Vyhledávání tRNA genů

Petr Dvořáček – xdvora0n@stud.fit.vutbr.cz

Tento dokument pojednává o jednoduchém vyhledávání a případné filtraci tRNA (transportní ribonukleonové kyseliny) genů v genomu DNA (2-deoxyribonuklenová kyselina). Geny tRNA slouží především pro transport aminokyselin do ribozomů. Tam díky nim a hlavně díky mRNA vznikají proteiny.

1 Úvod do problému

Libovolné geny, ať už mluvíme o mRNA, či tRNA, jsou po celé DNA roztroušeny na předem neurčených úsecích. Proto je vhodné tyto úseky odhadnout. V případě tRNA můžeme využít znalostí jako jsou velikosti délek jednotlivých úseků, pravděpodobnosti bází na daných pozicích a celkový tvar tRNA. O délkách úseků a pravděpodobnosti pojednává tabulka uvedená v zadání. Kde například vidíme, že na 14 pozici je nejpravděpodobnější bází adenin nebo guanin.

Tvar tRNA, který je ukázán na obrázku 1c, můžeme využít v pozdější filtraci genů. Ta probíhá tak, že se hledá komplement k bázi na dané pozici. Podle přednášek z bioinformatiky je Watson-Cirkovo párování bází (A–U, C–G) obohaceno o pár U–G, jenž se občas v RNA vyskytuje. Ovšem nemusí to platit furt, jak zobrazeno na obrázku 1c. Můžeme zde vidět, že došlo k páru A–G (znázorněný oranžově). To je nejspíše zapříčiněno tím, že se tato vazba nachází poblíž D smyčky. Takže buď nemusí vůbec existovat, nebo jeden vodíkový můstek není využit.

Nalezené tRNA geny mohou být porovnány s referenčními geny a to podle jejich pozice výskytu. Referenční tRNA geny byly nalezeny jinou metodou, než byla použita v tomto projektu.

