Identificació de seqüències i comprensió d'anotacions

Bases de dades en Biologia

En la actualitat es disposa d'una gran quantitat d'informació sobre seqüencia de proteïnes i acids nucleics. Aquesta informació es troba emmagatzemada a bases de dades de tipus divers:

Bases de dades primàries

Son aquelles que es fan servir com a repositoris de la informació de seqüencia obtinguda directament de les fonts, sense processar, o amb un processament nomes tècnic.

Les mes rellevants són:

GenBank (https://www.ncbi.nlm.nih.gov/genbank/): Sequències de DNA, proteina, genomes complets.

RefSeq (https://www.ncbi.nlm.nih.gov/refseq/): Seqüències de referència

ENA (European Nucleotide Archive, https://www.ebi.ac.uk/ena): Seqüencies de DNA

Uniprot (https://www.uniprot.org/): Seqüències de proteïna. Conté dues seccions: trEMBL: Seqüencies obtingudes automàticament; i SwissProt: seqüències de proteïna amb anotacions funcionals acurades.

Protein Data Bank (https://www.rcsb.org/): Estructures tridimensionals de macromolècules

Bases de dades secundàries

Aquestes bases de dades inclouen informació addicional que pot provenir d'anotació manual (com SwissProt), o de resultats d'anàlisi bioinformàtica.

Algunes de les més rellevants serien:

SwissProt: La secció anotada de Uniprot

Ensembl (http://ensembl.org): Lloc de referència per a informació sobre gens, transcrits, variants de seqüència.

Interpro (https://www.ebi.ac.uk/interpro/): Classificació de famílies de proteïna amb força informació funcional.

Comparació de seqüències

La operació més rellevant que es realitza en bioinformàtica és la comparació de seqüències. La semblança de seqüència entre dos gens o dues proteïnes ens permet inferir que les propietat de les mateixes seran semblants. En particular, fem servir la comparació de seqüències per:

- Identificar noves seqüències a partir de la semblança amb les seqüencies conegudes
- Avaluar el grau d'homologia, que a la seva vegada ens permet inferir la semblança de seqüencia i estructura
- Identificar patrons relacionats amb propietats funcionals com l'activitat enzimàtica.
- Predir estructures tridimensionals de proteïnes
- Predir lloc d'interacció proteïna-DNA

L'eina més habitual per identificar seqüències i localitzar homòlegs és BLAST (https://blast.ncbi.nlm.nih.gov/). Blast permet comparar seqüències de DNA o proteïna contra diverses bases de dades.

- Bases de dades no reduntants (nr). Les més extensives, contenen un representant de tots les seqüències conegudes
- Sequències de refèrencia: Compostes de seqüències seleccionades com a referència (correctament anotades i contrastades).
- SwissProt: Seqüències de proteïna anotades
- PDB: Seqüencies de protein amb informació estructural.