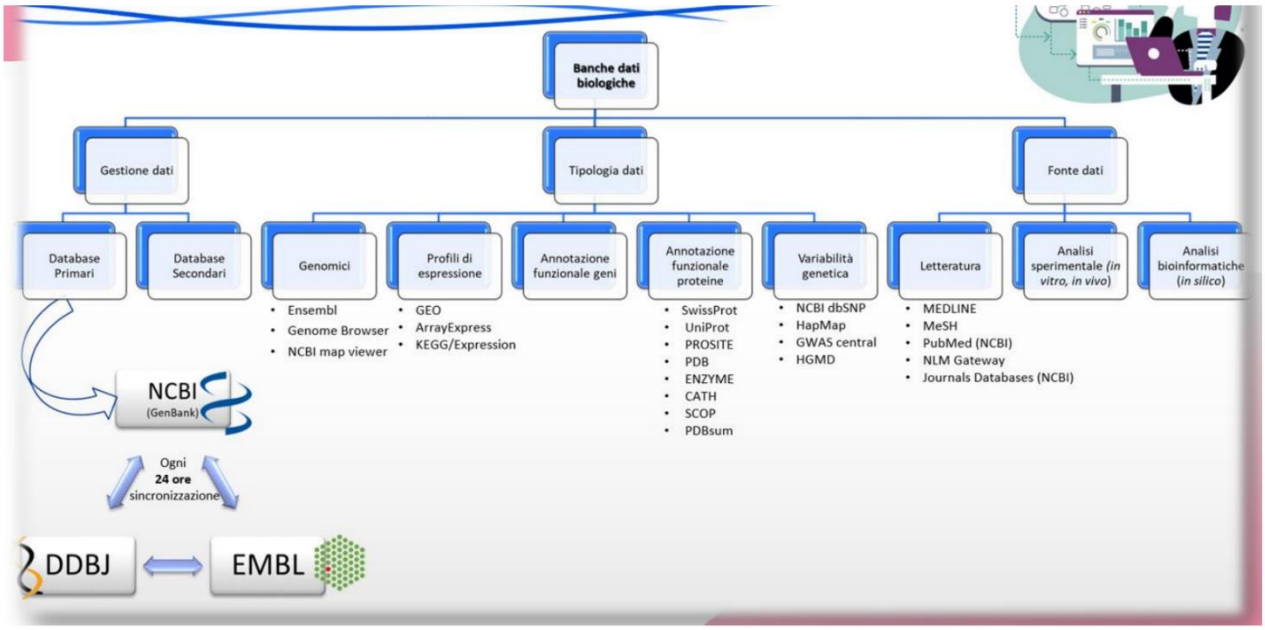
GENETICA LEZIONE N°14 11/05/2023
BANCHE DATI

EREDITARIETA'

Una banca dati biologica raccoglie informazioni e dati derivanti dalla letteratura e da analisi effettuate sia in laboratorio sia attraverso analisi bioinformatiche. Consente la consultazione e l'analisi delle informazioni in essa contenute e di ogni altra informazione a esse correlate e memorizzate in altre banche dati. La ricerca in una banca dati ha finalità statistiche e costituisce il punto di partenza di uno studio.
Ogni banca dati biologica è caratterizzata da un elemento biologico centrale che costituisce l'oggetto principale intorno al quale viene costruita la entry della banca dati.
Esempi di elemento centrale: la sequenza nucleotidica di DNA nelle banche dati da acidi nucleici.



Questa slide riassume quello che andremo ad analizzare nel corso della lezione.

Le banche dati si possono suddividere in 3 tipologie a seconda della loro funzione:

- Gestione dati: sono rappresentate da database primari. Sono sincronizzati a database europei, americani e giapponesi per formare i database secondari;
- Tipologia di dati:
    - Genomico (esempi in figura): il sistema ci restituisce informazioni e posizione del gene che sto analizzando;
    - Profili di espressione: sono dati che vengono conservati sui database citati in figura che ci forniscono informazioni sui risultati di laboratorio.
    - Annotazione funzionale di geni: i geni vanno annotati. Possono essere annotati in senso strutturale o funzionale: nel primo caso annoto la tipologia e le coordinate del gene (es: introne, esone, ecc…); nel secondo caso devo capire la mutazione a che disfunzione è correlata (= la funzione del gene).
    - Annotazione funzionale delle proteine: questi database solitamente danno le stesse informazioni, ma presentate in modo differente perché si basano su database primari differenti.
    - Variabilità genetica: si basano sullo studio delle varianti (polimorfismi). Ci indicano la frequenza, l'impatto sulla popolazione, ecc.
- Fonte dati:
    - Letteratura: il più importante è PubMed. È stato creato dal NCBI. Questi database consentono di consultare un catalogo di informazioni.

- Analisi sperimentali (in vivo o in vitro): la raccolta di tutti i risultati da un laboratorio,
- Analisi bioinformatiche (in silico): si valuta, per esempio, l'impatto funzionale di una variante a partire da un software che si basano sui database e ci dicono se quella variante ricade in un dominio funzionale o in un dominio conservato. La valutazione non è, quindi, pratica ma si basa sui software.

BANCHE DATI PRIMARIE E SECONDARIE

Basandoci sulle qualità e gestione delle informazioni contenute nelle banche dati, possiamo suddividerle in due classi principali:

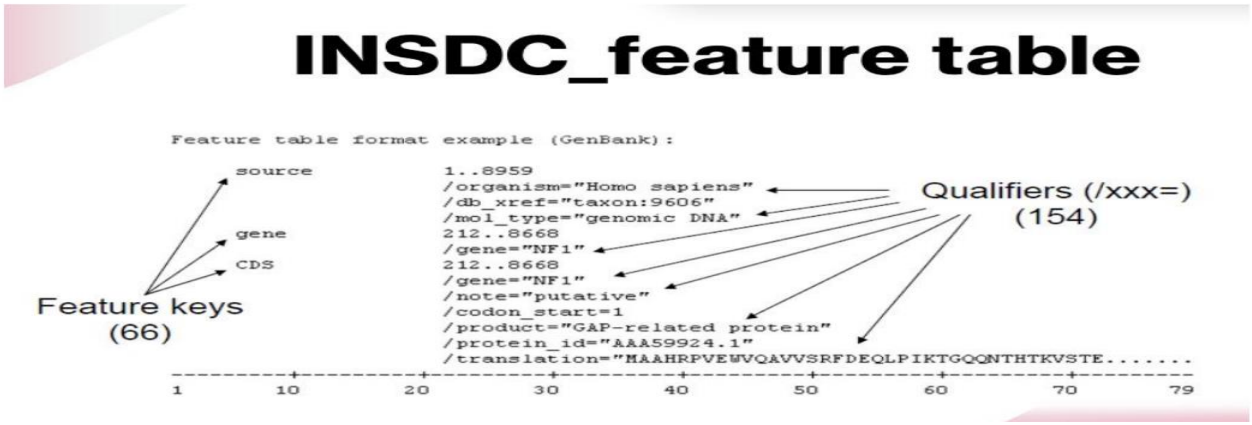
- Banche dati PRIMARIE (dette anche collettori primari): il loro ruolo è quello di raccogliere, giornalmente, tutte le informazioni che riguardano le biomolecole prodotte in tutti i laboratori del mondo e renderle disponibili.
- Banche dati SECONDARIE: dato che l'informazione contenuta nei collettori primari è sporca e ridondante, esse esaminano i dati dei collettori e correggono eventuali errori, includono informazioni aggiuntive e rendono disponibili i risultati di questo processo di raffinamento.

