

I Esercitazione di Bioinformatica **“Data retrieval”**

Seguire il percorso indicato nel testo sottostante e stilare una breve relazione con i contenuti recuperati e elaborati man mano usando diverse risorse informatiche.

A) Data retrieval da database dell'NCBI

NCBI (<http://www.ncbi.nlm.nih.gov/>) è un'importante risorsa da cui reperire informazioni biologiche. Fornisce un'interfaccia unificata per accedere a molti database differenti per contenuto e organizzazione, quali database di conoscenza (letteratura, geni, malattie), biosequenze (nucleotidiche, proteiche) e strutture (3D di proteine, domini proteici conservati).

A partire da <http://www.ncbi.nlm.nih.gov/guide/all/>

1) Ricerca bibliografica in NCBI Bookshelf (Literature):

- Dare un'occhiata a “Book List”.
- Cercare Cadherins in Cancer medicine.

2) Ricerca bibliografica in NCBI PubMed (Literature):

- Usare PubMed Advanced Search Builder, che permette di ricercare specifiche parole chiave in specifici campi e combinare le ricerche con operatori Booleani.
- Trovare l'articolo più recente sul rhabdomyosarcoma embrionale (embryonal rhabdomyosarcoma [Title]).
- Trovare l'articolo più recente sulla rivista Science riguardante processamento dell'RNA nel cristallino (lens)
- Trovare l'articolo più recente che abbia tra gli autori “Rajewsky-N”.

3) Retrieval di sequenze con NCBI Nucleotide (DNA & RNA; Genome, gene and transcript sequence data):

- Cercare la sequenza di riferimento del gene mitocondriale per il citocromo b umano (cytochrome; cytb) e scaricarla. Quanto e' lunga?

4) Retrieval di sequenze con NCBI Protein:

- Cercare la sequenza della PROTEINA umana alpha crystallin-A, annotarsi la lunghezza e copiare la sequenza in formato FASTA.

5) NCBI Genome:

Consultare le pagine raggiungibili da NCBI Genome (<http://www.ncbi.nlm.nih.gov/genome>), quali “Human genome”, “Viruses”, ecc.

Provare a raggiungere le informazioni relative ad uno specifico genoma completo:

→ *Propionibacterium acnes* KPA171202

Usare Taxonomy database per la classificazione (classe, sottoclasse, ecc.) utile alla navigazione dell'albero.

Quanti geni protein coding contiene questo genoma batterico?

Scaricare la seq in format FASTA della proteina più lunga codificata.

→ Homo sapiens (human) genome view, Annotation Release 106. Raggiungere le mappe del cromosoma Y, e cercare di capire quanto e' lungo e quanti geni annotate contiene. Usare "Table view" per identificarne alcuni inclusi nella regione di 4 megabasi adiacente a Yter e circa corrispondente alla banda Y11.3 e scaricare le descrizioni dei geni mediante "Download data".

Dalla tabella seguire il link a "LINC00278", orientarsi e scaricare la sequenza di riferimento (Accession NR_...) in format FASTA.

B) Dalla malattia genetica, al gene, attraverso il genoma.

Prenderemo in considerazione una malattia genetica e cercheremo di utilizzare i dati contenuti in diversi database (di malattie, geni, proteine e genomi) per reperire informazioni sul fenotipo, sui geni, sulle mutazioni che lo causano, ecc.

- 1) Collegarsi a **OMIM** e cercare "RETINITIS PIGMENTOSA"
 - a. Analizzare i risultati e commentarli (entries di geni; entries di malattie; ci sono piu' malattie diverse?).
- 2) Scegliere "RETINITIS PIGMENTOSA 9"
 - a. Analizzare le informazioni sulla patologia (pagina principale, clinical synopsis e gene map) per capire di che malattia si tratta, come si eredita e in quale cromosoma e regione genomica si trova il locus associato.
- 3) Seguire il link al locus e quindi alla "location" per arrivare alla visualizzazione grafica del contenuto di informazione della regione del genoma umano che contiene il gene. Con l'aiuto dell'esercitatore cercare di orientarsi nell' **UCSC Genome Browser**.

Cliccare sulla track del gene (colore blu).

- Di che gene si tratta? Che funzione ha? Che altre informazioni sul gene possono essere reperite in questo database ?
- Di quanti esoni e' composto il gene ?
- Quali sono i due geni "vicini" sul DNA genomico ?
- A partire dal campo "Sequence and Links to Tools and Databases" scaricare la sequenza della proteina codificata.