Základy praktické bioinformatiky

Téma 5/10

Nukleotidová bioinformatika I

Cíle:

Student bude schopen vyhledat a stáhnout nukleotidové sekvence vybraných genů a získat relevantní informace o těchto genech. Student bude schopen analyzovat základní vlastnosti vybraných nukleotidových sekvencí.

Hledání nukleotidové sekvence

GeneBank (NCBI: https://www.ncbi.nlm.nih.gov/nuccore)

Tato nukleotidová databáze je součástí amerického serveru NCBI. Spolu s evropskou EMBL a japonskou DDBJ představuje, jakožto primární databáze, hlavní zdroj všech nukleotidových sekvencí. Každá jednotlivá sekvence je obsažena v tzv. flatfile zápisu, který kromě vlastní sekvence obsahuje přístupové číslo (accession number), organismus, kódující oblast (CDS), rozdělení do exonů, pozici v chromozomu, odkazy na relevantní literaturu atd. Vlastní sekvence je dostupná ve fasta formátu, hned v úvodu po nakliknutí „FASTA“ obsahu. Dále umožňuje grafické znázornění sekvence (Graphic), kde je vidět rovnou rozdělení do exonů, kódující úsek sekvence. V odkazu vpravo lze zapnout „Highlight sequence features“, což umožní náhled na vybrané části sekvence. Aktivní (hnědě zabarvenou) část sekvence lze potom stáhnout přes odkaz FASTA v pravém dolním rohu.

Gene (NCBI: https://www.ncbi.nlm.nih.gov/nuccore)

Databáze Gene představuje sekundární nukleotidovou databázi, která obsahuje celou řadu přidaných informací ke každému zpracovávanému genu. Obsahuje seznam (pro rychlý přesun je možné použít rychlé odkazy vpravo) dostupných informací s možností prokliknutí do zdrojových databází např. lokalizace v chromozomu do databáze Mapviewer, odkazy na referenční sekvence proteinů či nukleotidů do databází Uniprot, či GeneBank. Obsahuje data ze studií expresí vybraného genu v různých tkáních, bibliografické odkazy, variace sekvenční atd.

Dále existuje celá řada specializovaných databází, kde lze získat potřebné informace o genech, především v souvislosti s nějakými patologiemi, s větší, či menší mírou přehlednosti. Jako příklad můžeme uvést **Ensembl**, Atlas genetiky (chromozomů) v onkologii či hematologii, **Genecards**, zaměřené na onkologii, obsahující také například komerční zdroje protilátek pro studované proteiny. Tyto databáze většinou slouží pro konkrétní potřeby a vyhledávání relevantních informací dle zaměření pracoviště.

Databáze SNP (jedno-nukleotidových polymorfismů)-veškeré polymorfismy, mutace nebo „odlišnosti“ identifikované v sekvencích jsou zahrnuty v příslušných databázích. NCBI portál obsahuje několik nástrojů k jejich prohledávání. Veškeré varianty jsou shromažďovány v databázi pod přístupovým kódem „rs…(číslo)...“

dbSNP (https://www.ncbi.nlm.nih.gov/snp)

Variation Viewer (https://www.ncbi.nlm.nih.gov/variation/view/)

Analýza nukleotidových sekvencí

V rámci portálu SMS Suite je možné provádět různé úpravy nuklotidových sekvencí podobně jako u proteinových sekvencí. Sekvence lze vkládat jednotlivě nebo ve větším počtu ve formátu fasta.

Portál zahrnuje například:

Filter DNA-program umožňující zbavit se všech „kontaminujících“ prvků v sekvencích-čísla mezery neznámé znaky apod.

Range Extractor DNA-umožňuje získat požadovanu část sekvence číselně zadanou “od..do“.

DNA Stats-program zpracovávající četnost jednotlivých bazí či dinukleotidů (sousedících dvou bazí).

Reverse complement-program, který přepíše druhý komplementární řetězec příslušné sekvence (tedy komplementární v orientaci 5´- 3´ tedy reversní k původní sekvenci).

Př. máme-li sekvenci 5´-AAGTCCAT-3´ pak reverzně komplementární sekvence je: 5´-ATGGACTT-3´