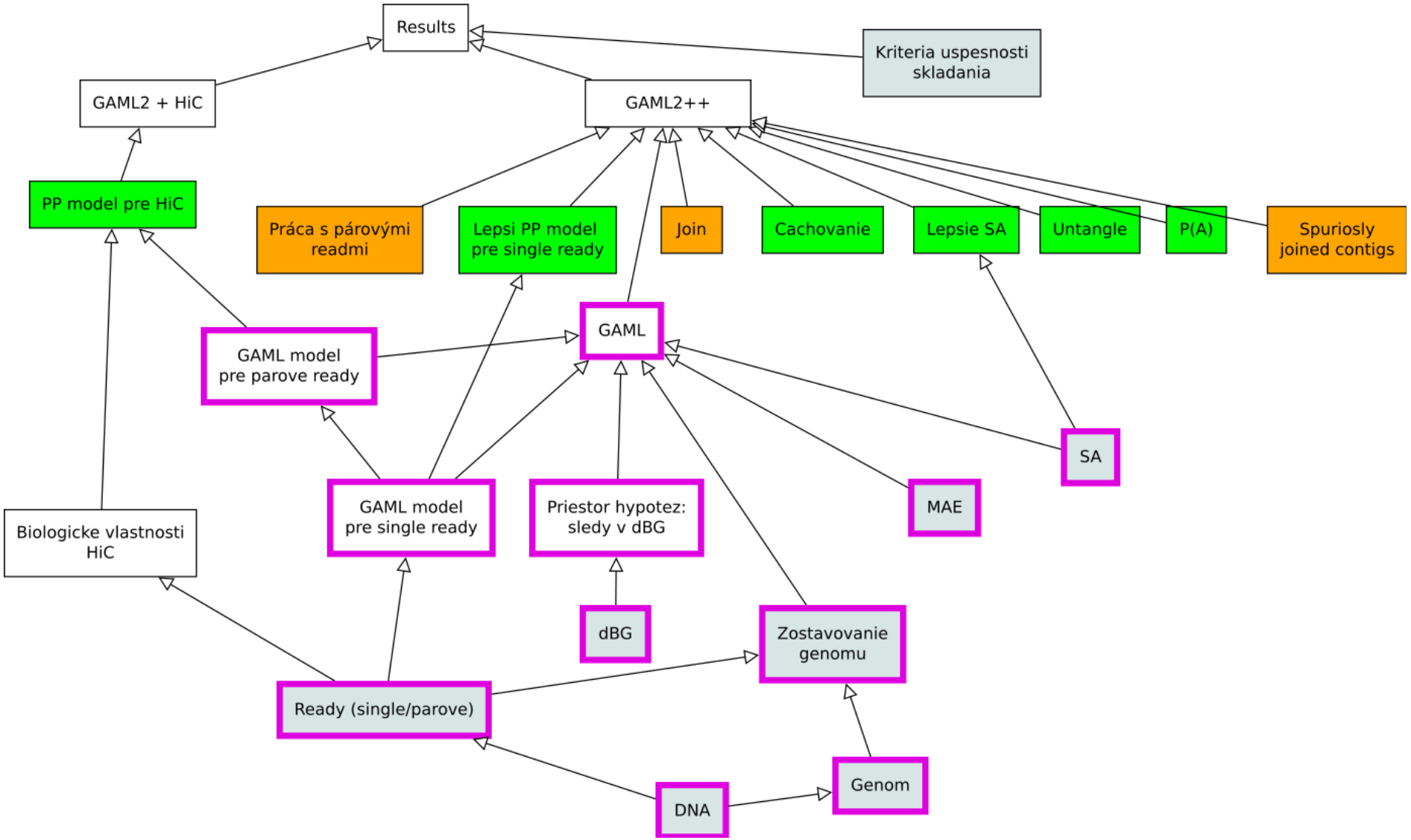
Vyhody

- Meranie kvality výsledkov (pôvodný účel tohto PP modelu)
- Možnosť kombinácie viacerych typov dát:

$$\arg\max_A P(A|R_1,\ldots,R_k) = \arg\max_A P(R_1|A)\ldots P(R_k|A)$$

Nevyhody

- Výpočtová náročnosť (v každej iterácii treba zarovnávať čítania ku novému genómu)


Results
Kriteria uspesnosti skladania
GAML2 + HiC
GAML2++
PP model pre HiC
Práca s párovými readmi
Lepsi PP model pre single ready
Join
Cachovanie
Lepsie SA
Untangle
P(A)
Spuriosly joined contigs
GAML
GAML model pre parove ready
SA
GAML model pre single ready
Priestor hypotez: sledy v dBG
MAE
Biologicke vlastnosti HiC
dBG
Zostavovanie genomu
Ready (single/parove)
Genom
DNA

