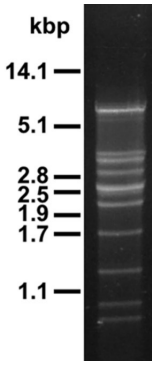
Domáca úloha č. 1 pre študentov informatických odborov
1-BIN-301: Metódy v bioinformatike
Termín: utorok 8.11.2022 22:00

Odovzdajte pdf cez Moodle

Príklad 1. Restrikčné enzýmy. Restrikčné enzýmy sú špeciálne enzýmy, ktoré sekajú DNA na každom výskyte určitej krátkej sekvencie. Napríklad enzým SmaI seká DNA v strede každého výskytu reťazca CCCGGG (čiže medzi C a G). Takéto pozície v sekvencii DNA budeme nazývať restrikčné miesta daného enzýmu. Predstavme si teraz, že máme v skúmavke veľmi veľa kópií tej istej DNA. Ak by sme na túto DNA nechali restrikčný enzým pôsobiť dostatočne dlho, prerezal by DNA na všetkých restrikčných miestach. Ak ho však necháme pôsobiť kratšie, niektoré restrikčné miesta zostanú neprerezané a dostaneme tak veľa fragmentov DNA rôznej dĺžky, pričom každý fragment začína aj končí buď na niektorom restrikčnom mieste alebo na jednom z koncov pôvodnej sekvencie DNA. Ak sme začali s veľmi veľa kópiami pôvodnej DNA, vo výslednej zmesi môžeme mať dokonca vo veľa kópiách zastúpené všetky fragmenty, ktoré predstavujú niektorý úsek medzi dvoma (nie nutne susednými) restrikčnými miestami alebo medzi restrikčným miestom a koncom chromozómu.

Takto získanú zmes fragmentov rôznej dĺžky môžeme analyzovať gélovou elektroforézou, pomocou ktorej získame odhady dĺžok fragmentov zastúpených v skúmavke. Nevieme však povedať, koľko kópií z ktorej dĺžky máme. Predstavme si teraz, že sekvenciu pôvodnej DNA nepoznáme a len na základe nameraných dĺžok fragmentov sa snažíme zistiť, koľko restrikčných miest sa v tejto DNA nachádza a v akých sú vzdialenostiach od začiatku sekvencie.



Obr. 1: Ukážka výsledku gélovej elektroforézy. Biely prúžok predstavuje skupinu fragmentov DNA približne rovnakej dĺžky. Mierka vľavo (v tisícoch bázových párov) bola získaná elektroforézou vzorky so známymi dĺžkami fragmentov (zdroj: Valach a kol. 2011).

Úloha:

a) Sformulujte tento problém ako informatický problém s jasne definovaným vstupom a výstupom. V tejto časti predpokladajte idealizovanú situáciu, v ktorej nevzniknú žiadne experimentálne chyby a dĺžky všetkých fragmentov vieme odmerať úplne presne.

   V definícii problému používajte informatické alebo matematické pojmy, ako napríklad čísla, množiny alebo postupnosti čísel, grafy, reťazce a pod., nie biologické pojmy. Potom však vysvetlite, ako vaša definícia informatického problému súvisí s popísaným biologickým problémom.

   Poznámky: Úloha môže mať viacero dobrých formulácií. Vôbec sa nemusíte zamýšľať nad tým, ako by sa vami definovaný problém dal riešiť.

b) Ukážte ilustratívny príklad vstupu a výstupu pre váš problém.

c) Zmeňte vašu formuláciu tak, aby brala do úvahy aspoň niektoré možné zdroje chýb, napríklad nepresné merania dĺžok fragmentov, chýbajúce fragmenty, ktoré by teoreticky mali vzniknúť alebo naopak fragmenty, ktoré nepochádzajú z cieľovej DNA, ale vznikli napr. kontamináciou vzorky. Opäť uveďte informatickú formuláciu problému a vysvetlite, ktoré z týchto zdrojov chýb berie do úvahy.

2. Homopolymérové zarovnanie. Homopolymér v DNA je postupnosť niekoľkých rovnakých báz, napríklad AAAAA alebo GGG. Vašou úlohou je nájsť algoritmus na výpočet optimálneho globálneho zarovnania so zvláštou skórovacou schémou, v ktorej vystupujú homopolyméry. Konkrétne vychádzame z predstavy, že počas evolúcie sa môžu do sekvencií pridávať a strácať homopolyméry, pričom inzercia alebo delécia homopolyméru sa udeje v jednom kroku. V našom zarovnaní zhoda bude mať skóre 0 a nezhoda bude zakázaná (bude mať teda skóre $-\infty$). Súvislá oblasť stĺpcov s pomlčkami medzi dvoma zhodami sa pri skórovaní rozdelí na niekoľko častí, každá predstavujúca inzerciu alebo deléciu homopolyméru v jednej sekvencii a každá takáto časť dostane skóre -1.

Napríklad zarovnanie v príklade nižšie vľavo bude mať skóre -5, pričom z prvej sekvencie zmažeme homopolyméry A, CC, G, GG a vložíme homopolymér GGG. Samozrejme, toto zarovnanie nie je optimálne, stačilo by zmazať homopolyméry A a CC a G-čka zarovnať k sebe, čím by sme dostali skóre -2, ako vidíme v zarovnaní napravo.

```
AACCG---GGGCT          AACCGGGGCT
A----GGG--GCT          A---GGGGCT
```

a) Navrhnite algoritmus na nájdenie optimálneho zarovnania s týmto skórovaním. Algoritmus podrobne popíšte, prípadne uveďte pseudokód, odhadnite jeho časovú zložitosť a stručne ho zdôvodnite. Pokúste sa nájsť algoritmus so zložitosťou $O(n^2)$, za malú bodovú stratu však môžete napísať aj horší. Dôležité je, aby váš algoritmus správne fungoval pre všetky vstupy. Pomôcka: skórovanie sa podobá na afínne skóre medzier.

b) Homopolymérovou dĺžkou sekvencie nazvime najmenší počet homopolymérov, z ktorých sa dá vyskladať konkatenovaním. Napr. prvá sekvencia v zarovnaní vyššie má homopolymérovú dĺžku 5, lebo sa dá vyskladať z AA, CC, GGGG, C a T. Ak váš algoritmus spustíte na sekvenciách s homopolymérovou dĺžkou n a m, aká môže byť najhoršia cena nájdeného zarovnania (ako funkcia n a m)? Nájdite aj všeobecný prípad, ako by mohli vyzerať vstupné sekvencie, ktoré majú zarovnania s takýmto skóre pre všeobecné hodnoty n a m.

c) Prof. Premúdrelý skúma dve sekvencie S a T a chcel by zistiť, koľko najmenej homopolymérových inzercií a delécii by muselo nastať v evolúcii, aby sa S zmenila na T. Tento počet označme e a optimálne skóre zarovnania týchto dvoch sekvencií v našej skórovacej schéme označme s. Profesor tvrdí, že e a $|s|$ sa vždy rovnajú.

Vysvetlite, prečo profesor nemá pravdu. Platí však jedna z nerovností $e \leq |s|$ resp. $e \geq |s|$. Ktorá a prečo?

Pomôcka: $S = ACGCA$, $T = AA$

d) Bonusová podúloha: Nájdite polynomiálny algoritmus, ktorý pre danú sekvenciu S spočíta minimálny počet homopolymérových delécií, ktorými môžeme z S dostať prázdny reťazec. Algoritmus podrobne popíšte, stručne zdôvodnite a odhadnite zložitosť.