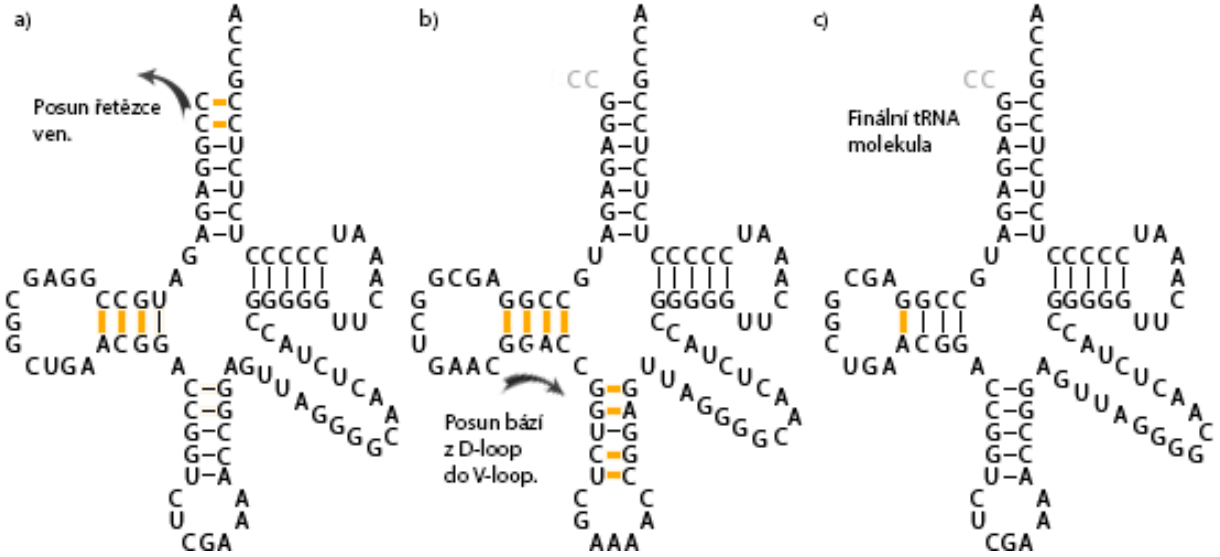
2 Implementace aplikace

V zadání bylo specifikováno, že má být použit FASTA formát. Tento formát obsahuje právě jeden řetězec DNA 5–3. Druhý řetězec DNA 3–5 musíme proto vytvořit z původního řetězce. Tvorba probíhá pomocí komplementarity bází a reverzaci řetězce. Ovšem při vyhledávání genů tento krok nehraje příliš významnou roli. Může být totiž využíto komplementarity a reverzaci hledaného genu. Takže na stejném DNA pouze zdvojnásobíme vyhledávání a nebude zapotřebí vytvoření reverzního řetězce DNA 3–5. Pomocí výše uvedené znalosti a tabulky uvedené v zadání byl vytvořen regulární výraz, který naleznete v souboru `get_tRNA.py`.

Tento regulární výraz, ale v Pythonu vyhledává hladově. Tím je myšleno, že nalezne největší možný podřetězec. Toto lze eliminovat následnou filtrací, která pracuje na základě naivního přístupu na predikci RNA struktury. Bylo využito postupného zmenšování nalezeného genu a porovnání bází ve *štopkách*. Na obrázku 1 je znázorněno, jak toto zmenšování probíhá:

1. Nejdříve se zkontroluje párování v A-stem.
2. Neproběhla-li kontrola v pořádku odřeže se první báze z řetězce a pokračuje se bodem 1. Viz obrázek 1a.
3. Zkontroluje se komplementarita D-stem a C-stem.
4. Neproběhla-li tato kontrola v pořádku, D-loop se zmenší a řetězec se jakoby posouvá k V-loopu, který se zvětšuje. Viz obrázek 1b.
5. Proběhlo-li i toto v pořádku pak se jedná o správný gen tRNA. Jinak se pokračuje bodem 1.

Rovněž bylo nutné zavést určitou toleranci k párování bází, jak bylo řečeno v sekci 1. Tolerance byla nastavena na, alspoň dvě chyby či méně.

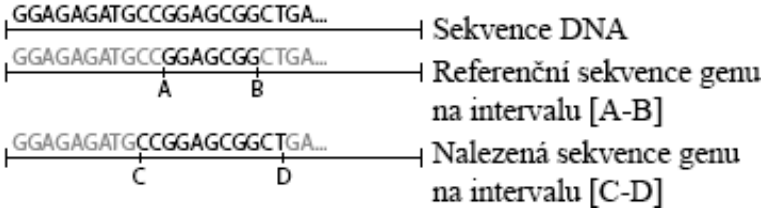


Obrázek 1: Filtrace genů pomocí komparativního přístupu.

Pro porovnání nalezených genů s referenčními bylo využito odchylek od původních startovních pozic genů a počtu jejich bází (koncových pozic genů). Porovnání pak odpovídá následujícímu vzorci

$$Error = \frac{abs(A - C) + abs(B - D)}{pocet\ bazi\ v\ referencnim\ tRNA}$$

, kde proměnná A odpovídá referenční počáteční pozici genu a proměnná B získané. Proměnná C odpovídá referenční koncové pozici a proměnná D získané. Navíc tato situace je znázorněna na obrázku 2.



Obrázek 2: Výpočet překryvu.

3 Závěrečné experimentální vyhodnocení

V zadaném genomu E. coli bylo nalezeno 91 genů tRNA. Z nichž 7 výskytů bylo falešných a 6 referenčních genů se nenašlo. Filtr falešné výskyty vyfiltroval a s ním jeden správně nalezený gen. Původní chybovost nalezených genů, která činila kolem 5.5%, byla pomocí filtrace redukována na téměř nulovou.

Jeden z nenalezených referenčích genů má skore nad 60. Jedná se o tRNA24 se stop kodonem bez konce xCCA. Tudíž nenese žádnou aminokyselinu a regulární výraz jej nenašel. Zbytek nebyl nalezen, neboť právě obsahuje mutace na těch místech, která z pohledu regulárního výrazu zůstanou neměnná. Tomu by se dalo předcházet přidáním nějaké tolerance do regulárního výrazu (či pravděpodobnosti). Dále by bylo vhodné využit znalostí promotorů v DNA (CAAT, TATA boxy). Další možné pokračování bych viděl ve vyhledávání mRNA genů u Eukarotyt, jenž úzce souvisí s tRNA.