Uno dei principali problemi è quello legato alla nomenclatura.
Non esiste uno standard nell'assegnazione dei nomi ai geni; uno stesso gene può avere diversi nomi o uno stesso nome può individuare più geni.
Occorre, quindi, un modo per individuare univocamente i geni e le proteine e per gestire la grande quantità delle informazioni ad esse legate: nelle banche dati primarie ogni elemento (gene, sequenza, etc.) è individuato univocamente da un accession number_parola contenente lettere, numeri e simboli (spazi vuoti non ammessi).

BANCHE DATI PRIMARIE

All'inizio nasce in Germania l'EMBL-EBI (conteneva oltre 500 enti di DNA e RNA). Nel 1982 nasce una banca simile in America (NCBI). Nel 1986 nasce un mirror della Gen Bank in Giappone.
I tre centri si unirono poi nell'International Nucleotide Sequence Database Collaboration nel 1992 stabilendo delle regole comuni nella gestione delle banche dati. I dati vengono integrati in database secondari (come RefSeq o Uniprot). I database secondari hanno la funzione di formattare in maniera standard i dati delle banche primarie.

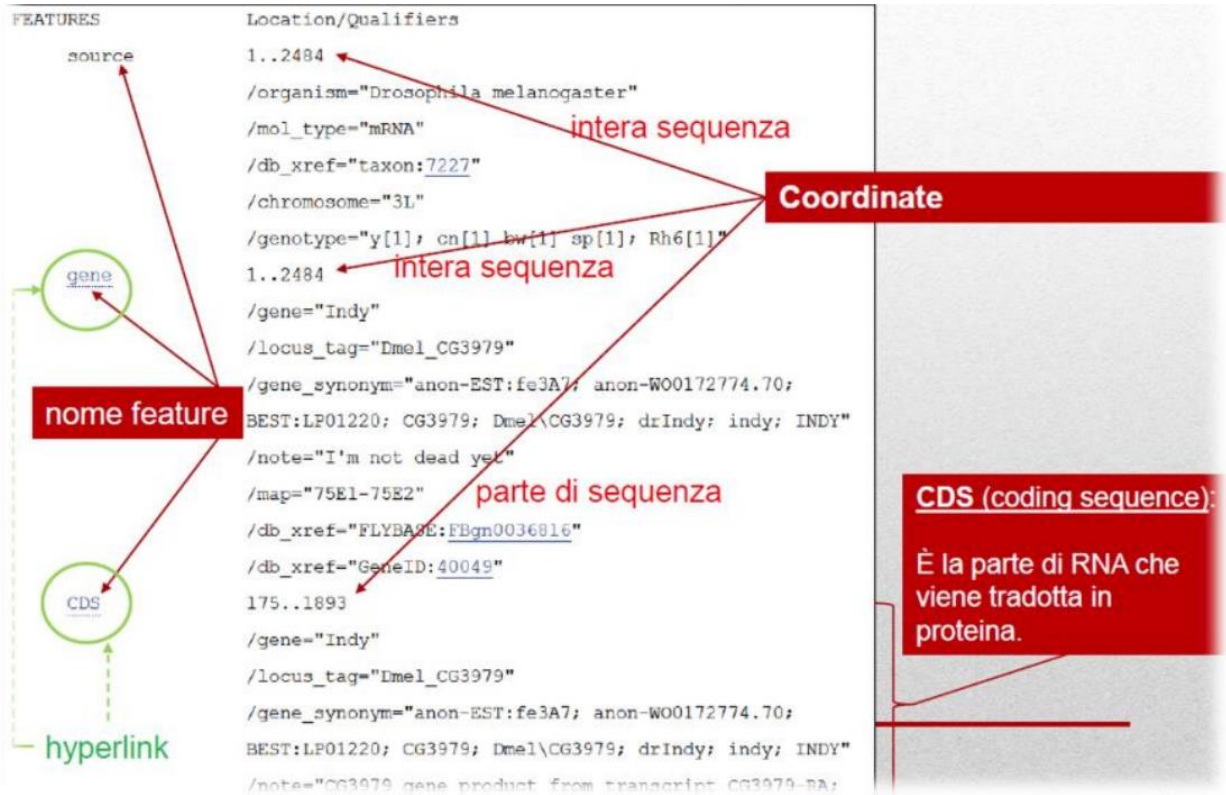
Le banche primarie sono la sorgente di tutte le altre banche dati e servono regole ferree e condivise affinché si possano sincronizzare "on a daily basis".
Per questo nasce il concetto di Feature Table. Nelle scorse lezioni abbiamo parlato del consorzio a Parigi che ha dato la nomenclatura citogenetica delle bande del cromosoma, anche in questo caso per uniformare i database primari si sono stabilite delle regole conservate in una documentazione.
Innanzitutto, si è stabilita la "Feature Table", ovvero, la facciata che si presenta quando dobbiamo leggere una sequenza o una proteina: deve essere presente un'organizzazione; dunque, vi deve essere scritto quello di cui stiamo parlando ; l'autore che ha scoperto la molecola e tante altre informazioni.
Nascono due formati diversi: il flat- file che è un testo semplice, formattato e non interattivo; HTML (linguaggio a marcatore per ipertesti), ovvero, una pagina web interattiva e di facile consultazione. In entrambe abbiamo delle cross references che ci riportano ad altri database.
Non sono pagine di facile lettura, ma ogni database è fornito di descrizioni.



L'immagine rappresenta un formato GenBank della feature table.
1..889 è la risorsa di origine e rappresenta il numero di nucleotidi (va da 1 a 8959).
I qualifiers indicano che tipo di informazione si trova in un dato campo (definiti da uno /).
Le feature keys sono: sorgente, gene e sequenza codificante (CDS). La CDS è la parte di RNA che viene tradotta in proteina (esoni).