Nový typ dát — frekvencie chromatínových interákcií

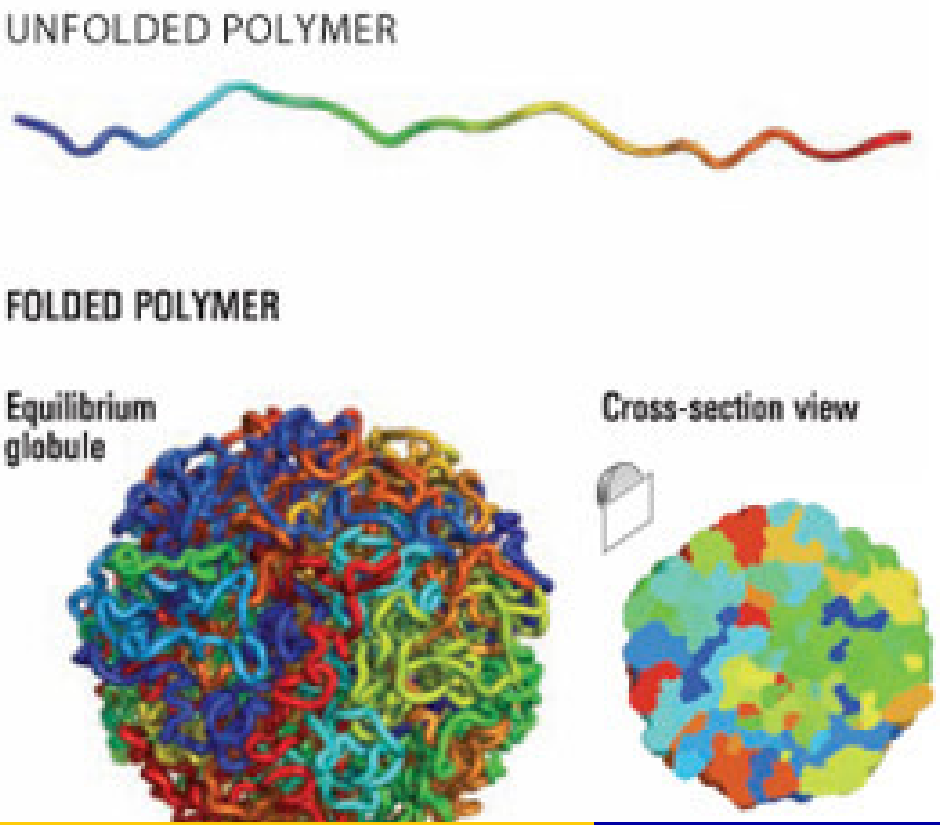
3D štruktúra DNA

DNA v bunke je uložená ako kompaktná štruktúra, nazyvaná *chromatín* (Berkum et al. 2010, Science)

Nový typ dát — frekvencie chromatínových interákcií

3D štruktúra DNA

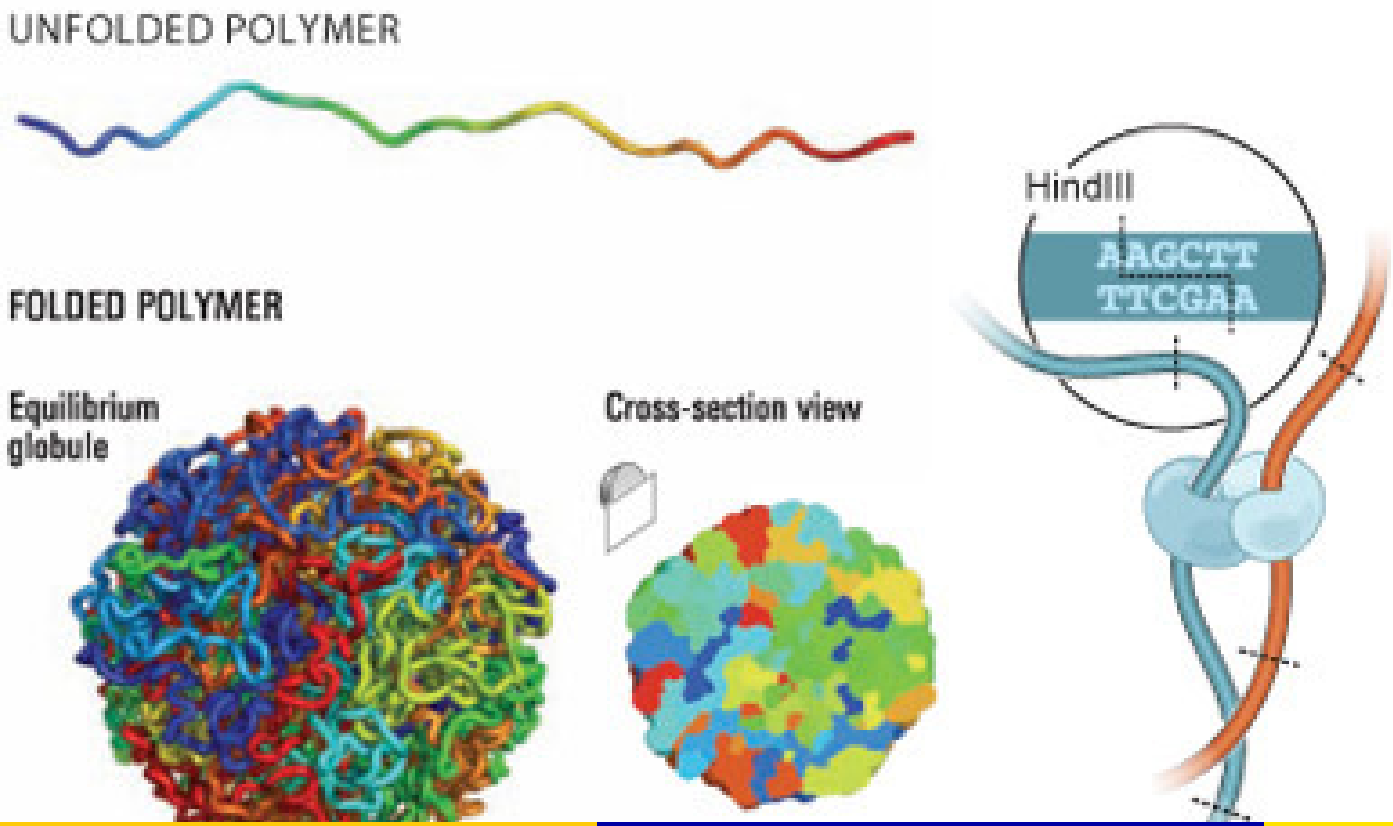
DNA v bunke je uložená ako kompaktná štruktúra, nazyvaná *chromatín* (Berkum et al. 2010, Science)



Nový typ dát — frekvencie chromatínových interákcií

3D štruktúra DNA

DNA v bunke je uložená ako kompaktná štruktúra, nazyvaná *chromatín* (Berkum et al. 2010, Science)



Ako vyzerajú dáta?

