Domáca úloha č. 1 pre študentov prírodovedných zameraní
2-AIN-501: Metódy v bioinformatike
Termín: streda 12.11.2025 22:00

Odovzdajte pdf cez Moodle

1. Skórovacie schémy. Na stránke `https://colab.research.google.com/github/fmfi-compbio/mbi//blob/master/notebooks/cb-aln1.ipynb` program v Pythone, ktorý sa dá spustiť v online prostredí Google Colab bez inštalácie softvéru na váš počítač.[1] Tento program spočíta lokálne zarovnanie pre dve krátke DNA sekvencie a zvolené skórovanie. Ak existuje viac lokálnych zarovnaní s najvyšším skóre, vypíše prvých 10 z nich.

V políčku pod nadpisom `Input and settings` môžete editovať časť programu, kde nastavíte sekvencie, ktoré sa majú zarovnať a skórovaciu schému. Potom spustite program gombíkom `Run all` a pozrite si výsledky pod nadpisom `Found alignments`.

Skúste zarovnať nasledujúce dve sekvencie s rôznymi skórovacími schémami a nájdite odpovede na otázky nižšie. Vždy aj vysvetlite, prečo by sme intuitívne takúto odpoveď očakávali.
Sekvencia X: `AGGACTGTGTTCACGGCCTACACATTA`
Sekvencia Y: `ATTTTATACAGTCGTCCCCGTCTCTTGGGAGG`

a) Uvažujme skórovacie schémy, v ktorých je skóre zhody (match) 10 a skóre nezhody (mismatch) a medzery (gap) sú nejaké celé záporné čísla. Pri akých typoch skórovania dostávate ako výsledok veľmi krátke lokálne zarovnania? Prečo je to tak? Uveďte príklad s najkratším zarovnaním, aké sa vám podarilo nájsť a pri akej skórovacej schéme tento príklad nastal.

b) Uvažujte opäť skórovanie, v ktorom je skóre zhody 10 a skóre nezhody a skóre medzery sú nejaké celé záporné čísla. Pri akých typoch skórovania dostávate ako výsledok dlhé lokálne zarovnania, ktoré obsahujú veľkú časť vstupných sekvencií? Prečo je to tak? Uveďte príklad zarovnania, ktoré obsahuje čo najviac zo vstupných sekvencií a pri akej skórovacej schéme tento príklad nastal.

c) Pridajte do jednej zo sekvencií niekoľko (čo najmenší počet) nových nukleotidov tak, aby sa pri skórovaní z časti b) zarovnali obidve sekvencie celé. Vysvetlite, ako ste pri riešení úlohy postupovali.

2. Zostavovanie kontigov z BACov. V minulosti sa pri sekvenovaní používala metóda BAC-by-BAC, v ktorej sa kusy sekvencie dĺžky 100-200kbp naklonovali ako bacterial artificial chromosome (BAC) a každý sa zvlášť sekvenoval a zostavoval. V tejto úlohe máme niekoľko BACov, z ktorých je potrebné zostaviť jeden kontig. Vo všetkých častiach popíšte, ako ste dospeli k odpovediam.

a) Stiahnite z Genbanku sekvencie AC191982, AC191985, AC191986. Z akého sú organizmu?

b) Pomocou nejakého programu na hľadanie lokálnych zarovnaní nájdite medzi kontigmi prekryvy a určite, v akom poradí a v akej orientácii sa majú BACy nachádzať vo výslednom kontigu.

   Napíšte návod ako zostaviť výsledný kontig (ktoré časti z ktorých BACov by ste použili vo výsledku, vrátane súradníc). Aký dlhý bude výsledný kontig? Ukážte, ako ste k riešeniu dospeli.

   Pomôcka: Môžete použiť web server programu yass `https://bioinfo.univ-lille.fr/yass/yass.php`, ktorý umožňuje zarovnávať páry sekvencií a vypisovať výsledok v rôznych formátoch, ako napríklad dotplot (na vizuálne posúdenie podobnosti) alebo tabuľka (na určenie súradníc zarovnaných úsekov).

c) Zistite, na ktorom chromozóme a na akých súradniciach sa tento kontig nachádza v príslušnom organizme. Na tento účel použite aplikáciu BLAT na UCSC genome browseri `https://genome-euro.ucsc.edu/`. Stačí pomocou BLATu mapovať úplný začiatok a úplný koniec kontigu.

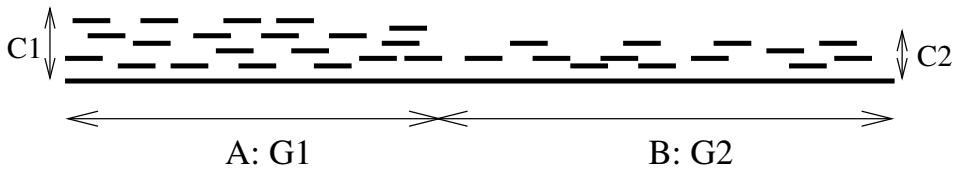
   Skúste toto mapovanie v najstaršej a najnovšej dostupnej verzii príslušného genómu (z 2006 a 2019). Ako sa dĺžka tohto úseku zmenila medzi verziami a ako sa porovná s dĺžkou vášho kontigu?