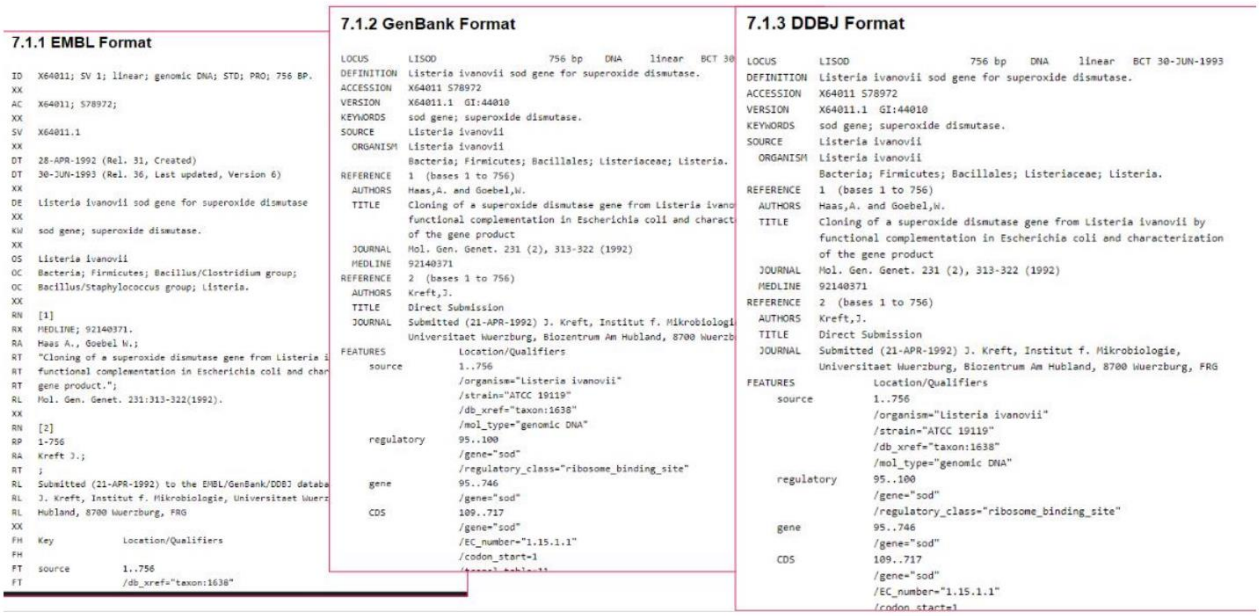
In questo caso è una sequenza amminoacidica che ha il gene NF1 la cui parte codificante va da 212 nucleotidi a 8668.

L'immagine sottostante ci fa vedere il tutto in maniera più descrittiva.



La risorsa, nell'immagine, va da 1 a 2484 nucleotidi.

FORMATTAZIONE DEI FLAT FILES



7.1.1 EMBL Format

```
ID   X64011; SV 1; linear; genomic DNA; STD; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   Listeria ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria i
RT   functional complementation in Escherichia coli and char
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databa
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerz
RL   Hubland, 8700 Wuerzburg, FRG
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
```

7.1.2 GenBank Format

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivano
            functional complementation in Escherichia coli and charact
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologi
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzb
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     regulatory      95..100
                     /gene="sod"
                     /regulatory_class="ribosome_binding_site"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
```

7.1.3 DDBJ Format

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     regulatory      95..100
                     /gene="sod"
                     /regulatory_class="ribosome_binding_site"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
```

Questi formati sono leggermente diversi, ma le informazioni sono le stesse (e anche le feature keys e i qualifiers). Le immagini ci mostrano come viene rappresentato nei tre formati lo stesso locus.

Nel formato EMBL ci sono dei codici aggiuntivi verso sx:

- Ogni sequenza inviata al database viene contrassegnata da un accessionnumber(AC) permanente per l'entry;
- Viene anche assegnata una entry name (ID);
- DT: date di creazione e aggiornamento;
- KW: parole chiave per la descrizione;
- OS: nome della specie;
- OC: classificazione tassonomica;
- RN, RA, RT, RL: informazioni sulla bibliografia;
- FT: regioni funzionalmente caratterizzate;
- SQ: sequenza nucleotidica.

Gen Bank e DDBJ sono più simili.

La sequenza in modalità ENTRY (o GenBank) è rappresentata da righe che contengono 60 caratteri (es. nucleotidi) che sono divisi in gruppi da 10 caratteri per facilitare i conteggi. Ogni riga inizia con il numero del primo carattere (nucleotide) della riga stessa.

Le informazioni sulla sequenza possono essere visualizzate anche in formati diversi:

- Summary: solo informazioni principali;

- FASTA/FASTA (text) : solo sequenza;
- GenBank (full): informazioni estese.

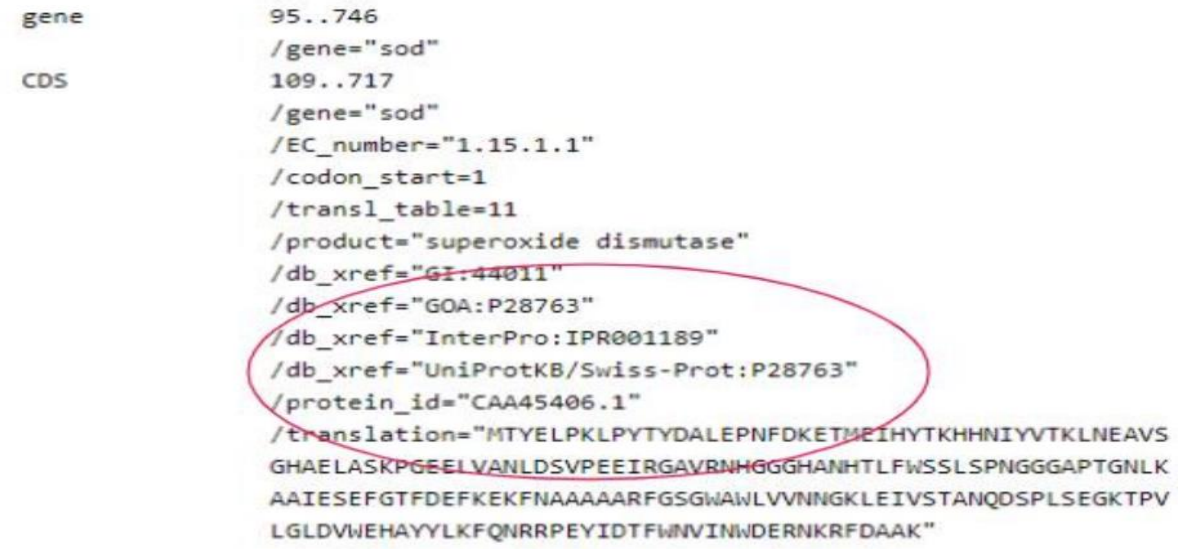
FORMATO FASTA (text)

Il formato FASTA mostra solo la sequenza (senza spazi vuoti e numeri). Inoltre, non contiene caratteri invisibili (a differenza delle altre modalità di visualizzazione).
In questo formato la sequenza può essere usata come input per programmi che effettuano analisi su di essa (programmi in silico).
La prima riga dà le informazioni sulla sequenza (inizia con >); dalla seconda inizia la sequenza vera e propria.

Una volta che abbiamo le informazioni sulla sequenza nucleotidica, possiamo esportare e salvare il film (anche mandarlo).

INTERATTIVITÀ E DATI INCROCIATI

L'interattività ha un ruolo centrale per una banca dati, perché permette di navigare tra le sue entries e quelle di altri database. Sia i flat-file sia le pagine web sono ricchi di cross-references, riferimenti che mandano ad altre banche dati generiche o specializzate.