- Na prvý pohľad sú ako obyčajné párové čítania
- Jednotlivé časti však môžu pochádzať z rôznych chromozómov

Pozorovania o charaktere HiC dát [Kaplan a Dekker, 2014, Nature]

- *cis-trans-ratio*: čítania z jedného chromozómu (cis-) sú častejšie ako z dvoch rôznych (trans-)
- *distance-dependent decay*: distribúcia vzdialenosti medzi častiami cis-čítaní má exponenciálny charakter

PP model pre HiC dáta

$$P_{hic}(r|A) = p_{cis} \cdot P_{cis}(r|A) + (1 - p_{cis}) \cdot P_{trans}(r|A)$$

PP model pre HiC dáta

$$P_{hic}(r|A) = p_{cis} \cdot P_{cis}(r|A) + (1 - p_{cis}) \cdot P_{trans}(r|A)$$

$$P_{cis}(r|S) \approx \frac{1}{2|S|} \sum_{\substack{a^{(1)} \in BAL(r_1,S) \\ a^{(2)} \in BAL(r_2,S)}} p_{single}\left(a^{(1)}\right) p_{single}\left(a^{(2)}\right) f_{exp}\left(|a_p^{(1)} - a_p^{(2)}|; \hat{\lambda}_{MLE}(R,S)\right)$$

PP model pre HiC dáta

$$P_{hic}(r|A) = p_{cis} \cdot P_{cis}(r|A) + (1 - p_{cis}) \cdot P_{trans}(r|A)$$

$$P_{cis}(r|S) \approx \frac{1}{2|S|} \sum_{\substack{a^{(1)} \in BAL(r_1,S) \\ a^{(2)} \in BAL(r_2,S)}} p_{single}\left(a^{(1)}\right) p_{single}\left(a^{(2)}\right) f_{exp}\left(|a_p^{(1)} - a_p^{(2)}|; \hat{\lambda}_{MLE}(R,S)\right)$$

$$\hat{\lambda}_{MLE}(R,S) := \frac{\sum_{(r_1,r_2) \in R} |BAL(r_1,S)| \cdot |BAL(r_2,S)|}{\sum_{(r_1,r_2) \in R} \sum_{\substack{a^{(1)} \in BAL(r_1,S) \\ a^{(2)} \in BAL(r_2,S)}} \left|a_p^{(1)} - a_p^{(2)}\right|}$$

PP model pre HiC dáta

$$P_{trans}(r|S_1, S_2) \approx \sum_{\substack{a_1 \in BAL(r_1, S_1) \\ a_2 \in BAL(r_2, S_2)}} \frac{p_{single}(a_1)}{2|S_1|} \frac{p_{single}(a_2)}{2|S_2|}$$

$$P_{trans}(r|A) \approx \sum_{1 \le i < j \le |A|} \frac{|S_i| \cdot |S_j|}{\sum_{1 \le u < v \le |A|} |S_u| \cdot |S_v|} \left(\frac{1}{2} P(r|S_i, S_j) + \frac{1}{2} P(r|S_j, S_i) \right)$$

PP model pre HiC dáta

$$\phi(r,S) := \sum_{\substack{a^{(1)}\in BAL(r_1,S)\\ a^{(2)}\in BAL(r_2,S)}} p_{single}\left(a^{(1)}\right) p_{single}\left(a^{(2)}\right) f_{exp}\left(|a_p^{(1)} - a_p^{(2)}|;\hat{\lambda}\right)$$

$$\psi(r,S_1,S_2) := \sum_{\substack{a_1\in BAL(r_1,S_1)\\ a_2\in BAL(r_2,S_2)}} p_{single}(a_1)p_{single}(a_2)$$

$$\Phi(r,A) := \sum_{S\in A}\phi(r,S)$$

$$\Psi(r,A) := \sum_{1\le i<j\le |A|}(\psi(r,S_i,S_j)+\psi(r,S_j,S_i))$$

PP model pre HiC dáta

$$P_{hic}(R|A) \approx \prod_{r \in R} \left(\frac{p_{cis}}{2 \sum_{S_i \in A} |S_i|} \Phi(r, A) + \frac{1 - p_{cis}}{8 \sum_{1 \leq u < v \leq |A|} |S_u| \cdot |S_v|} \Psi(r, A) \right)$$

