Korekcia dlhých čítaní s vysokým počtom chýb

Marcel Schichman
Vedúci: Mgr. Tomáš Vinař, PhD.
Konzultant: Mgr. Vladimír Boža

Chybové čítania

- Výstup sekvenačneho zariadenia PacBio RS II
- Lacné dlhé čítania (do 14 tis. báz) s vysokou chybovosťou
- Vysoké pokrytie genómu umožňuje chyby opravovať

Problém

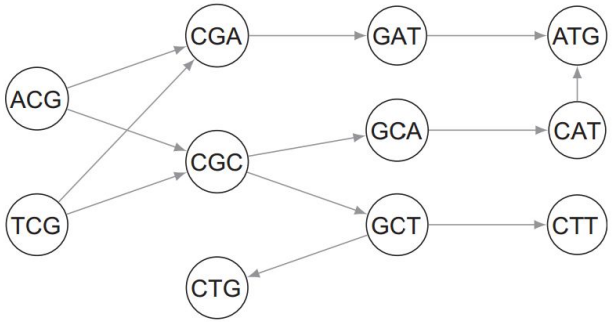
- VSTUP: množina chybových čítaní rovnomerne rozmiestnených po genóme
- VÝSTUP: k vstupným čítaniam priradené opravené čítania

Štandardný algoritmus

- Na opravované čítanie sa zarovnajú prekrývajúce čítania
- Jednoducho povedané – najčastejšia hodnota pre stlpec

LoRDEC

- Hybridná korekcia
- Využíva krátke čítania s nízkou chybovosťou
- De Bruijnov graf



LoRDEC* + LoRMA

- Namiesto namiesto pomocných čítani používa často sa vyskytujúce *k*-tice
- Spúšťa sa opakovane s rastúcim *k*
- Výsledné sekvencie zdokonalené štandardným algoritmom

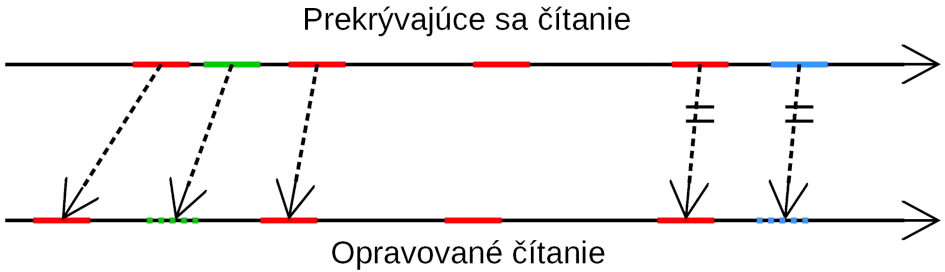
Jadrá korekcie

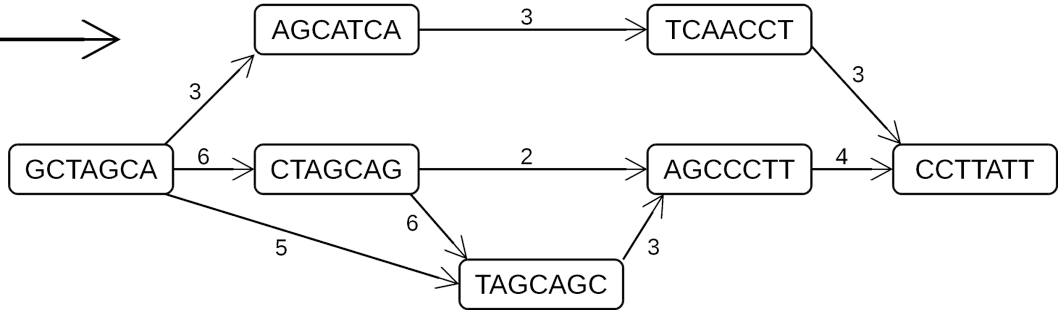
- Skupinu jadier zarovnania medzi dvomi čítaniami s približne rovnakými rozostupmi považujeme za pravdepodobne správne

- Ak dve čítania majú medzi sebou viacero jadier na blízkej diagonále, považujeme ich za prekrývajúce sa