Nell'immagine il database ha usato la banca dati InterPro e nella riga successiva si esprime l'utilizzo della banca dati UniProt.
Ogni sequenza deriva, quindi, da un database originario che viene specificato. La qualifier che viene usata per specificarlo è: /db_xref="database:identifier"

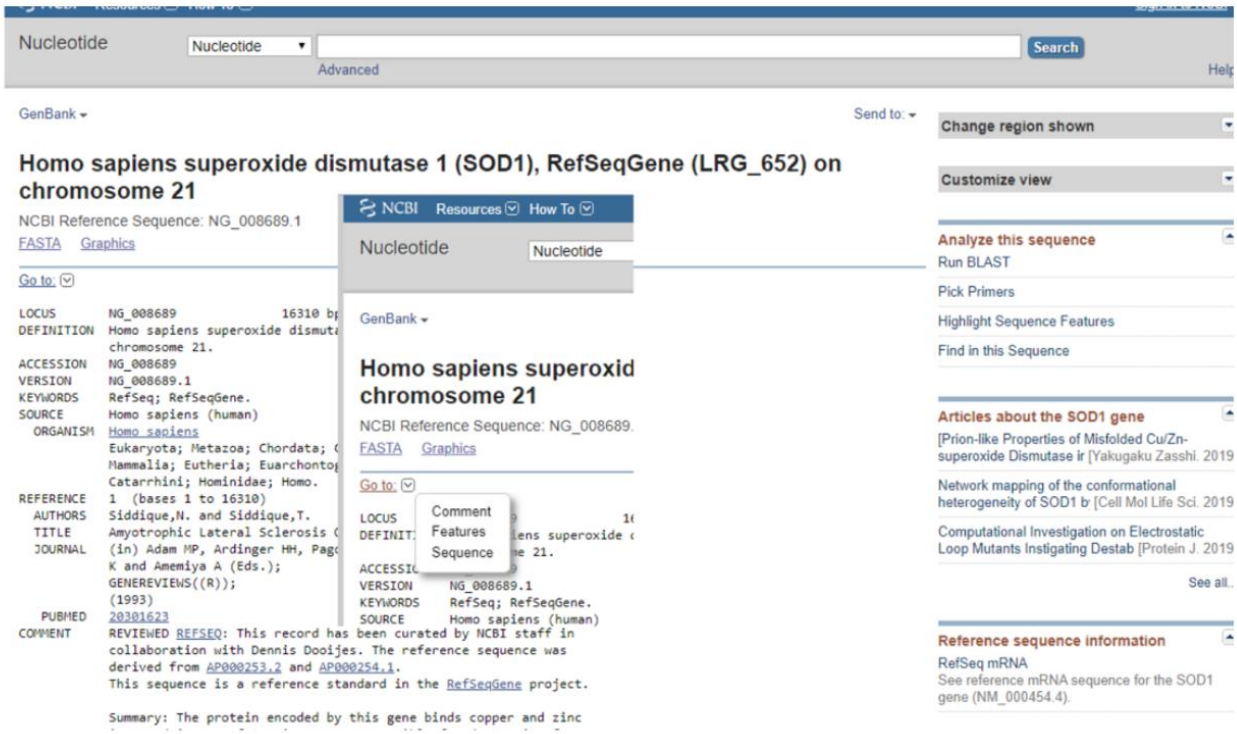
MODALITÀ DI RICERCA DEI DATI

Per realizzare l'estrazione di dati esistono vari sistemi (sistemi di interrogazione), fra cui i più efficienti sono SRS (sequence retrieval system) (EBI) ed ENTREZ (NCBI). L'interrogazione di una banca dati può avvenire in maniera banale, inserendo il nome ricercato in una finestra di tipo textsearch oppure tramite la sottomissione di formsin cui inserire varie informazioni sulla nostra ricerca. La logica di criterio è quella booleana che effettua intersezioni (operatore AND), somme (operatore OR), ed esclusioni (operatore BUT NOT), di insiemi di dati.

Il sistema SRS può essere scaricato e usato offline. Esso consente di interrogare più banche dati contemporaneamente e sfrutta i meccanismi di codifica di cross-referencing e consente la navigazione tra le banche dati.
Entrez è un database sviluppato dalla NCBI attraverso cui si possono esplorare distinti altri database. È un sistema disponibile via web per la ricerca e l'estrazione di dati da banche diverse.
Entrez, a differenza di SRS, è un modello chiuso in cui non è possibile scaricare via internet, o ottenere un software che gestisce l'intero sistema, né è possibile duplicare il sito su altri computer, né installare proprie banche dati personali. Entrez è usato soprattutto da ricercatori in quanto è un sistema di riferimento per la ricerca bibliografica sulla banca dati Medline che è la più completa banca dati bibliografica del settore bio-medico.

Le FAQ aiutano a capire come si utilizzano questi database.

Nucleotide serve per conoscere la sequenza nucleotidica di un gene. REFSEQ è il secondario.



NG sta per genomico.
"Go to" ci porta alla feature che vogliamo consultare (le pagine sono lunghe e dispersive).

DIFFERENZE TRA I DUE

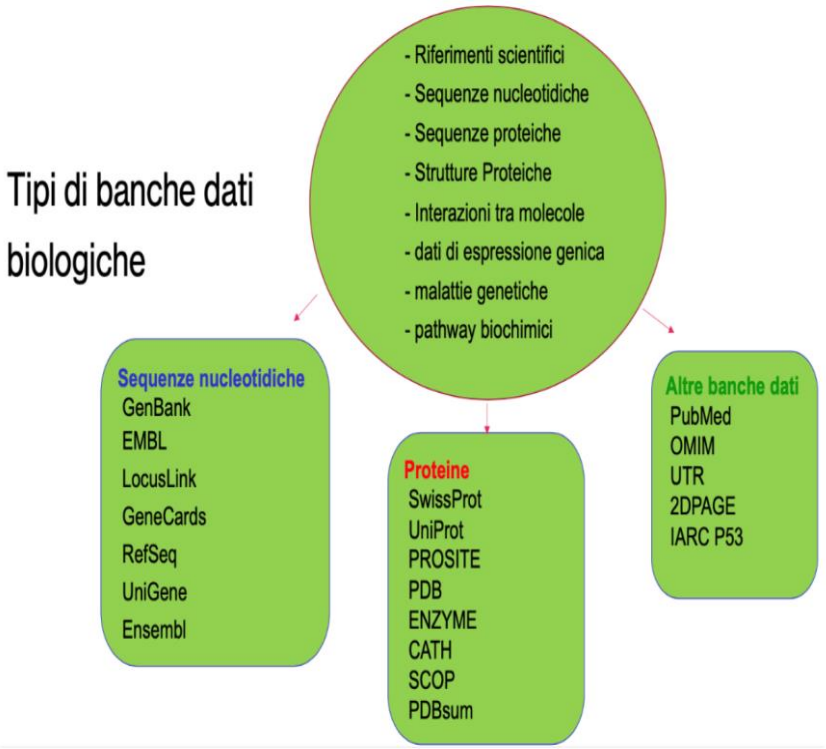
ENTREZ: È un sistema disponibile sul sito dell'NCBI (www.ncbi.nlm.nih.gov) per interrogare ed estrarre dati dalle più varie banche dati esistenti. Non è commercialmente disponibile e quindi non può essere scaricato e disinstallato localmente, né è possibile modificare le banche dati implementate sul sistema.

