M. V. Cubellis

Modulo di Biomarcatori

BANCHE DATI

Le banche dati biologiche si possono dividere in:

- primarie (di archivio)
- secondarie (di annotazione e/o curate)

UNIPROT

Banca dati che fornisce sequenze e informazioni sulle proteine. E'divisa in due parti :

1. trEMBL→insieme di sequenze proteiche non annotate
2. swissprot→ contiene sequenze proteiche annotate manualmente e riviste/curate

il nome di una proteina è costituito da due parti :

- la prima parte differisce : se ci si trova in trEMBL corrisponde ad una sigla ; se ci si trova in swissprot è un qualcosa che riguarda la proteina
- la seconda parte identifica la specie di provenienza

MIM/OMIM

E' una banca dati che contiene tutti i disturbi di natura mendeliana e si concentra principalmente sul rapporto tra fenotipo e genotipo. In OMIM sono contenuti circa 12000 geni e tutte le informazioni sono aggiornate quotidianamente.

PDB (protein data bank)→ primaria

Conserva tutte le strutture tridimensionali di proteine derivanti dalla cristallografia a raggi X e dalla NMR.

Cristallografia a raggi X→tecnica della cristallografia in cui l'immagine, prodotta dalla diffrazione dei raggi X attraverso lo spazio del reticolo atomico in un cristallo, viene registrata e quindi analizzata per rivelare la natura del reticolo

Risonanza magnetica nucleare (NMR).

I primi righi che riguardano la struttura di una proteina iniziano con la parola "ATOM" che indica il tipo di atomo e la sua posizione nello spazio.

PDBSUM→secondaria

È un database che fornisce una visione del contenuto di ciascuna struttura 3D depositato nella Protein Data Bank (PDB).

PUBMED

Comprende più di 22 milioni di citazioni di natura biomedica ricavate da MEDLINE e da libri on-line. Inoltre sono spesso inclusi collegamenti ad articoli interi.

dbSNP

è un archivio pubblico gratuito dove sono inserite molte le variazioni genetiche, polimorfismi, contenuti all'interno di una specie e tra le diverse specie (la maggior parte sono polimorfismi patologici). In questa banca dati è possibile effettuare ricerche avanzate, come in PDB.

SNP→ polimorfismi di un singolo nucleotide.

EMBL, GENBANK, DDBJ→ banche dati primarie

sono banche dati nucleotidiche . EMBL è una banca dati europea, GENBANK americana e DDBJ giapponese.

ENSEMBL

E' un progetto nato nato tra EMBL-EBI allo scopo di fornire banche dati genomiche di veretebrati e di altre specie genomiche. Tutte le informazioni sono disponibili gratuitamente on-line.

Le sequenze ricreano i cromosomi, ci si può muovere su di essi e si può conoscere quali geni sono a monte e a valle della nostra sequenza.

ORPHANET

Banca dati che raccoglie tutte le malattie mendeliane rare. E' una banca dati orientata al pubblico.

Pfam

Il database Pfam è una grande collezione di famiglie di domini di proteine, ognuno rappresentato da allineamenti di sequenze multiple.

Le proteine sono generalmente composti da una o più regioni funzionali, comunemente chiamati domini. Diverse combinazioni di domini dar luogo alla vasta gamma di proteine presenti in natura. L'identificazione di domini che si verificano all'interno delle proteine possono quindi fornire comprensione della propria funzione.

CONCETTO DI OMOLOGIA:

due proteine si dicono omologhe se si sono evolute da un ancestore comune. Si dividono in:

- Paraloghe→ proteine che svolgono la stessa funzione in specie uguali

- Ortologhe→ proteine che svolgono la stessa funzione in specie diverse

SRS SERVER

Motore di ricerca per banche dati. Permette una ricerca molto dettagliata, attraverso la scelta e l'inserimento delle parole chiave più appropriate in appositi campi di ricerca.

BLAST

Programma che confronta le sequenze nucleotidiche o proteiche mediante allineamenti e ne propone quelli più significativi.

CLUSTAL W

Permette di confrontare sequenze proteiche o nucleotidiche attraverso allineamenti multipli.

POLYPHEN

è uno strumento che predice il possibile impatto di una sostituzione amminoacidica sulla struttura e la funzione di una proteina umana attraverso allineamenti multipli con specie diverse.

UCSC GENOME BROWSER

E' una banca dati che concentra i suoi contenuti su sequenze di riferimento e bozze di lavoro riguardanti una vasta collezione di genomi.

BLAT

È una versione un po' rivista di blast. Strumento che riesce a mappare velocemente la sequenza genomica desiderata e fa allineamenti con specie diverse presenti nel browser. È studiato appositamente per fare ricerche su nucleotidi, ed in particolare è molto accurato nel prevedere i siti di splicing a altre strutture genomiche. In generale è molto utilizzato nelle ricerche su interi genomi e nucleotidi (di specie più o meno silimi), mentre blast resta più affidabile per la ricerca su proteine.

COME E' FATTO UN FILE PDB ?

La parte del file che contiene le vere info sulla proteina è così strutturato :

- ATOM→ indica il tipo di atomo e la sua posizione nello spazio
- Numero→numero dell'atomo all'interno della struttura
- N, C, O, CA, CB, CG→ rispettivamente la natura dell'atomo : azoto, carbonio, carbonio alfa, carbonio beta, carbonio gamma.
- Amminoacido di partenza

- A (alfa) e B (beta)→ fa riferimento alla catena alfa o beta della proteina
- Numero→ indica da quale posizione si inizia a lavorare dopo la cristallografia. NB alcune volte si parte, per esempio, dal numero 12 in quanto dopo la cristallografia si possono perdere parti di proteine. In poche parole : si annota solo ciò che il cristallografo vede.
- Nelle tre colonne successive si indicano le coordinate spaziali dell'atomo : X, Y e Z.

MODELING PER OMOLOGIA

Il modelling per omologia si basa sull'osservazione che proteine che presentano un buon livello di similarità di sequenza, in genere mostrano un buon livello di similarità di struttura. Nella maggioranza dei casi, due proteine che mostrano un'identità di sequenza superiore al 30% conservano una similarità di struttura che consente di utilizzarne una come modello per poter costruire il modello di omologia dell'altra ; inoltre tanto più è alta l'identità di sequenza, tanto maggiore risulta la similarità delle loro strutture tridimensionali e aumenta, di conseguenza, l'affidabilità del modello ottenibile.

In particolare, proteine con almeno il 50% di identità di sequenza in genere mantengono circa il 90% dei residui in posizioni strutturalmente conservate.

Non è sempre semplice costruire un modello tridimensionale buono per una proteina.

Un buon programma è MODELER, ma non è accessibile a tutti. Esistono diverse procedure, alcune di queste in rete, che sono state sviluppate al fine di rendere più accessibile il compito ad utenti non esperti. Uno di questi è SWISS MODEL, che consente di identificare possibili modelli (template) per la propria sequenza proteica e di produrre modelli 3D di buona qualità.

COME LAVORA SWISS MODEL ?

- Cerca la proteina
- Cerca una struttura omologa in PDB
- Allinea le due sequenze