Základy praktické bioinformatiky

Téma 2/10

Proteinová bioinformatika I

Cíle:

Student bude schopen vyhledat a stáhnout požadovanou sekvenci proteinu. Na základě sekvence bude schopen určit (nalézt nebo spočítat) fyzikálně-biochemické vlastnosti proteinu.

Hledání proteinových sekvencí

Uniprot (http://www.uniprot.org/)

Uniprot představuje sekundární databázi, která obsahuje proteinové sekvence revidované (v rámci Swiss-Prot databáze) a nerevidované (v Tremble databázi), které vznikly pouhým překladem sekvencí nukleotidových. V databázi Uniprot je k nalezení kromě názvu proteinu a genu, také přístupové číslo (např.P15559) unikátní pro každý protein. Dále zde můžeme nalézt funkci proteinu, buněčnou lokalizaci, strukturní informace, „cross-reference" do jiných databází a především sekvence všech isoforem proteinu.

Sekvence jsou zobrazovány ve formátu pro přehlednost rozdělenou čísly:

```
        10         20         30         40         50
 MVGRRALIVL AHSERTSFNY AMKEAAAAAL KKKGWEVVES DLYAMNFNPI
        60         70         80         90        100
 ISRKDITGKL KDPANFQYPA ESVLAYKEGH LSPDIVAEQK KLEAADLVIF
       110        120        130        140        150
 QFPLQWFGVP AILKGWFERV FIGEFAYTYA AMYDKGPFRS KKAVLSITTG
```

Z hlediska další práce se sekvencemi, je tento formát ale nepoužitelný, Uniprot umožňuje náhled nebo pro přímé stažení i **FASTA formát** každé sekvence:

```
>sp|P15559|NQO1_HUMAN NAD(P)H dehydrogenase [quinone] 1 OS=Homo sapiens GN=NQO1 PE=1 SV=1
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL
KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV
FIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFC
GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
```

(FASTA formát=styl zápisu sekvencí, v prvním řádku je >, následovaný názvem sekvence (může být cokoliv, na dalším řádku začíná vlastní sekvence)

V sekci „**Pathology & Biotech**" jsou k nalezení informace o propojení příslušného proteinu s různými patologickými stavy, případně propojené odkazy do chemických databází obsahujících léčiva (př. Drugbank), která mají s daným proteinem nějakou spojitost (př.ligandy, inhibitory…)

V sekci „**Cross-references**“ jsou k nalezení odkazy do dalších databází, včetně odkazů na referenční sekvence s unikátními přístupovými kódy do databáze NCBI a to začínající „NP_...“ pro proteinové sekvence a „NM_...“ pro sekvence nukleotidů (příslušné kódující sekvence).

NCBI Protein database (https://www.ncbi.nlm.nih.gov/protein)

Primární databáze, která obsahuje veškeré proteinové sekvence (nerevidované). Sekvence se opakují a nejsou příliš přehledně dohledatelné (ideální je dostat se na příslušnou sekvenci odkazem skrz databázi Uniprot).

Hned v úvodu zobrazení konkrétního proteinového vstupu „entry“ je odkaz na **FASTA** formát sekvence, čistý zápis pro další aplikace. (Na konci vstupu je tato sekvence zobrazená s čísly pro přehlednost.) **Graphics** umožňuje grafické zobrazení sekvence, včetně různých vlastností, důležitých oblastí apod.

Dále je identifikační číslo konkrétní sekvence (accession number), většinou „NP_...“, počet aminokyselin a název proteinu. Následuje většinou odkaz na zdroj nukleotidové sekvence, ze které příslušný protein vznikl (REFSEQ „NM_...“).

Tato databáze obsahuje navíc **reference na literaturu** týkající se daného proteinu, které je možné si prohlížet přímo přes odkaz na Pubmed (PMID číslo).

V pravém sloupci jsou možnosti různých typů analýz sekvencí, BLAST (hledání podobnosti), identifikace domén nebo přímo hledání úseků v dané sekvenci.

Analýza proteinových sekvencí

V rámci portálu **SMS Suite** je možné provádět různé úpravy a počítat různé vlastnosti proteinových sekvencí. Sekvence se mohou vkládat jednotlivě nebo po větším počtu ve formátu fasta.

Portál zahrnuje například:

Filter protein-program, který „očistí“ proteinovou sekvenci od čísel či mezer a vrací čistý fasta formát.

Range extractor protein-program, který umožní vybrat požadovanou část sekvence. Zadanou například pořadím aminokyselin (od..do), nebo pořadím od konce apod.

Protein Isoelectric Point-počítá izoelektrický bod dané aminokyselinové sekvence

Protein Molecular Weight-počítá molekulovou hmotnost dané aminokyselinové sekvence

Protein Stats-počítá statistiku zastoupení jednotlivých aminokyselin zadané sekvence

V rámci portálu Expasy.org lze počítat vlastnosti zadaných sekvencí například programem „**ProtParam**“, který po zadání čisté sekvence (ne fasta formát) vypočítá molekulovou hmotnost, izoelektrický bod, statistiku atd.

Simulace štěpení proteasami

V rámci portálu Expasy.org lze simulovat štěpení zadaných sekvencí různými proteasami programem „**Peptide Cutter**“. Při základním nastavení program predikuje štěpení všemi možnými proteasami. Ve výstupu abecedně ukazuje, jaké proteasy štěpí kolikrát a za jakými aminokyselinami (dle pořadí v sekvenci).

Zadání lze omezit jen na vybrané enzymy (například Trypsin) a při zaškrtnutí požadavku tabulky jednotlivých peptidů, získáme také přehled jak dlouhé a jakou hmotnost by měly vzniklé fragmenty po štěpení.