SRS- Sequence Retrieval System: Il nome può suggerire un uso limitato a "sequenze". In realtà è un sistema utilizzabile su qualunque tipo di database. Molti centri di ricerca hanno installato SRS sul proprio web server utilizzandolo per offrire un servizio di consultazione di banche dati. Uno dei sistemi SRS più curati è quello presente sul sito dell'EBI(www.ebi.ac.uk).

ESERCIZIO [Ogni sequenza ha un accession number].

1. Ricercare la sequenza nucleotidica che corrisponde all'Accession number M10051 in EMBL.
2. Ricercare in GenBank tutte le entry che corrispondono al termine "myoglobin"
3. Confrontare le entry nelle due banche dati se ricerchiamo in entrambe all'Accession number M10051

Insieme alla prof abbiamo guardato e cercato degli elementi sui vari siti per dimostrazione



Ci sono diversi formati, si confrontano i risultati per vedere se corrispondano. Se si apre NCBI si mette il nome del gene su research e si seleziona "nucleotide" si avrà la sequenza nucleotidica, se si selezionano altre informazioni si possono vedere varianti, gene, genoma etc. la modalità di ricerca è quindi molto semplice. Si scrive il nome del gene e poi si preme search.
NCBI da' moltissime informazioni; inoltre, si possono consultare le pubblicazioni su PubMed, e possiamo cercare i specifici geni. Se si ha a che fare con una mutazione bisogna comprendere di che si tratti e in cosa sia implicata, in quale proteina, che funzione abbia questa proteina e se essa sia correlata ad una malattia. Ci sono diverse banche su ogni argomento, quelle precedentemente citate sono quelle maggiormente usate.

Le banche dati non possono essere relegate ad una lezione: sono entità complesse e ognuna ha le sue peculiarità. Non è nemmeno possibile memorizzare i dettagli di utilizzo di ogni banca dati, perché vengono aggiornati periodicamente (se si cerca una determinata cosa, è possibile che tempo dopo non la si trovi più, perché è aggiornata).

La più usata è PubMed che, a differenza di google, è attendibile ed è il sito per eccellenza dove andare a fare una ricerca scientifica, che può essere tramite:

- autore (utilizzando il nominativo del ricercatore);
- rivista (o un articolo, alcuni dei quali scaricabili gratuitamente);
- parole chiave

Esercizio: cerca gli articoli che nel 2023 riguardino ALS (sclerosi laterale amiotrofica) su PubMed Dovranno uscire 56980 risultati, che vanno dal 1925 al 2023. Si possono selezionare gli abstract, o si può selezionare il tipo di articolo: ad esempio se si va su "clinical trial" risulteranno circa mille elementi.

Single Nucleotide Polimorfism (SNP) permette di vedere tutti i polimorfismi relativi a quel gene, tutte le varianti fino ad oggi trovate, in quali popolazioni e i gli impatti.

Esiste anche la Banca Dati dei Geni: se si inseriscono i geni (ad esempio SOD1) vengono rilasciate le relative informazioni. Il Genome Browser dà le coordinate, la posizione sul cromosoma e lo visualizza.
BLAST: permette di vedere se quel gene è allineato con lo stesso gene in altre specie
DOWNLOAD: permette di scaricare i risultati.

Dalla ricerca escono tutti i geni trovati, quindi nel topo, nell'uomo, nelle altre specie… a meno che non si metta un filtro selezionando solo quello che darà risultati di mio interesse (es homo sapiens)

dbVar: è lo studio delle varianti dell'organismo e tutte le regioni varianti, le coordinate della posizione sul genoma.

Un database molto utilizzato in ambito medico è OMIM, che è stato fondato dal DR McKusick e che contiene informazioni non solo su malattie genetiche di tipo autosomico, bensì anche su malattie associate ad alterazioni dei cromosomi sessuali e dei mitocondri. Anche questo database presenta dei cross link ad altre banche dati.
Si inserisce il nome della malattia o il suo specifico numero Omim e si può fare anche una ricerca avanzata avendo un riassunto sulle caratteristiche cliniche o una visualizzazione sulla mappa, in modo gratuito.

Per dati genomici abbiamo Genome data Viewer (non molto usato) che è sempre NCBI e che permette di vedere l'origine e l'evoluzione di un dato gene e se una mutazione è stata già scoperta e da chi e quando.

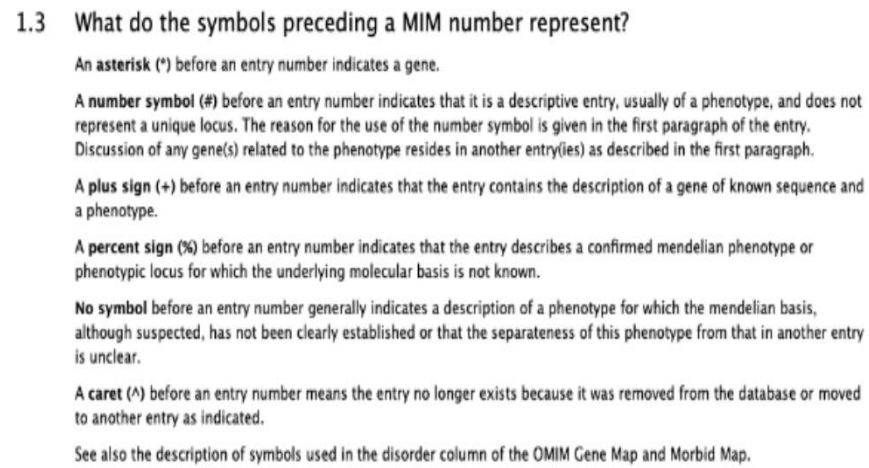
Esercizio:
è stata posta alla tua attenzione una famiglia con una rara condizione genetica denominata teleangectasia emorragica ereditaria (HHT). Usando internet come reperiresti le informazioni riguardo:

- Alle caratteristiche cliniche
- Al tipo di eredità
- Al nome del/i gene/i responsabile/i e la localizzazione cromosomica
- Ai laboratori che eseguono il test genetico
- A eventuali associazioni di pazienti

Bisogna cercare su OMIM che è il prediletto in ambito medico. Avremo una serie di entries (2351) e bisogna restringere il campo, aggiungiamo quindi informazioni. Ogni simbolo ha un significato:

- asterisco *: la malattia è associata ad un gene noto
- cancelletto #: si descrive un fenotipo
- percentuale %:mutazione mendeliana legata ad un locus

Le numerazioni indicano il tipo di ereditarietà.
Se si clicca su 2 #187300 indica il locus, la locazione sul cromosoma e la correlazione fenotipo genotipo ma, non essendoci l'* non si tratta di un gene noto.



1.3 What do the symbols preceding a MIM number represent?

An **asterisk** (*) before an entry number indicates a gene.