[1] Spúšťanie programu v Colabe vyžaduje prihlásenie do vášho konta na Gmail. Ak také konto nemáte alebo ho nechcete použiť, môžete skúsiť spustiť program na stránke `https://jupyter.org/try-jupyter/lab/` V menu zvoľte `File/Open from URL` a zadajte `https://fmfi-compbio.github.io/mbi/notebooks/cb-aln1.ipynb`

3. Pokrytie genómu čítaniami Pri plánovaní sekvenovania určitého genómu sú dôležité nasledujúce parametre: dĺžka genómu G, dĺžka čítania L a počet čítaní N. Súčet dĺžok čítaní bude teda NL a jedna báza genómu bude v priemere pokrytá NL/G čítaniami. Túto hodnotu označíme C, pokrytie genómu. Keďže však pri sekvenovaní sú jednotlivé čítania na genóme rozmiestnené náhodne, na niektorých miestach ich bude viac a na iných menej a môže sa dokonca stať, že niektoré bázy nebudú pokryté žiadnym čítaním. Pomocou pomerne jednoduchých vzťahov sa dá odvodiť, koľko pri určitých parametroch očakávame v genóme takýchto nepokrytých báz. Tieto vzorce sú však odvodené za idealistických podmienok, kde každá báza v genóme má rovnakú šancu byť začiatkom čítania. Tento model rozmiestnenia čítaní nazveme M1.

V reálnych dátach však často dochádza k výchylkám, kde oblasti s extrémne vysokým alebo nízkym obsahom GC majú menšie pokrytie čítaniami než oblasti s obsahom GC blízkym 50%. V modeli M2 budeme uvažovať genóm rozdelený na časti A a B, pričom časť A má veľkosť G_1 a priemerné pokrytie C_1 a časť B má veľkosť G_2 a priemerné pokrytie C_2. Pozrieme sa, aké majú rozdiely v priemernom pokrytí efekt na počet báz, ktoré nie sú vôbec pokryté žiadnym čítaním.

Na stránke `https://compbio.fmph.uniba.sk/vyuka/mbi-data/du1/du1-b.xls` nájdete súbor pre Excel (alebo tabuľkový procesor v LibreOffice a pod.), v ktorom môžete meniť parametre G, L a C, ako aj C_1 a G_1 a počítať, koľko nepokrytých báz očakávame v genóme v modeloch M1 a M2. Veľkosť oblasti B sa vždy dopočíta tak, aby celková veľkosť bola rovnaká ako v modeli M1 a podobne pokrytie oblasti B sa dopočíta tak, aby celkový počet čítaní bol rovnaký ako v M1. Veľkosť genómu G je nastavená na milión báz a dĺžka čítania L na 100.



a) Uveďte v tabuľke, koľko báz s pokrytím 0 očakávame v celom genóme pri priemernom pokrytí $C = 1, 5, 10, 15$ v modeli M1. Aký trend pozorujete? Prečo to tak asi je?

b) Nastavte celkové pokrytie na $C = 10$, veľkosť oblasti A nastavte na $G_1 = 500\,000$, t.j. $G/2$ a skúšajte pokrytie časti A s hodnotami $C_1 = 11, 12, 13, 14, 15$. Uveďte v tabuľke, koľko nepokrytých báz očakávame v celom genóme v modeli M2 pre tieto hodnoty C_1. Aký trend pozorujete? Máme viac alebo menej nepokrytých báz ako v modeli M1? Prečo to tak asi je?

c) Nastavte celkové pokrytie $C = 10$, pokrytie oblasti A na $C_1 = 12$ a skúšajte nastaviť veľkosť časti A s hodnotami $G_1 = 300\,000, 400\,000, 500\,000, 600\,000, 700\,000$. Uveďte v tabuľke, koľko nepokrytých báz očakávame v celom genóme v modeli M2 pre tieto hodnoty G_1. Aký trend pozorujete? Vidíme pri nerovnomernom rozdelení genómu na časti A a B väčší alebo menší počet nepokrytých báz? Prečo to tak asi je?

V jednotlivých častiach skúste neformálne zdôvodniť, prečo sú pozorované trendy intuívne očakávateľné. Nie je potrebné v tomto zdôvodnení uvádzať žiadne vzorce alebo výpočty.

Poznámka: pre kontrolu uvádzame údaje v jednotlivých častiach pre iné nastavenia než požadované. Ak v časti a) zvolíme $C = 8$, dostaneme výsledok 335,33. V časti b) ak $C_1 = 16$, dostávame 9154,21. Ak v časti c) zvolíme $G_1 = 800\,000$, dostávame 27058,43.