

Nástroj na hodnotenie kvality atomizácie

Cieľ

Tvorba analytického nástroja určeného na hodnotenie kvality genomickej atomizácie, teda rozkladu genómových sekvencií na menšie homogénne úseky - atómové segmenty - na základe lokálnej podobnosti. Atomizácia, realizovaná pomocou algoritmu IMP (implementovaného v nástroji GEESE), umožňuje identifikovať potenciálne homologické regióny medzi sekvenciami. Úlohou nástroja bude overiť, ako presne a konzistentne tento proces prebehol a poskytnúť kvantitatívne aj vizuálne ukazovatele kvality.

Na základe čoho hodnotíme kvalitu atomizácie?

Hodnotiaca matematická schéma pozostáva z troch zložiek:

- 1.) Pokrytie atómami (rozsah genómu prokrytý predikovanými atómami).
- 2.) Kvalita zarovnania členov danej skupiny k ich reprezentantovi.
- 3.) Percento zarovnaní, ktoré sa správne priradujú k danej triede.

Hodnoty jednotlivých metrík sú normalizované na tvar [0,1].

Implementácia matematickej schémy?

1.) Pokrytie atómami

- Pokrytie definujeme ako:

$$C := (\text{počet báz pokrytých atómami}) / (\text{celkový počet báz})$$

2.) Kvalita zarovnania (body 2 a 3 naraz pomocou Minimap2)

- Na vyhodnotenie zarovnania využívame Minimap2:
 - target: representatives.fa
 - query: genomes_file.fa
 - parametre:
 - -x asm20 - optimalizované pre sekvencie s diverzitou do 20%
 - -c - výstup obsahuje CIGAR reťazce
 - -p 0.1 - zobrazí aj slabé alebo čiastočné zarovnania
- Týmto získame PAF súbor s informáciami o všetkých lokálnych výskytoch reprezentantov v genómoch.
- Následne PAF vyfiltrujeme podľa kritérií:
 - minimálna podobnosť 95%,
 - minimálna dĺžka zarovnania 500 bp.
- Zostávajúce zarovnania slúžia ako podklad na vytvorenie gold standard atomizácie. Následne ich prevedieme z formátu PAF do formátu GEESE, aby

bolo možné výslednú atomizáciu priamo porovnať s predikovanými atómami a vizuálne overiť prípadné nepresnosti.

Princíp fungovania nástroja

Nástroj na vstupe dostane:

- FASTA súbor so vstupnými genómovými sekvenciami,
- predikovanú atomizáciu vytvorenú pomocou GEESE.

Cieľom je vypočítať celkové skóre kvality atomizácie, ktoré pozostáva z:

1.) Skóre pokrytia

- Podiel báz v genóme pokrytých predikovanými atómami.

2.) Skóre zarovnania

- Získava sa porovnávaním predikovanej atomizácie s “pravdivou” (gold standard) atomizáciou. Gold standard sa vytvorí nasledovne:

1.) Z každej skupiny predikovanej atomizácie vyberieme reprezentanta.

2.) Reprezentant je atóm, ktorý má najvyššiu kombinovanú podobnosť s ostatnými atómami danej skupiny.

- Podobnosť sa určuje pomocou Mash (sám so sebou porovnanie FASTA súboru skupiny).

3.) Všetky vstupné genómy zarovnáme na reprezentantov príkazom:

```
minimap2 -x asm20 -c -p 0.1 representatives.fa genomes_file.fa -> minimap2_alignments.paf
```

4.) Po prefiltrovaní zarovnaní podľa kvality a dĺžky získame mapovanie reprezentantov na genómy.

5.) Z toho vytvoríme gold standard atomizáciu na porovnanie s predikciou.

Výsledky sú následne vizualizované tak, aby bolo jasne vidieť:

- ako sa jednotlivé reprezentanty zarovnávajú na genóm,
- kde sa v genóme nachádzajú pôvodné predikované atómy.

Metriky

Používateľ si môže zvoliť medzi dvoma typmi metrík presnosti, každá s vlastnou vizualizáciou:

1.) Bázová štatistika (base-level)

- Metriky sa počítajú podľa dĺžky prekryvu v bázach.
- Hodnotia sa intervaly bez ohľadu na percentuálny podiel prekryvu.
- Využívajú sa metriky: TP, FP, FN, Precision, Recall, F1-score.

2.) Intervalová štatistika (interval-level) (predvolené, dôležitejšie)

- Každý atóm sa považuje za jednu jednotku.

- Atóm z predikcie je správny, ak sa prekrýva s atómom z gold standardu aspoň na 80% (overené cez pomer dĺžky prieniku ku dĺžke zjednotenia).
- Metriky sa počítajú na úrovni celých segmentov.
- Využívajú sa metriky: TP, FP, FN, Precision, Recall, F1-score.

Používateľ si môže tiež zvoliť:

- výpočet po jednotlivých skupinách, alebo
- výpočet pre všetky skupiny spolu (predvolené, dôležitejšie).

Poznámky

Plánujeme implementovať aj pomocnú funkciu na generovanie umelých datasetov.

Napríklad náhodným poprehadzovaním 10% atómov medzi klastrami vytvoríme umelo zhoršenú atomizáciu. Nástroj by mal následne priradiť nižšie skóre, čím sa overí jeho korektnosť.

Na čo bude slúžiť hodnotenie atomizácie?

Hodnotiaci nástroj bude slúžiť ako podklad pre vývoj metódy, ktorá umožní identifikovať kandidátske homologické regióny medzi sekvenciami bez nutnosti poskytnúť GEESE-u ako vstup kompletný súbor párových zarovnaní všetkých sekvencií.