A **number symbol** (#) before an entry number indicates that it is a descriptive entry, usually of a phenotype, and does not represent a unique locus. The reason for the use of the number symbol is given in the first paragraph of the entry. Discussion of any gene(s) related to the phenotype resides in another entry(ies) as described in the first paragraph.

A **plus sign** (+) before an entry number indicates that the entry contains the description of a gene of known sequence and a phenotype.

A **percent sign** (%) before an entry number indicates that the entry describes a confirmed mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known.

No symbol before an entry number generally indicates a description of a phenotype for which the mendelian basis, although suspected, has not been clearly established or that the separateness of this phenotype from that in another entry is unclear.

A **caret** (^) before an entry number means the entry no longer exists because it was removed from the database or moved to another entry as indicated.

See also the description of symbols used in the disorder column of the OMIM Gene Map and Morbid Map.

La stessa ricerca si può fare ovviamente anche su PubMed dove avremo anche i risultati di letteratura.
Per vedere le caratteristiche cliniche, tipo di eredità nome dei geni responsabili etc. si può si Omim, vedere se c'è un locus (o dei loci), se c'è un gene responsabile.
Per quanto riguarda invece i laboratori che eseguono il test genetico, si utilizza di più Orphanet, che darà come risultato la spiegazione della patologia e i test genetici legati ad essa (in Italia o all'estero), utili anche per l'indirizzamento del paziente (verso ulteriori esami).

Esercizio

Un bambino di 8 anni presenta macrocefalia e disturbi comportamentali nell'ambito dello spettro autistico. L'analisi del cariotipo, così come l'analisi array CGH, è risultata normale. Quali database possono aiutarci nel formulare una diagnosi e attivare un percorso finalizzato alla conferma della stessa?
In questo caso la prima scelta ricade su OMIM perché fornisce nozioni cliniche, collegandosi alle malattie ad oggi conosciute. "Macrocefalia" + autism vengono inseriti nella barra di ricerca e danno come risultato una serie di elementi. Si può rifinire la ricerca, inserendo 'macrocefalia+ autismo' e ne risulteranno solo due entries, si evincerà che la sindrome sarà collegata ad un gene conosciuto, sul quale verrà fatta un'analisi su orphanet ad esempio o su face2gene (a pagamento), che utilizza anche delle fotografie, si può paragonare quindi anche l'aspetto morfologico del paziente con l'entry.

Compbio dà la possibilità invece di inserire tutte le peculiarità e i sintomi del paziente permettendo di capire a quale patologia potrebbero corrispondere. Questo database permette di selezionare i termini che identificano le caratteristiche cliniche principali del paziente, all'interno di una lista predefinita. (si può selezionare disfagia etc…) o selezionare (se noto) che la malattia ad esempio sia autosomica dominante.

https://compbio.charite.de/phenomizer/

Esercizio

Si rende necessario l'allestimento di una nuova metodica per effettuare un test genetico che confermi il sospetto clinico di neurofibromatosi 1. A tal fine, occorre recuperare le seguenti informazioni:

- Dimensioni della sequenza codificante il gene;
- Numero di esoni presenti nel trascritto più lungo o più abbondantemente espresso;
- Eventuale presenza di trascritti alternativi e la loro struttura;
- Identificazione dei domini funzionali della proteina codificata dal gene;
- Profilo di espressione del gene nei principali tessuti e organi.

Con Genome Browser possiamo ricercarlo.
Su NCBI possiamo cercare la sequenza codificante il gene, selezionando "gene" (NF1) darà l'annotazione e ci dirà che è posizionato sul braccio lungo del cromosoma 17, e la sequenza in cui è racchiuso, però bisogna capire la parte codificante. Per i trascritti ci si affida spesso alla sequenza più lunga perché è la più completa. Alcune sequenze sono degradate, quindi non formano la sequenza completa.

Un aspetto interessante dei database è poter vedere se queste sequenze, in questo caso nucleotidiche, se allineate, presentino omologia e in quale quantità; ci permette di individuare una specie filogenetica e di calcolare la distanza tra i geni e se, nelle specie che confronto, si presenti sempre un determinato dominio di una determinata proteina. Se la sequenza si ritrova sempre (è quindi conservata) vuol dire che è molto importante funzionalmente. Se un nucleotide cambia più spesso di un altro vuol dire che il gene tollera quei cambiamenti e quel tratto funzionale non è importantissimo, non è quello che da' funzione alla proteina e

non è conservato. L'assunzione fondamentale è che due sequenze identiche, formate dagli stessi elementi nella stessa successione, avranno un comportamento identico. Limitandoci alla sequenza nucleotidica in prendiamo in analisi altri fattori come elementi traduzionali.

CONFRONTO FRA SEQUENZE

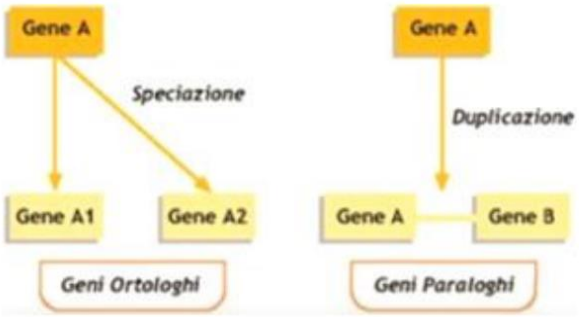
La funzione di una proteina è correlata ai meccanismi traduzionali, quindi non dipende esclusivamente dalla sequenza di nucleotidi, ma anche da come essa viene tradotta. È importante confrontare due sequenze biologiche perché ciò permette:

- La costruzione di alberi filogenetici (filogenesi molecolare);
- La ricerca di similarità in banche dati (evoluzione dei singoli genomi);
- L'identificazione di domini funzionali (caratterizzazione di proteine con funzione sconosciuta);
- La possibilità di identificare mutazioni responsabili di un fenotipo.

Se viene scoperta una sequenza, di cui non è mai stata fatta esperienza in precedenza, si procede con la ricerca di tale sequenza in altre specie. Ciò consente di capire che se la sequenza presenta una funzione specifica in un organismo molto probabilmente la manifesterà anche nel sistema (modello) in esame.

Similarità e Omologia

- Omologia: è la relazione filogenetica tra due sequenze (due entità condividono una stessa origine filogenetica), costituisce quindi un carattere qualitativo;
- Similarità: ha un significato più generico che indica una somiglianza nella composizione a prescindere dalle ragioni che la determinano, costituisce quindi un carattere quantitativo (quanti nucleotidi sono simili ad un'altra sequenza);



Spesso due sequenze omologhe hanno un elevato grado di omologia, ma possono divergere per mutazione ed evoluzione: possono restare omologhe anche se non troppo simili. Le regioni che tendono a restare simili sono quelle più importanti per l'attività della proteina.

