Daniel Vodák

IV110 Projekt z bioinformatiky I – zadání

Cílem mého projektu je vytvoření programu pro vyhledávání strukturních podobností mezi proteiny na vstupu a proteiny z databáze PDB. Z reprezentativní podmnožiny PDB databáze budou nejprve vybrány (procházením FASTA sekvencí) proteiny, které vyhoví požadavkům na sekvenční podobnost (vyhledem ke vstupu) v místech potenciálních kontaktů; u proteinů, které projdou tímto sítem, bude dále podrobně zkoumána strukturní podobnost (s proteiny na vstupu).

Vstup:

Data mapující kontakty v proteinech; formát řádků vstupu:

ID POS1 POS2 S1 S2

kde:

- ID je PDB identifikátorem proteinu,
- S1 je první sekvencí účastnící se kontaktu,
- S2 je druhou sekvencí účastnící se kontaktu,
- POS1 je pozicí kontaktní sekvence S1,
- POS2 je pozicí kontaktní sekvence S2.

Výstup:

Data mapující kontakty a sekundární strukturu v proteinech podobných proteinům na vstupu; formát řádků výstupu:

ID POS1 POS2 S1 S2 *CARMSD *ARMSD *TA *NDS

kde:

- ID je PDB identifikátorem proteinu,
- S1 je první sekvencí účastnící se kontaktu,
- S2 je druhou sekvencí účastnící se kontaktu,
- POS1 je pozicí kontaktní sekvence S1,
- POS2 je pozicí kontaktní sekvence S2,
- CARMSD je RMSD vzdálenost alfa uhlíků v páteřích kontaktních sekvencí (udaná v angstromech)
- ARMSD je RMSD vzdálenost všech atomů kontaktních sekvencí (udaná v angstromech)
- TA jsou torzní úhly v oblastech kontaktních sekvencí,
- NDS jsou sekundární struktury v oblastech kontaktních sekvencí.

- Prvky výstupu označené * jsou volitelné (viz přepínače).
- Výpisy CARMSD, ARMSD, TA a 2NDS jsou uvozeny identifikátorem kontaktní sekvence (S1/S2).
- Do výstupních výpisů jsou zahrnuty i vstupní proteiny (pro možnost porovnání torzních úhlů a sekundárních struktur); nalezené proteiny jsou porovnávány vždy s tím vstupním proteinem, který jim bezprostředně předchází ve výstupním výpisu.

Nepovinné parametry:

re1 re2 – regulární výrazy upřesňující/omezující nároky na podobnost se sekvencemi S1 a S2 vstupních proteinů (při neuvedených parametrech bude vyžadována dokonalá shoda kontaktních sekvencí)

Volitelné přepínače:

-G – výpis souhrnných statistik na zvláštních řádcích výstupu;

-C <reálné číslo c> – výpis hodnot CARMSD (případné c slouží jako horní limit pro výpis);

-A <reálné číslo a> – výpis hodnot ARMSD (případné a slouží jako horní limit pro výpis);

-T – výpis hodnot torzních úhlů;

-N – výpis sekundárních struktur;

-Q – kompletní výpis.