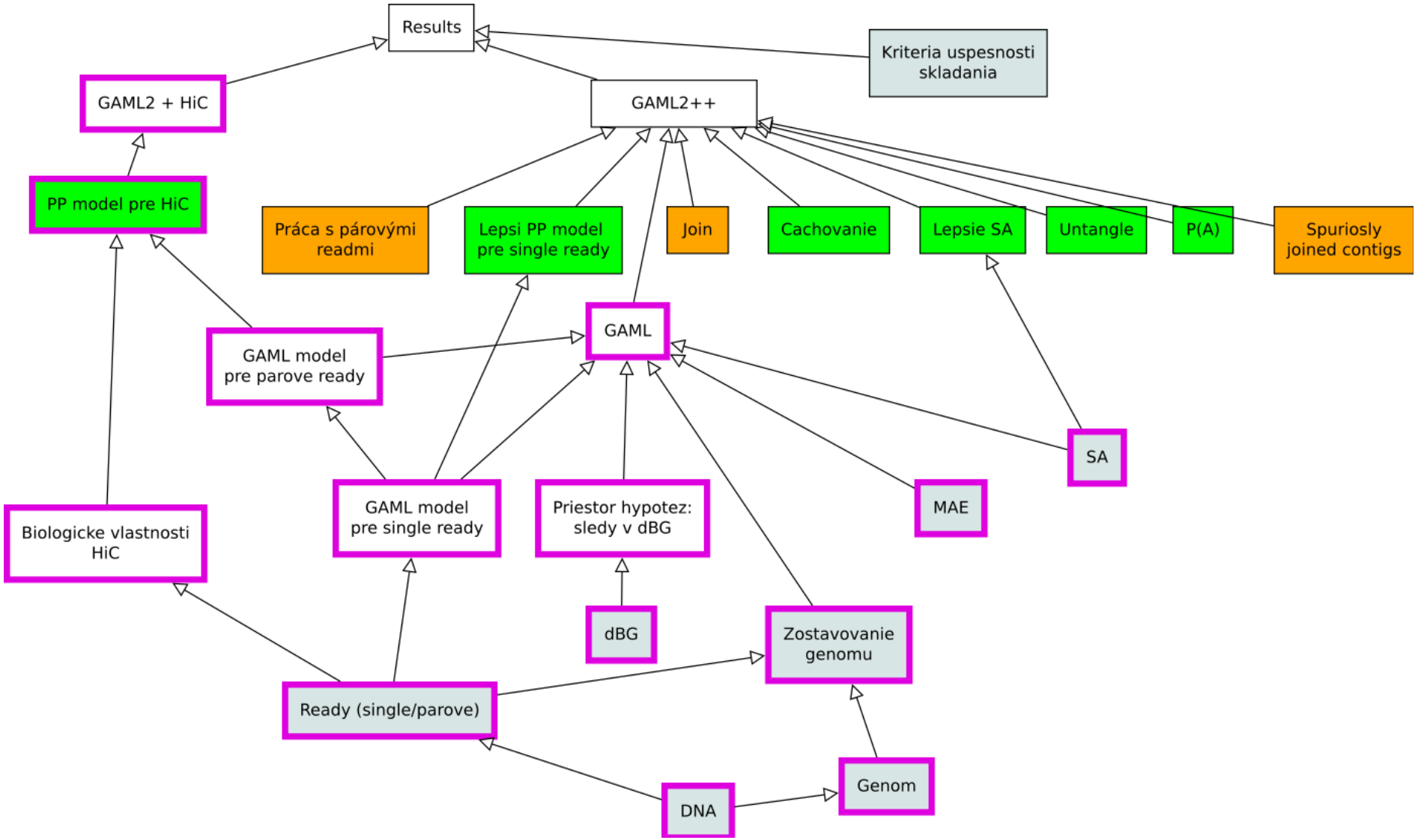
PP model pre HiC dáta

$N := A' - A,\ D := A - A',\ K := A \cap A'$

$$\Phi(r, A') = \Phi(r, A) - \Phi(r, D) + \Phi(r, N)$$

$$\Psi(r, A') = \Psi(r, A) - \Psi(r, D) - \sum_{S_1 \in D, S_2 \in K} (\psi(r, S_1, S_2) + \psi(r, S_2, S_1)) +$$
$$+ \Psi(r, N) + \sum_{S_1 \in N, S_2 \in K} (\psi(r, S_1, S_2) + \psi(r, S_2, S_1))$$

O(n) zložitosť od počtu sledov!


Results
Kriteria uspesnosti skladania
GAML2 + HiC
GAML2++
PP model pre HiC
Práca s párovými readmi
Lepsi PP model pre single ready
Join
Cachovanie
Lepsie SA
Untangle
P(A)
Spuriosly joined contigs
GAML
GAML model pre parove ready
SA
GAML model pre single ready
Priestor hypotez: sledy v dBG
MAE
Biologicke vlastnosti HiC
dBG
Zostavovanie genomu
Ready (single/parove)
Genom
DNA

Zlepšenia

1. Reimplementácia ťahu "Join with advice"
2. Cache pre vypočítané zarovnania
3. Lepší model pre čítanie, generované z daného miesta
4. Apriórna pravdepodobnosť genómu $P(A)$
5. Trest pre zle spojené useky
6. Podpora párových čítaní
7. Uprava ťahu "Random extend"
8. Nový ťah "Untangle crossed paths"
9. Vylepšené simulované žíhanie

Lepší model pre čítanie, generované z daného miesta

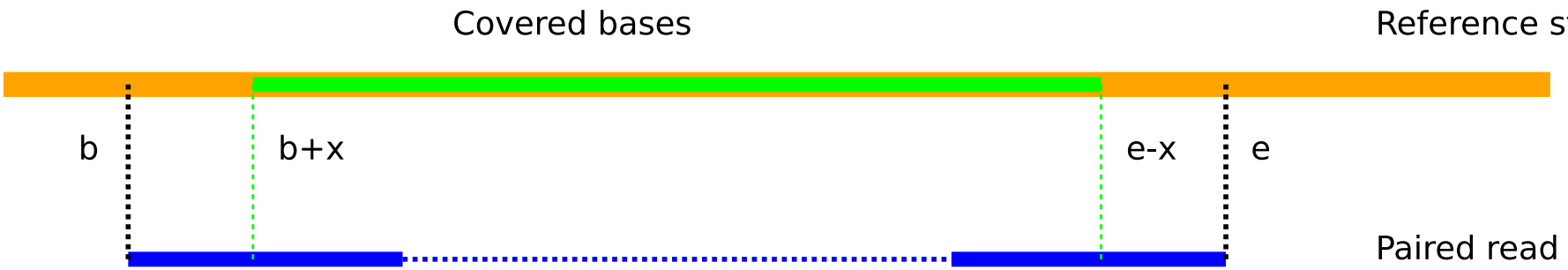
$$b_{new}(a) := \varepsilon_i^{a_i} \varepsilon_d^{a_d} \varepsilon_s^{a_s} (1 - 4\varepsilon_i - \varepsilon_d - 3\varepsilon_s)^{a_m}$$

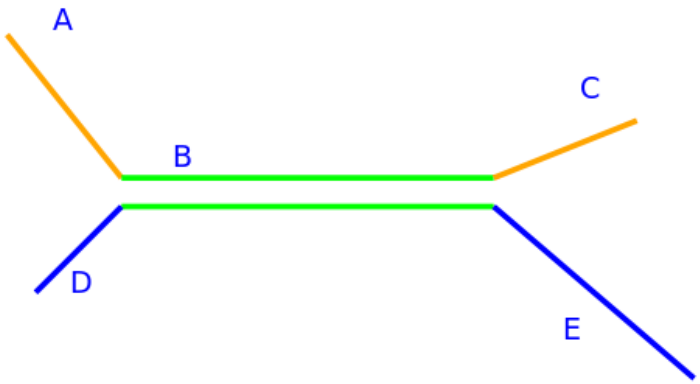
Lepší model pre čítanie, generované z daného miesta

$$b_{new}(a) := \varepsilon_i^{a_i} \varepsilon_d^{a_d} \varepsilon_s^{a_s} (1 - 4\varepsilon_i - \varepsilon_d - 3\varepsilon_s)^{a_m}$$