Vi sono due tipi di omologia:

- due sequenze omologhe si definiscono ortologhe se appartengono a due specie diverse e il loro processo di divergenza ha avuto origine in seguito al processo di speciazione da cui le due specie suddette hanno avuto origine, ovvero se la sequenza originale da cui entrambe le sequenze omologhe derivano era presente nel più recente progenitore (definito cenancestore, oppure LUCA, Last Universal Commor Ancestor) da cui le due specie derivano.
  Derivano dunque da un unico gene ancestrale e sono geni che codificano per la stessa funzione in specie diverse. Per esempio Solv1 è presente sia nella specie umana che nel topo.
  Vi è anche una rappresentazione diversa: generalmente i geni ortologhi umani si indicano con la lettera maiuscola, mentre nelle altre specie sono in minuscolo;
- due sequenze omologhe si definiscono paraloghe se il loro processo di divergenza ha avuto origine in seguito a un processo di duplicazione genica. Sono presenti nello stesso genoma ma codificano per prodotti diversi.

Mediante confronti di geni simili tra genomi diversi, e di geni simili dello stesso genoma, si può stabilire se due geni sono ortologhi o paraloghi, e da qui ricostruire la probabile storia evolutiva.

Allineamento di stringhe

Cominciamo con l'affrontare il problema più generale dell'allineamento di una coppia di stringhe. Date due stringhe acbcdb e cadbd, in che modo possiamo stabilire quanto sono simili? La similarità scaturisce dall'allineamento ottimale delle due stringhe. Ecco un possibile allineamento:

a c - - b c d b

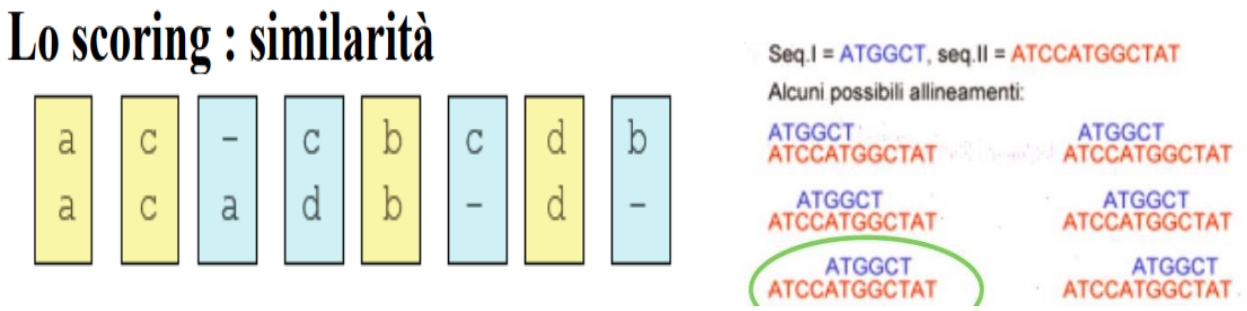
- c a d b - d -

Il carattere speciale "-" rappresenta l'inserimento di uno spazio, che sta a significare una cancellazione nella sequenza o, equivalentemente, un'inserzione nell'altra sequenza (Operazioni di INDEL). Quindi in generale questo vuoto indica che non c'è un similare nel compartimento complementare.

a c - c b c d b

- c a d b - d -

Similarità e distanza L'allineamento è fondamentale per valutare il grado di correlazione tra stringhe, poiché consente di calcolare la similarità o la distanza. Alta similarità → bassa distanza,
Bassa similarità → elevata distanza.



In generale è possibile valutare il grado di similarità o la distanza tra due stringhe, assegnando un punteggio (score) all'allineamento utilizzando un'opportuna scoring function.
Per esempio, se assegniamo un punteggio di +2 per ogni match esatto e un punteggio di -1 per ogni mismatch o indel, la similarità tra le due sequenze secondo l'allineamento considerato sarà:
S = 4*2 + 4*(-1)
Questa è la logica utilizzata da molti algoritmi di vari software, che prendono le sequenze dai database e le analizzano, allineano e ne studiano l'omologia.

Delle sequenze rappresentate nella figura a lato, qual è l'allineamento ottimale? Quella cerchiata in verde, perché contiene il maggior numero di lettere allineate. Di conseguenza, l'allineamento ottimale si ha quando matchano più coppie possibile.

Allineamento ottimale
Come trovare l'allineamento ottimale?
Il metodo più ovvio per determinare l'allineamento ottimale tra due sequenze consiste nel costruire tutti i possibili allineamenti e valutare quello con lo score più alto, tuttavia questo è un approccio impraticabile. Si consideri che allineare sequenze di appena 20 caratteri (lunghezza inusuale per una sequenza nucleotidica,

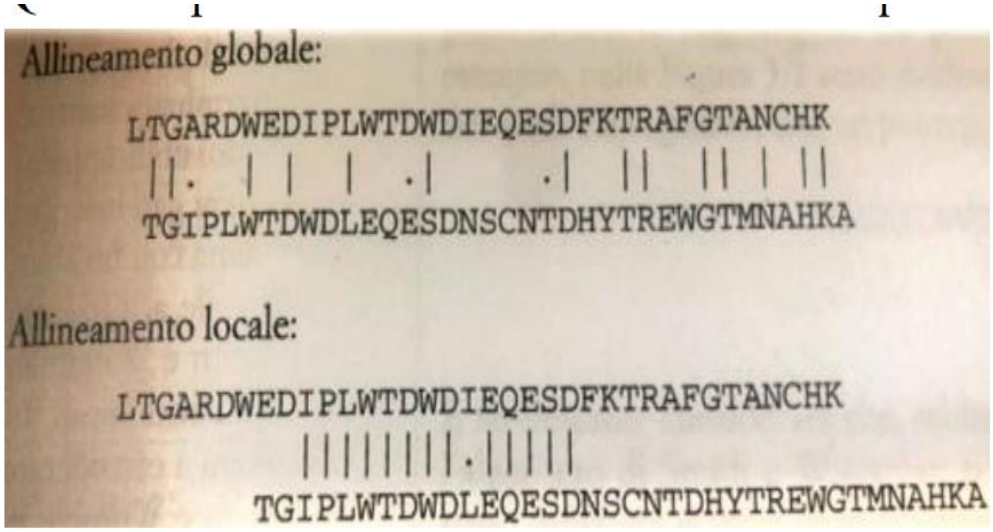
che solitamente è formata da un numero molto maggiore di caratteri) richiederebbe un tempo sicuramente inaccettabile.
La necessità di ricercare in banche dati di sequenze ha determinato l'esigenza di disporre di algoritmi rapidi di allineamento.
La crescita esponenziale delle dimensioni delle banche dati di sequenza ha portato allo sviluppo di programmi che velocemente effettuano ricerche di similarità grazie a soluzioni euristiche (basate su assunzioni non certe ma molto probabili).
Dunque, più gli algoritmi utilizzati sono veloci meno sono affidabili, infatti più sono veloci più la certezza del dato è minore. Questi database ci sono vari algoritmi utilizzati che non tratteremo in questo corso, verranno ad essere utilizzati quelli di default.

