Vyhledávání tRNA genů

Milan Munzar
xmunza00@stud.fit.vutbr.cz

1 Úvod

Tento text popisuje návrh a implementaci algoritmu na vyhledávání tRNA genů uvnitř genomu. Program je napsán v jazyce Python 2.7 a tvoří ho dva skripty. Skript `get_tRNA.py` nalezne tRNA geny a skript `compare_tRNA.py` porovná nalezené geny s geny z Genomic tRNA Database(GTD). Program je testován na záznamech k bakterii Escherichia coli CFT073.

2 Implementace vyhledávání tRNA genů

Pro hledání tRNA genů jsem využil konzervanosti hledaných oblastí [1]. Tato vlastnost dovoluje sestavit regulární výraz a aplikovat jej na celý genom v sense i anti-sense směru (alg: 1). Hledání tRNA genů v anti-sense směru vyžaduje vytvoření komplementárního řetězce a jeho otočení. Výhoda této implementace je, že algoritmus pracuje s lineární časovou složitosti.

Algoritmus je implementován v souboru `get_tRNA.py` a využívá možností knihovny `re`. Vstupem programu je soubor se sekvencí genomu ve formátu fasta. Výstupem na standartní výstup je množina záznamů ve formátu multifasta, odpovídající nalezeným tRNA genům.

Algoritmus 1 Použitý regulární výraz na hledání tRNA genů

```
1  pattern = """
2  [AUCG]{13}                                  # 1 - 13
3  A                                           # 14
4  (A|G)                                       # 15
5  [AUCG]{1,3}                                 # 16 - 17A
6  G                                           # 18
7  [AUCG]{11,14}                               # 19 - 31
8  (A|C|U)                                     # 32
9  U                                           # 33
10 (?P<Anticodon>[AUCG]{3})                    # 34 - 36 Anticodon
11 (A|G)                                       # 37
12 [AUCG]{11,31}                               # 38 - 52
13 GUUC                                        # 53 - 56
14 (G|A)                                       # 57
15 A                                           # 58
16 [AUCG]                                      # 59
17 (U|C)                                       # 60
18 C                                           # 61
19 [AUCG]{12}                                  # 62 - 73
20 CCA                                         # 74 - 76
21 """
```

3 Výsledky porovnání s GTD

Požadavkem pro úspěšné vyřešení problému je dosáhnutí alespoň 80% překryvu se záznamy GTD kódujících 20 standartních aminokyselin se skóre alespoň 60. Dalším požadavkem je, že program by neměl vypisovat více než 20% falešných výskytů (záznamy jež nejsou v GTD). Oba tyto požadavky jsou splněny jak je vidět na výstupu programu:

```
$ ./compare_tRNA.py seq/found_tRNA.multifasta seq/known_tRNA.multifasta
tRNA_6|AspCUA|630445|76|+
tRNA_14|AlaCGA|1027722|87|+
tRNA_75|LeuAAC|1812665|90|-
tRNA_76|SerUCA|1742003|83|-
tRNA_77|ArgGCG|1697066|89|-
tRNA_90|SerAGC|408294|93|-

Escherichia_coli_CFT073_chr (236813-236889)  Glu (TTC) 77 bp  Sc: 49.26
Escherichia_coli_CFT073_chr (434982-435062)  Undet (???) 81 bp  Sc: 28.45
Escherichia_coli_CFT073_chr (1211152-1211228)  Arg (TCT) 77 bp  Sc: 59.54
Escherichia_coli_CFT073_chr (1342387-1342463)  Arg (TCG) 77 bp  Sc: 48.60
Escherichia_coli_CFT073_chr (4274308-4274398)  SeC(p) (TCA) 91 bp  Sc: 75.19

neporovnaných get_tRNA.py: 6/90
neporovnaných GTD: 5/89
```

Program produkuje 7% falešných výskytů a nenalezne 6% záznamů v GTD. Všechny nenalezené záznamy obsahují bud' nestandartní aminokyselinu nebo mají skóre nižší jak 60.

4 Závěr

Podařilo se mi naimplementovat algoritmus pro hledání tRNA genů. Program splňuje požadavky ze zadání projektu. Výskyt falešných položek a nenalezení všech záznamů z GTD je způsobeno jednoduchostí použité metody. Falešné výskyty by se mohli redukovat přidáním dodatečné filtrace nalezených záznamů. Pro nalezení většího počtu záznamů z GTD je možno vylepšit stávající algoritmus, případně použít složitější metodu hledání tRNA genů.

Reference

[1] SAKS, Margaret E., CONERY, John S. *Anticodon-dependent conservation of bacterial tRNA gene sequences.* RNA 13.5 (2007): 651-660.