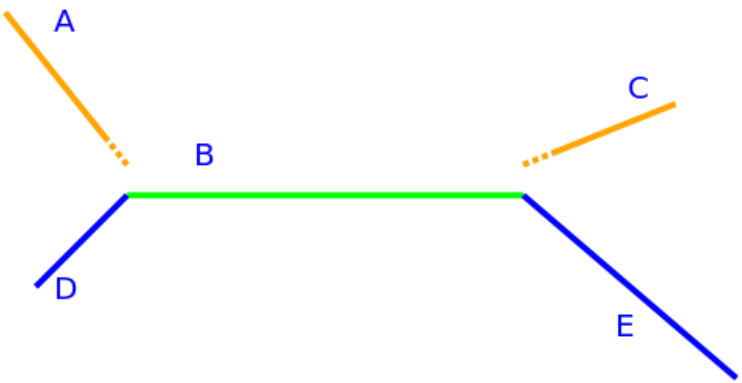
Apriórna pravdepodobnosť genómu $P(A)$

$$P(A) :\approx \prod_{S \in A} 4^{-|S|} = 4^{-\sum_{S \in A} |S|}$$

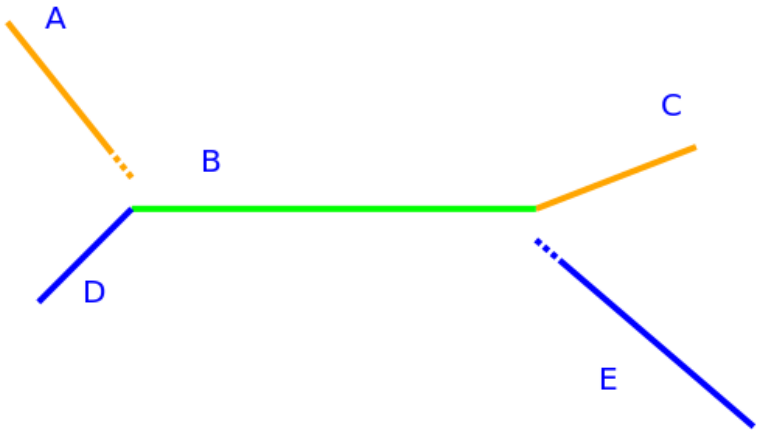

Covered bases
Reference strand
b
b+x
e-x
e
Paired read



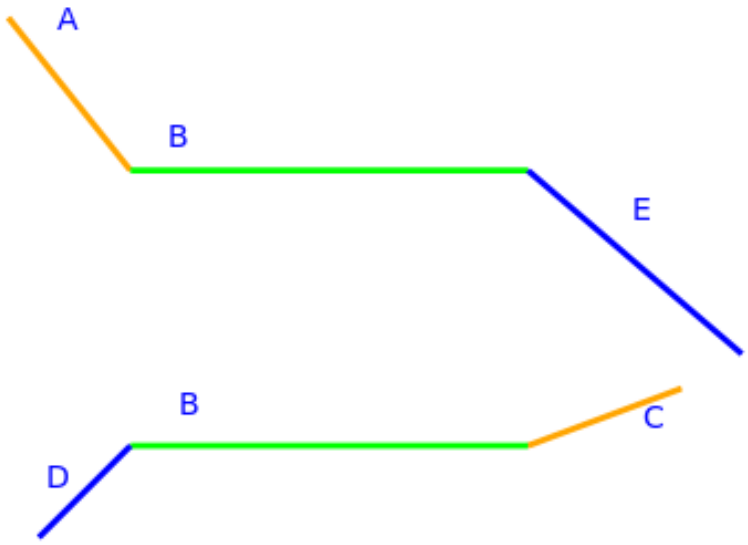
(a) Starting position



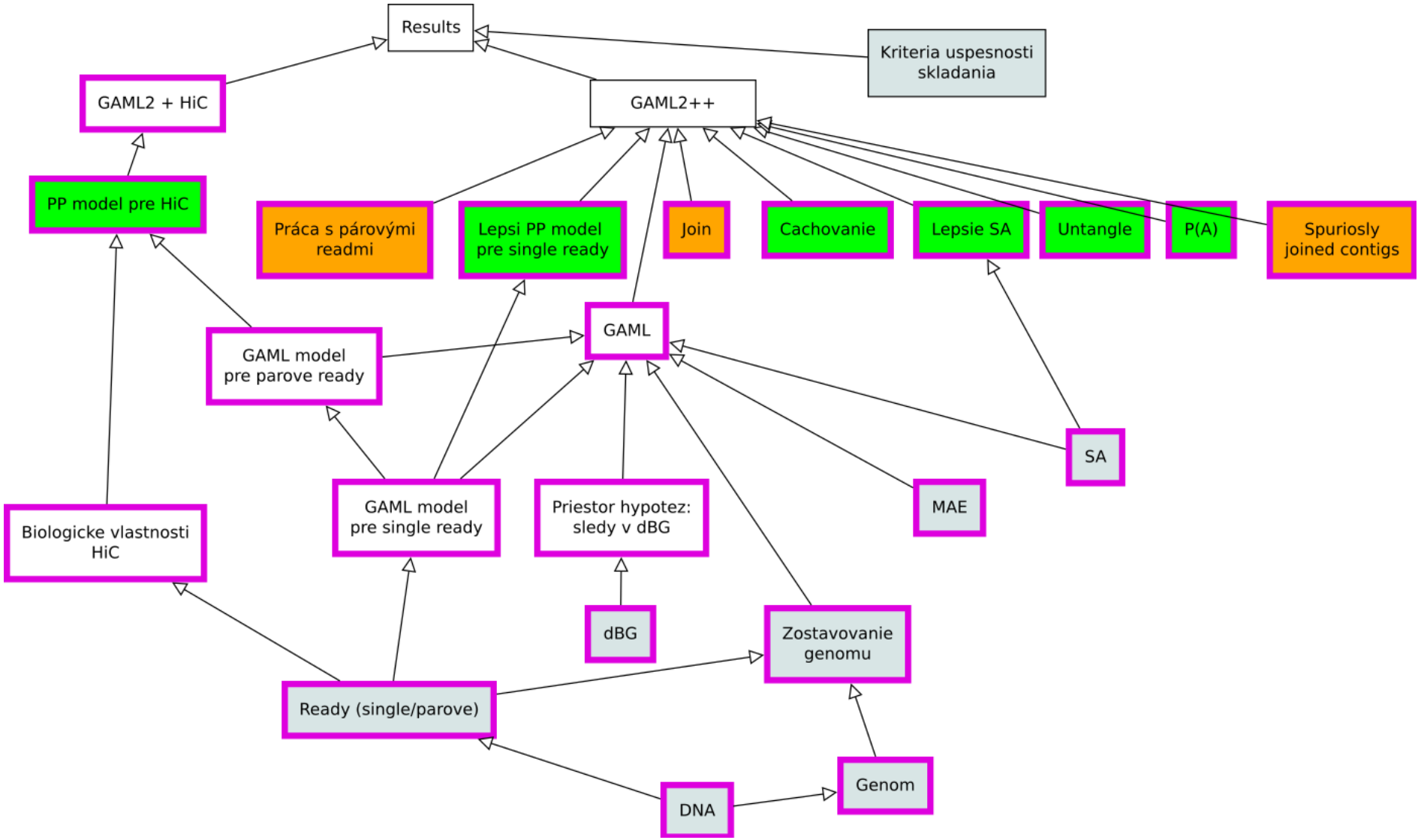
(b) Move type 1



(c) Move type 2

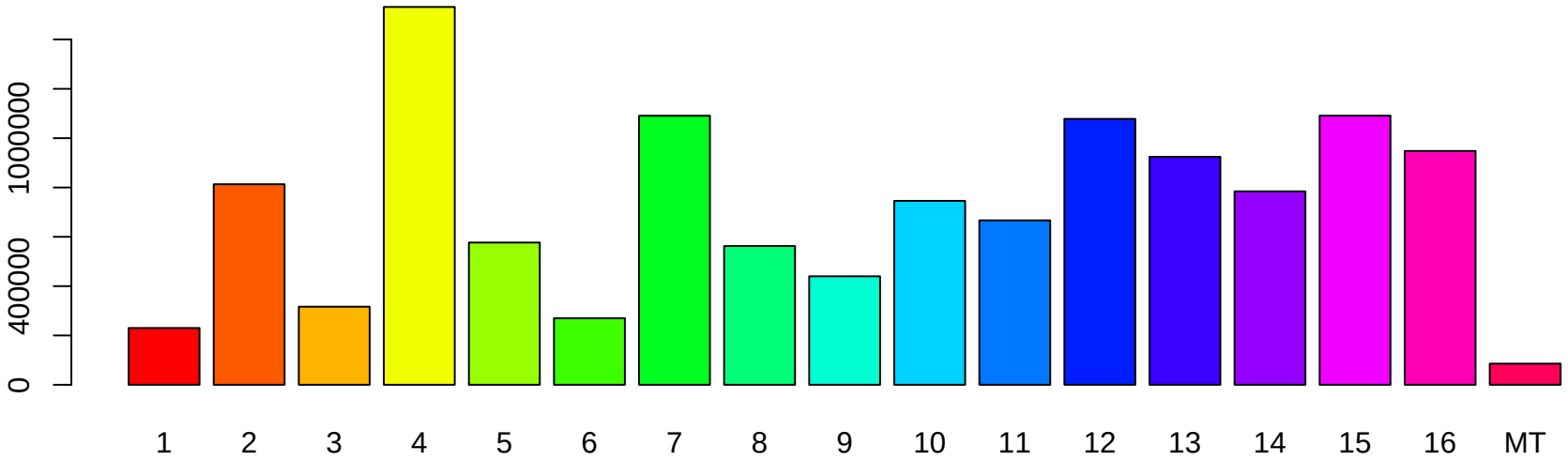


(d) Move type 3


Results
Kriteria uspesnosti skladania
GAML2 + HiC
GAML2++
PP model pre HiC
Práca s párovými readmi
Lepsi PP model pre single ready
