Petr Jestřábek

IV114 Projekt z Bioinformatiky – zadání projektu

Cílem projektu je vytvořit program, který vyhledá s pomocí databáze UniProt proteiny s podobnými vlastnostmi jako mají ty, jež jsou zadány na vstupu. Součástí výstupu bude celková statistika.

Vstup:

Vstupem bude seznam sekvencí proteinů z databáze PDB ve tvaru ID POS1 POS2 SEQ1 SEQ2, kde:

- ID je identifikací proteinu,
- POS1 resp. POS2 je počáteční pozice první resp. druhé sekvence v proteinu,
- SEQ1 resp. SEQ2 je první resp. druhá sekvence, kde jsou jednotlivé aminokyseliny znázorněny příslušnými písmeny (A – alanin, C – cystein, G – glycin, ...).

Výstup:

Výstupem programu bude seznam proteinů obsahujících shodnou či podobnou (podle nastavených parametrů) dvojici sekvencí jako protein na vstupu. Tyto budou vyhledány v lokálním souboru databáze UniProt (ve formátu FASTA). Seznam bude ve tvaru ID POS1 POS2 SEQ1 SEQ2, kde:

- ID je identifikací proteinu,
- POS1 resp. POS2 je počáteční pozice první resp. druhé sekvence v nalezeném proteinu,
- SEQ1 resp. SEQ2 je první resp. druhá sekvence, kde jsou jednotlivé aminokyseliny znázorněny příslušnými písmeny (A – alanin, C – cystein, G – glycin, ...)

Dále bude k výstupu připojena celková statistika:

- Určení, nakolik se liší vzdálenosti mezi sekvencemi,
- typické skóre podobnosti,
- typická vzdálenost mezi sekvencemi.

Přepínače:

Program bude možné spustit s následujícími volbami:

- `-s PERCENT`

  Volitelné skóre podobnosti sekvencí v procentech (0-100%). Implicitně je s = 100.

- `-d REAL (d je z intervalu <0, 1>)`

  Vzdálenost, ve které se mohou vyskytovat zadané sekvence. Pro d=1 je vzdálenost určena přímo vstupem, pro d=0.5 budou sekvence hledány ve vzdálenosti poloviční až dvojnásobné, pro d=0 budou hledány v jakékoliv vzdálenosti. Implicitně je d = 1.

- `-r`

  Sekvence se mohou v proteinu vyskytovat i v opačném pořadí (SEQ2 SEQ1) než jak jsou zadány na vstupu. Implicitně je volba vypnuta.