Bioinformatika

Biologická data a jejich typy

Biologická data představují základní vstup pro bioinformatiku a zahrnují široké spektrum informací popisujících strukturu, funkci, variabilitu a chování živých organismů. Nejčastěji se setkáváme s:

- Genomickými daty (sekvence DNA a RNA)
- Proteomickými daty (informace o složení, množství a modifikacích proteinů)
- Metabolickými daty (údaje o metabolických drahách a produktech)
- Fenotypovými daty (popis znaků a vlastností organismu)

Správné zpracování těchto dat je klíčové pro pochopení biologických procesů, evoluce, chorob i pro biomedicínský výzkum.

Transkripce, translace a replikace

Transkripce

Transkripce je proces, během něhož je genetická informace uložená v DNA přepisována do molekuly RNA. Tento enzymaticky řízený proces zajišťuje RNA polymeráza a probíhá ve třech základních krocích:

- Iniciace: RNA polymeráza se naváže na specifickou oblast DNA zvanou promotor.
- Elongace: Postupné přidávání ribonukleotidů podle komplementarity s DNA.
- Terminace: Uvolnění hotové RNA po dosažení terminátorové sekvence.

Transkripce je klíčová pro regulaci genové exprese.

Translace

Translace je proces, kterým se informace v mRNA překládá do pořadí aminokyselin v bílkovině. Probíhá na ribozomech ve třech krocích:

- Iniciace: Navázání malé ribozomální podjednotky na start kodon mRNA (AUG).
- Elongace: tRNA přináší aminokyseliny podle kodonů; vznikají peptidové vazby.
- Terminace: Uvolnění polypeptidu po dosažení stop kodonu (UAA, UAG, UGA).

Replikace

Replikace je proces, kdy vzniká přesná kopie DNA před dělením buňky:

- Iniciace: Otevření dvoušroubovice DNA v místě originu replikace.
- Elongace: DNA polymeráza syntetizuje nové vlákno.
- Terminace: Oddělení dvou molekul DNA po dokončení kopírování.

Biologické databáze

Biologické databáze uchovávají, organizují a zpřístupňují rozsáhlé soubory dat (sekvenování genomů, analýza proteinů aj.). Mezi nejvýznamnější patří:

- GenBank (sekvence nukleových kyselin)
- Protein Data Bank (PDB) (3D struktury proteinů)
- Ensembl (anotované genomy)

Sequence alignment (zarovnání sekvencí)

Zarovnání sekvencí je základní bioinformatická technika, která hledá shody/podobnosti mezi sekvencemi DNA, RNA nebo proteinů.

Důvody pro zarovnání sekvencí:

- Zjištění homologie (společného původu) či funkční podobnosti
- Identifikace důležitých oblastí (např. aktivní místa enzymů)
- Odhalení evolučních vztahů a mutací

Typy zarovnání:

- Globální zarovnání (Needleman-Wunsch): srovnává celé sekvence
- Lokální zarovnání (Smith-Waterman): hledá nejpodobnější úseky v rámci sekvencí

Skórování zarovnání:

- Skórovací matice (BLOSUM, PAM) a penalizace za mezery
- Různé matice pro různé evoluční vzdálenosti (např. BLOSUM62 pro blízkou homologii)

BLAST a heuristické algoritmy pro vyhledávání podobných sekvencí

BLAST (Basic Local Alignment Search Tool) je nejrozšířenější nástroj pro rychlé vyhledávání podobných oblastí v databázích. Pracuje ve dvou fázích: rychlé nalezení krátkých identických/podobných úseků (slov), pak rozšíření a přesnější zarovnání.

- Varianty: blastn (DNA-DNA), blastp (protein-protein) aj.
- Statistické ukazatele: E-value (očekávaný počet náhodných shod), p-value
- Další: FASTA (podobný, někdy citlivější na vzdálenější homologii)

Dotplot a vizuální porovnávání sekvencí

Dotplot je grafická metoda pro porovnání dvou sekvencí. Na osách jsou znaky obou sekvencí, v místech shody je tečka. Diagonály označují podobné úseky; lze odhalit inverze, duplikace, posuny.

Homologie, ortologie a paralogie

- Homologie: sekvence mají společného předka
    - Ortologie: geny vzniklé speciací (různé druhy, stejná funkce)
    - Paralogie: geny vzniklé duplikací v jednom organismu (mohou diverzifikovat)

Rozpoznání homologických vztahů je klíčové pro přenos funkční anotace a evoluční analýzy.

Strukturní srovnávání a predikce struktury proteinů

Strukturální srovnávání

Porovnávání 3D struktur proteinů může odhalit podobnosti nezjevné na úrovni sekvence a napomáhá objasnit evoluční a funkční vztahy.

Predikce proteinové struktury

- Homologní modelování: modelování podle známé struktury podobného proteinu (templátu)
- Ab initio modelování: bez známého templátu, čistě na základě fyzikálně-chemických principů
- Threading: hledání nejvhodnějšího známého foldingu bez významné sekvenční podobnosti

Moderní přístupy (např. AlphaFold) využívají kombinace algoritmů a strojové učení.

Substituční matice BLOSUM

BLOSUM (BLOcks SUbstitution Matrix): sada substitučních matic pro zarovnávání proteinů. Odvozeny z bloků evolučně diverzifikovaných sekvencí; umožňují hodnotit pravděpodobnost záměn aminokyselin při evoluci. Např. BLOSUM62 je vhodná pro blízké příbuznosti.

Predikce sekundární struktury proteinů

- Sekundární struktura = alfa-helixy, beta-listy, smyčky
- Metody predikce: Chou-Fasman, GOR – využívají pravděpodobnosti výskytu aminokyselin v daném typu struktury.

Multiple Sequence Alignment (MSA)

MSA znamená zarovnání tří a více sekvencí současně. Pomáhá najít konzervované (důležité) oblasti pro evoluční studie, funkční anotace i predikci proteinových struktur.

- Progresivní metody: např. CLUSTAL
- HMM metody: využití skrytých Markovovských modelů