Join
Cachovanie
Lepsie SA
Untangle
P(A)
Spuriosly joined contigs
GAML
GAML model pre parove ready
SA
GAML model pre single ready
Priestor hypotez: sledy v dBG
MAE
Biologicke vlastnosti HiC
dBG
Zostavovanie genomu
Ready (single/parove)
Genom
DNA

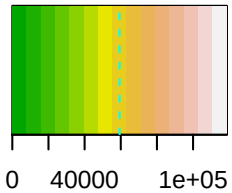
Dáta

Organizmus: Saccharomyces cerevisiae, 16 chromozómov + 1 mitochondriálny

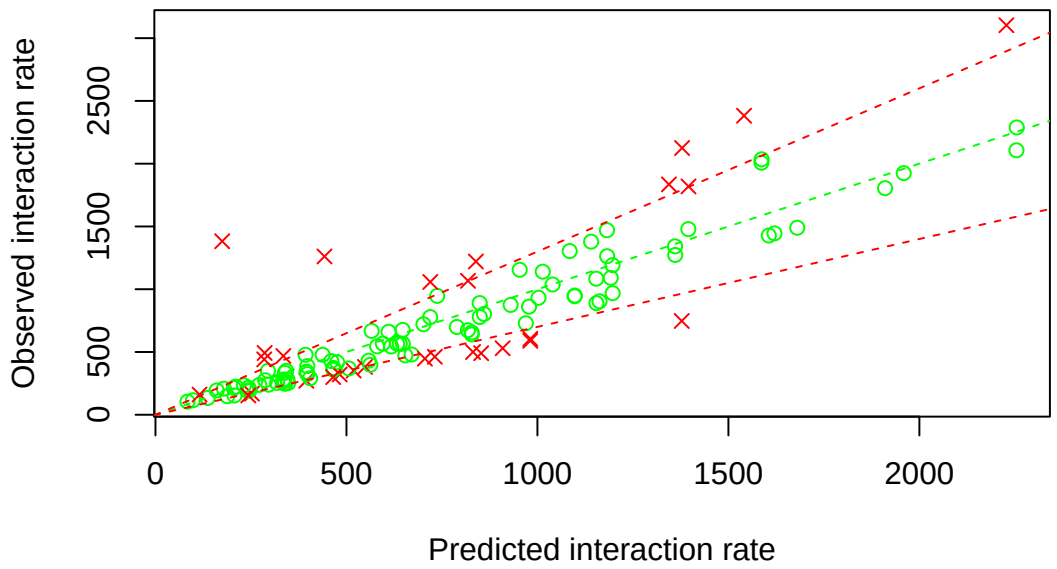


SRR4446972 — paired reads, $125 + 125$ bp each. Total number 6162477. Mean coverage is 128. Subsampled to 1 million reads $\rightarrow$ mean coverage $=$ 21