Graf jadier

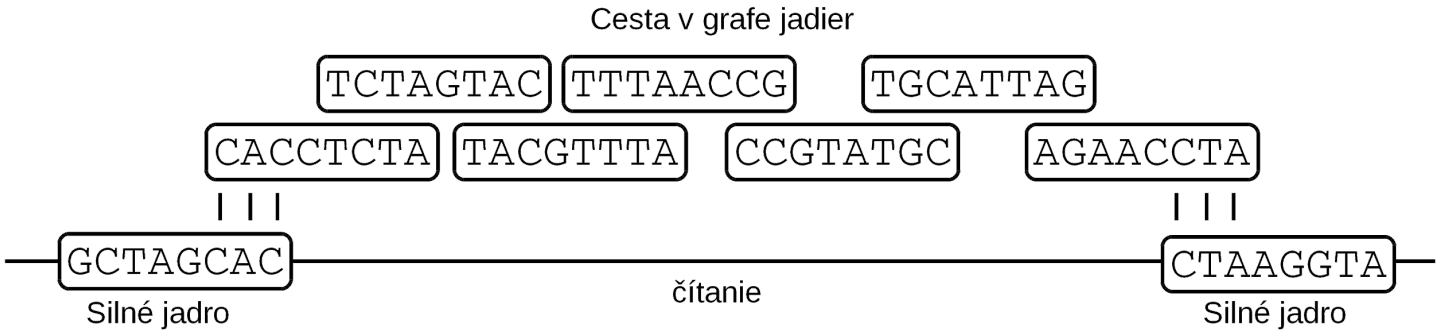
- Zoberie všetky jadrá opravovaného čítania a s ním sa prekrývajúcich čítaní
- Odhadne pozície jadier prekrývajúcich čítaní na opravovanom čítaní
- Z jadier zostrojí De Bruijnov graf





Rekonštrukcia

- Hľadá cestu cez jadrá s minimálnou editačnou vzdialenosťou od pôvodnej sekvencie
- Dijkstrov algoritmus s priestorom pozícií:
  {pozície na čítaní} x {jadrá} x {pozície v jadre}



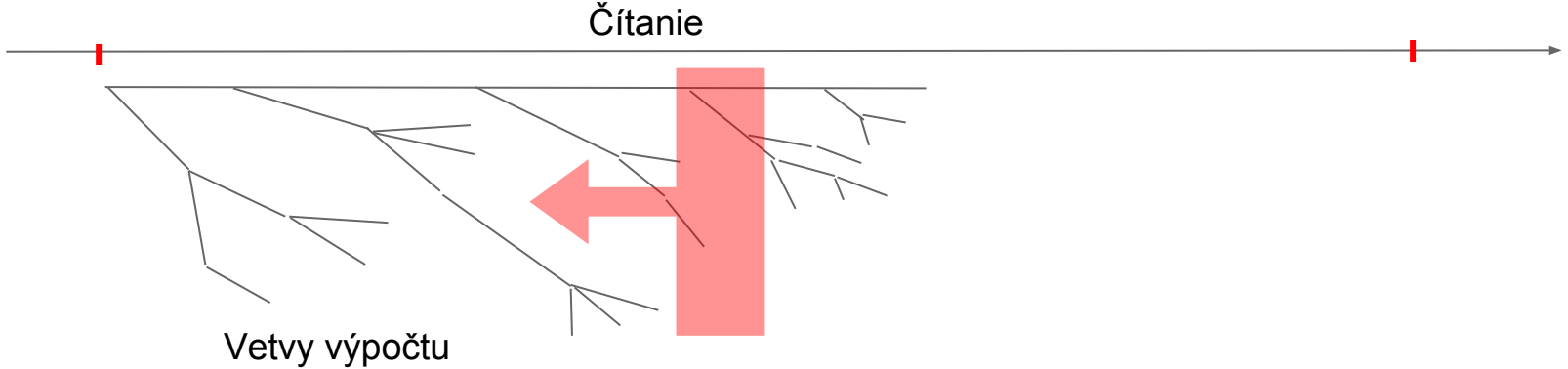
Penalizácia

- Kroky v rámci jadier:
  - Substitúcie
  - Inzercie
  - Delécie

- Pri prechádzaní z jadra na iné:
  - Dĺžka skoku
  - Vzdialenosť od očakávanej pozície

Heuristiky

- Odstraňujeme vetvy výpočtu príliš zaostávajúce za najďalej siahajúcou vetvou



- Od každej vetvy výpočtu vyžadujeme aby mala za posledných *d* báz aspoň *m* zhôd

Výsledky

- Priemerné pokrytie: 50
- Veľkosť vstupných dát: ~ $5*10^6$ bp

	LoRMA	Náš algoritmus	Druhá iterácia
Čas	23,078 s	739,667 s	+287.851 s
Zhoda s genómom (%)	98,72	94,60	95,56
Veľkosť výstupu (%)	5,22	50,13	41,96

Ďakujem za pozornosť