Domáca úloha č. 2 pre študentov informatických odborov
1-BIN-301: Metódy v bioinformatike
Termín: utorok 29.11.2022 22:00

Cez Moodle odovzdajte textovú časť v pdf formáte a zdrojový kód vášho programu

Komparatívna genomika. Túto úlohu naprogramujte v niektorom z jazykov Java, C, C++ alebo Python. Prípadné použitie iných jazykov najskôr konzultujte s vyučujúcimi. Použite iba štandardné knižnice vášho jazyka. Dobre čitateľný kód odovzdajte v jednom zip súbore prostredníctvom systému Moodle. Požadované výsledky prípadne diskusiu v otázkach b), d), e), príp. f) odovzdajte na papieri. Dáta k tejto úlohe (súbory `strom.txt`, `exony.txt` a `cftr.txt`) nájdete na webstránke predmetu. V závislosti od efektívnosti vašej implementácie môže výpočet v časti d) pomerne dlho trvať, takže si nechajte dosť času.

a) Implementujte funkciu, ktorá dostane na vstupe zakorenený strom s dĺžkami hrán, rýchlosť mutácií α a jeden stĺpec viacnásobného zarovnania s riadkami zodpovedajúcimi listom stromu a Felsensteinovým algoritmom spočíta pravdepodobnosť, že v tomto strome Jukes-Cantorov model vygeneruje práve tento stĺpec zarovnania. Medzery v zarovnaní (označené pomlčkami) interpretujte ako chýbajúce dáta, t.j. vo Felsensteinovom algoritme nastavte pre takýto list v $A[v,a] = 1$ pre každú bázu a.

   V Jukes-Cantorovom modeli je pravdepodobnosť, že po čase t bude báza taká istá $P(A|A,t) = (1+3e^{-\frac{4}{3}\alpha t})/4$ a že sa zmení na inú konkrétnu bázu (napr. z A na C) $P(C|A,t) = (1-e^{-\frac{4}{3}\alpha t})/4$. V Jukes-Cantorovom modeli má každá báza v ekvilibriu pravdepodobnosť $\frac{1}{4}$, čo využijete ako q_a vygenerovania určitej bázy a v koreni stromu.[1]

b) Akú pravdepodobnosť vráti vaša funkcia pre strom v súbore `strom.txt` a stĺpec zarovnania, v ktorom sú všetky bázy rovnaké? Uveďte výsledok pre $\alpha = 1$ a $\alpha = 0.1$.

   Strom je v súbore zadaný ako postupnosť trojíc (vrchol,rodič,dĺžka hrany). Vrchol Root je koreňom stromu a nemá teda rodiča.

c) Napíšte funkciu, ktorá dostane zarovnanie niekoľkých sekvencií a fylogenetický strom a nájde hodnotu parametra α, pre ktorú má toto zarovnanie najväčšiu pravdepodobnosť (predpokladajte, že jednotlivé stĺpce zarovnania sa vyvíjú nezávisle podľa Jukes-Cantorovho modelu s tou istou hodnotou α). Pri hľadaní optimálnej hodnoty α jednoducho vyskúšajte hodnoty α od 0.1 po 2 s krokom 0.1 a vyberte α^*, pre ktoré dostanete najvyššiu pravdepodobnosť vygenerovania daného zarovnania. (V praxi sa samozrejme používajú zložitejšie a presnejšie optimalizačné metódy.)

   Nakoľko funkciu budete spúšťať na pomerne dlhé sekvencie, pravdepodobnosť vygenerovania celého zarovnania môže byť veľmi malá a môžu nastať numerické problémy s príliš malými číslami. Preto pravdepodobnosť vygenerovania každého stĺpca zlogaritmujte a pracujte ďalej s pravdepodobnosťami v logaritmickom tvare (násobenie pravdepodobností nahraďte sčítaním ich logaritmov).

d) Napíšte program, ktorý načíta dlhé viacnásobné zarovnanie, rozdelí ho na neprekrývajúce sa okná dĺžky $w = 100$ a pre každé okno pomocou funkcie z časti c) spočíta optimálnu hodnotu α^*.

   Spustite váš program na zarovnanie v súbore `cftr.txt`. V tomto súbore je zarovnanie časti CFTR regiónu z 14 cicavcov. Meno organizmu je dané na riadku začínajúcom znakom >, potom nasleduje séria riadkov so sekvenciou, do ktorej sú povkladané pomlčky. Okrem báz A,C,G,T sa môže vyskytovať aj znak N, označujúci neznámu bázu. Takéto znaky spracujte rovnako ako pomlčky.

   V domácej úlohe uveďte nájdenú hodnotu α^* pre prvých 10 okien, (prvé okno pokrýva stĺpce $1,\ldots,100$, druhé okno stĺpce $101,\ldots,200$, atď).

e) Súbor `exony.txt` obsahuje súradnice exónov kódujúcich proteíny v ľudskej sekvencii v našom zarovnaní (pričom bázy číslujeme od 1 a **nepočítame pomlčky**). Rozdeľte okná z úlohy d) na dve skupiny: tie, ktoré sa prekrývajú s exónom (hoci aj jedinou bázou) a tie, ktoré sa neprekrývajú.

   Pre každú skupinu okien zostavte histogram optimálnej hodnoty α^* a vykreslite ho (napr. v Exceli alebo nejakom štatistickom či vykresľovacom softvéri). Okomentujte výsledky: aké vidíte rozdiely medzi histogramami, sú také, ako ste čakali, čo ich asi spôsobuje?

[1] Pozor, vo veľa programovacích jazykoch $4/3 = 1$.

f) Za max. 3 bonusové body môžete analyzovať poskytnuté dáta ešte nejakým iným spôsobom. Napríklad môžete skúsiť menší počet organizmov, meniť veľkosť okna, pozrieť sa na vlastnosti okien s obzvlášť veľkou alebo malou hodnotou α a pod. Popíšte váš postup, uveďte výsledky a okomentujte ich.