Sequence comparison

projekt pro jednoduché porovnání dvou biologických sekvencí

Části projektu

- stažení a zpracování sekvencí z existujících online databází
- porovnání daných dvou sekvencí pomocí různých algoritmů a parametrů
- vygenerování PDF reportu s výsledky porovnání

Stažení sekvencí

- program umí stahovat sekvence z těchto existujících databází
  - uniprot (Universal Protein Resources)
  - ncbi (National Center for Biotechnology information)
  - ebi (European bioinformatics institute)
  - ddb (DNA data bank of Japan)

- podporované formáty - fasta

Stažení sekvencí

- poté co jsou sekvence staženy, snaží se je program normalizovat, aby se daly lépe porovnat
- tato normalizace je přidána jako poslední záchrana a nepředpokládá se, že by byla reálně často využívána

Porovnání sekvencí

- jednoduchý histogram znaků
- algoritmy pro globální a lokální alignment, použití match/mismatch parametrů, gap opening/extension parametrů a také existujících biologických matic
- podporované matice: PAM40, PAM80, PAM120, PAM250, BLOSUM62

- parametry programu se dají řetězit - lze tedy stejné sekvence porovnat více možnostmi při jednom volání skriptu, pro lepší analýzu

Generování PDF reportu

- použití ReportLab knihovny
- není lehké se do ní dostat (userguide má “jen” 125 stran)
- je to ovšem velmi efektivní knihovna, s mnoha možnostmi (grafy, tabulky, grafika, …)

Shrnutí

- fungující stahování sekvencí z online databází
- implementovány porovnávací algoritmy s různými možnostmi parametrů
- úspěšné použití robustní PDF knihovny

Všechen kód k projektu je dostupný z
https://github.com/gyfis/sequence-comparison

Díky za pozornost!