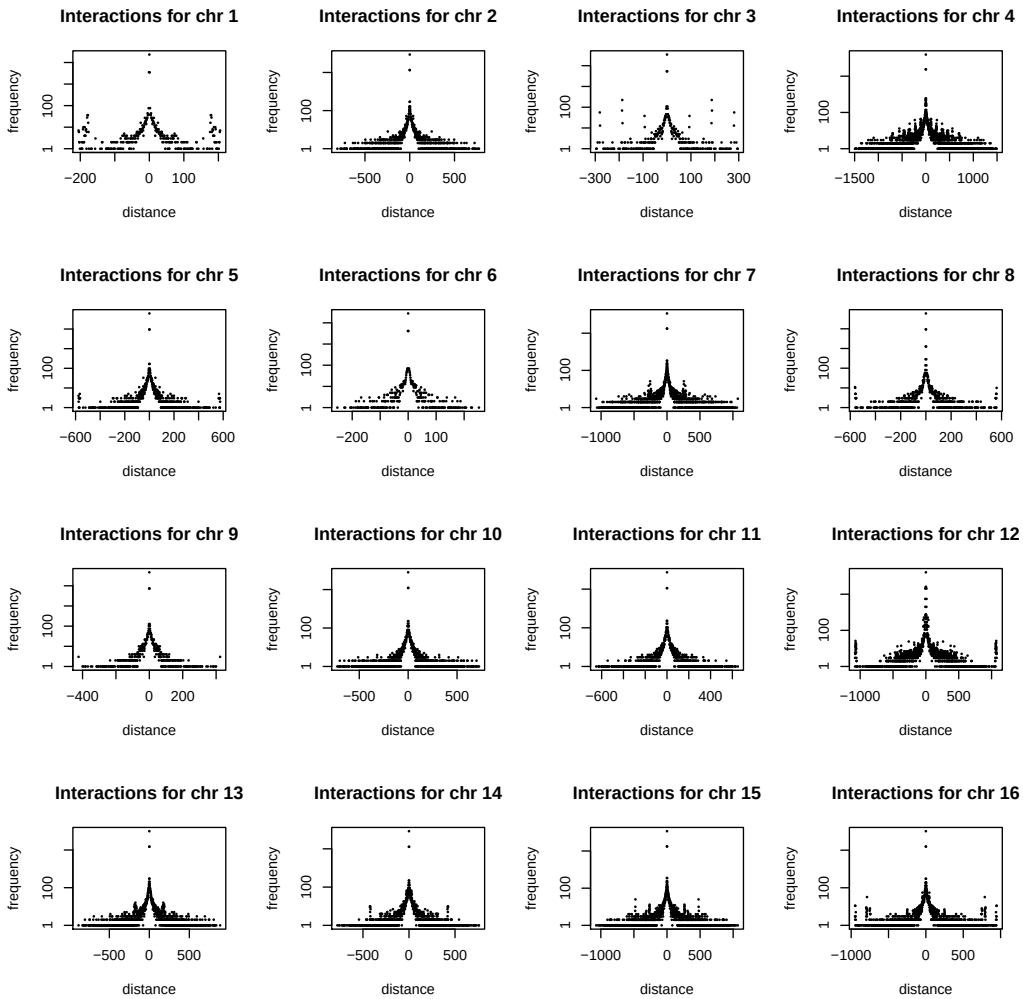
SRR5077811 — Hi-C reads, $102 + 102$ bp each. Total amount 91776947. Mean coverage is 1506. Subsampled to 1 million reads $\rightarrow$ mean coverage $=$ 16.



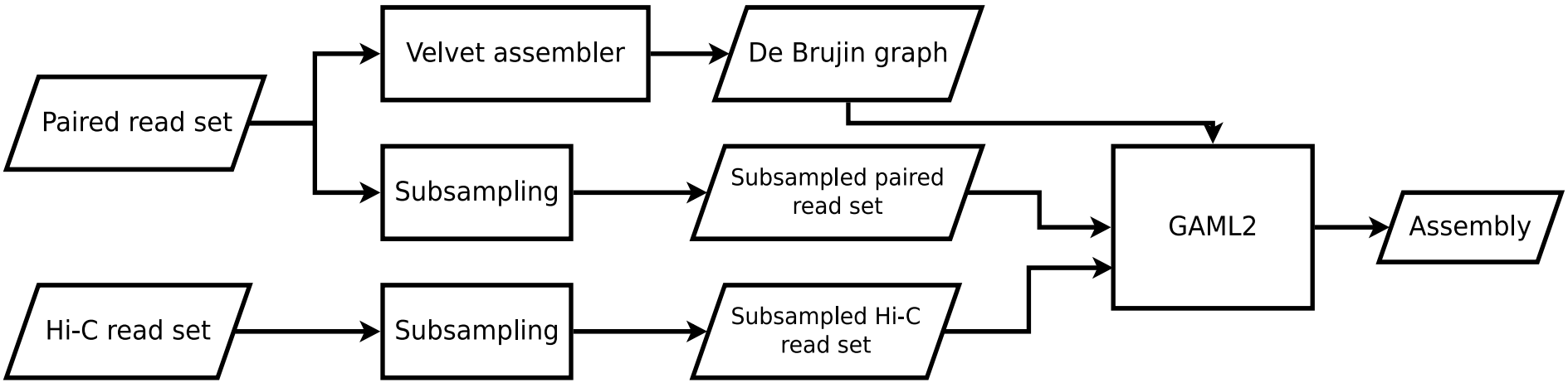
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	MT
1	13267	168	116	418	208	105	334	1382	134	235	154	468	275	215	248	350	0
2	168	49214	253	1490	566	240	1192	544	321	674	462	1263	1140	803	969	1038	6
3	116	253	19722	473	195	161	358	193	148	254	441	428	271	281	373	291	2
4	418	1490	473	90272	1091	431	2107	904	530	2382	747	3103	1805	1445	2289	1925	11
5	208	566	195	1091	34995	224	890	478	350	545	352	1221	778	661	779	947	5
6	105	240	161	431	224	15610	388	214	194	238	154	478	250	492	324	282	1
7	334	1192	358	2107	890	388	65665	656	676	944	589	2008	1342	1084	1428	1479	10
8	1382	544	193	904	478	214	656	35262	265	668	369	1067	721	567	640	1058	2
9	134	321	148	530	350	194	676	265	26480	1261	343	566	381	300	566	398	8
10	235	674	254	2382	545	238	944	668	1261	44904	480	1304	875	699	950	1155	13
11	154	462	441	747	352	154	589	369	343	480	41709	729	500	447	605	490	2
12	468	1263	428	3103	1221	478	2008	1067	566	1304	729	118766	1835	1379	2035	2125	15
13	275	1140	271	1805	778	250	1342	721	381	875	500	1835	56620	861	1273	1470	5
14	215	803	281	1445	661	492	1084	567	300	699	447	1379	861	47334	889	932	5
15	248	969	373	2289	779	324	1428	640	566	950	605	2035	1273	889	66011	1819	11
16	350	1038	291	1925	947	282	1479	1058	398	1155	490	2125	1470	932	1819	58112	6
MT	0	6	2	11	5	1	10	2	8	13	2	15	5	5	11	6	965



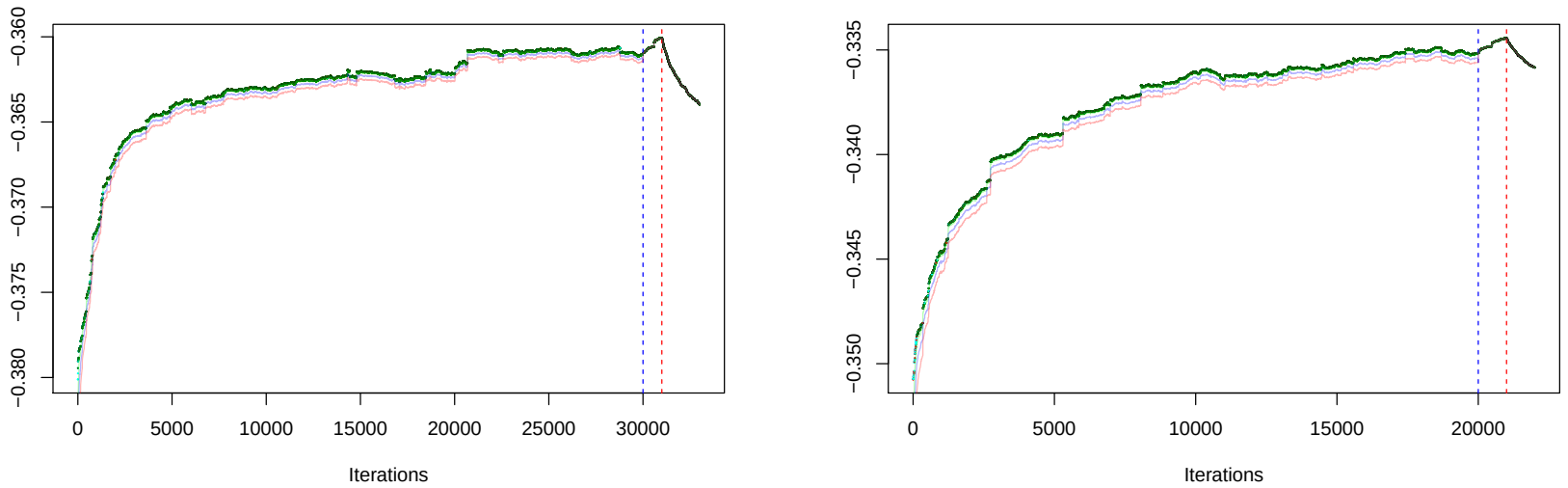
Obr.: Trans interaction rate estimation. Green line shows the prediction of $h_{i,j}$ based on chromosome size, red lines show the boundaries of 70% and 130% relative errors. Overall, 91 out of 120 rates were predicted with relative error in the boundaries (green circle stands for successful estimation and red cross for failure).



Obr.: Insert sizes for cis-reads on every chromosome (y axis is logarithmic)



Obr.: Experiment pipeline



(a) Paired reads only.

(b) Paired and Hi-C reads.

Obr.: Score during the computation of GAML2. Vertical blue line marks the start of hill climbing phase, vertical red line marks the start of cleaning phase. Transparent blue line stands for score decrease, which will be allowed by the temperature with probability 10%. Transparent red line — 1%.

Metric	Velvet output	Only paired	Paired + Hi-C
# contigs ($\geq$ 500 bp)	1450	1098	986
# misassemblies	2	76	102
Total length	11 217 354	11 264 014	11 289 460
N50	14 809	18 720	21 075
L50	241	176	170
NA50	14765	17564	19608
LA50	242	197	181

Tabuľka: Quality metrics for experimental results (by Quast analyser).

- Implementovať ďalšie ťahy do simulovaného žíhania
- Lepšie cacheovanie
- Paralelné vypočty
- Nové typy dát
- Používať hĺbku pokrytia pri zostavovaní
- Lepší PP model pre Hi-C dáta
- Predikcia karyotypu genómu

Ďakujem za pozornosť!