Tipi di allineamento
Quanti tipi di allineamento possono dare questi genome?
Possiamo avere un andamento:

- Pairwise: di due sequenze;
- Globale: si allineano le sequenze intere (una sequenza genomica di un gene e una sequenza genomica di un altro gene);
- Locale: si allineano sottosequenze delle sequenze di partenza (all'interno di un gene devo eseguire l'allineamento di una singola parte di esso);
- Multiple: di più di 2 sequenze;

Allineamento globale e locale

- Quanto sopracitato fa intendere che le stesse sequenze possono allinearsi diversamente: due sequenze Pairwise si possono allineare in modo globale o locale:



- Globale: cerca di estendere l'appaiamento delle basi lungo le intere sequenze;
- Locale: identifica regioni di similarità all'interno di sequenze che possono essere molto diverse.

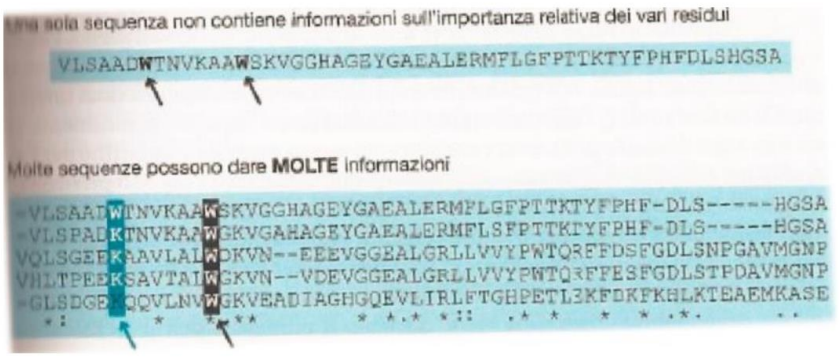
Esempio:

Due proteine che hanno sequenze diverse ma un dominio proteico in comune, usiamo l'allineamento locale; un dominio proteico è un esempio di 'lunga sottosequenza comune'.

Le lineette nella foto a lato contrassegnano i match di un allineamento globale, in questo caso la d la t matchano in numero maggiore.

In quella locale la stessa sequenza viene allineata per una sequenza interna; quindi, in realtà cerca di trovare l'allineamento ideale per una stringa più definita della sequenza.

Questo può essere utile quando è necessario identificare la presenza all'interno della proteina di una particolare sequenza funzionale identica oppure quando è necessario focalizzare l'attenzione su una parte specifica di un genoma.



Nell'immagine a lato vi è invece l'esempio di un allineamento multiplo. È presente una stringa di amminoacidi, di cui quello colorato è il triptofano. L'allineamento multiplo da

informazioni se è presente una stringa singola? No, una sequenza amminoacidica non da informazioni sufficienti, è necessario che vi siano più sequenze amminoacidiche. Per esempio, l'allineamento di molte sequenze amminoacidiche di specie diverse permette di affermare che la K (che dovrebbe essere la lisina) è conservata in tutte le specie (homo sapiens, drosophila, cavallo…).

Se in tutte le specie è presente la lisina mentre nell'essere umano nella stessa posizione ritroviamo il triptofano, la sequenza si dice non completamente conservata. Dall'altra parte, se nella posizione in esame compare lo stesso amminoacido in più specie, l'amminoacido è detto conservato(nelle diverse specie è sempre lo stesso). Ciò comporta che quella parte di sequenza è probabilmente così importante che la presenza di una mutazione, sconvolgerebbe la struttura e il funzionamento della proteina.

Ricapitolando: il fatto di trovare sempre lo stesso amminoacido è molto importante e lo deduciamo non da una singola stringa di amminoacidi ma dall'allineamento di molte stringhe amminoacidiche. Per essere informativo un allineamento deve contenere una distribuzione di sequenze che possono essere sia strettamente che lontanamente correlate.

Software di supporto all'analisi
Esistono diversi tools specializzati che automatizzano l'allineamento di sequenze, alcuni dei quali disponibili in rete:

- BLAST (integrato in GENEBANK), FASTA: ricerca di zone di omologia locale in coppie di sequenze;
- ALIGN: ricerca di allineamento globale fra due sequenze;

Software di supporto all'analisi
Esistono diversi tool specializzati che automatizzano l'allineamento di sequenze, alcuni dei quali disponibili in rete:

- **Allineamento Pairwise**:
  - EMBOSS;
  - BLAST;
  - FASTA;
  - ..
- **Allineamento Multiplo**:
  - CLUSTAW;
  - T-COFFEE;
  - ANTICLUSTAL;

https://www.ebi.ac.uk/Tools/msa/
Esempi di programmi che usano criteri euristici (FASTA, BLAST..)

BLAST

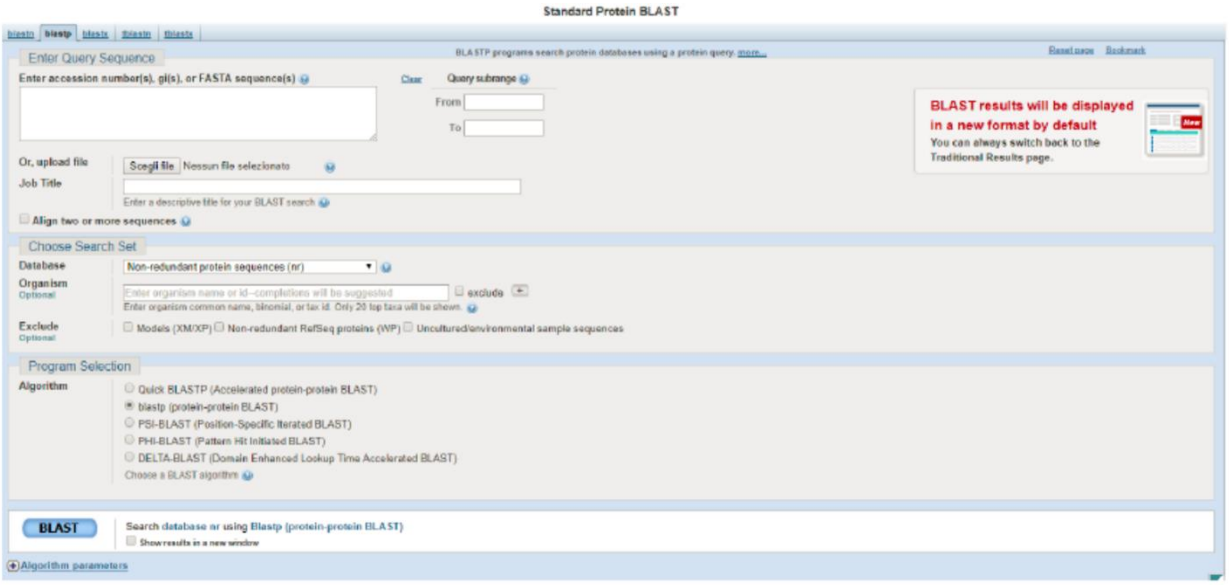
BLAST (**Basic Local Alignment Search Tool**), sviluppato dal National Center for Biotechnology Information, NCBI):

- **allineamento locale**;
- **estremamente veloce**;
- **parte cercando brevi frammenti della sequenza, che poi prova ad estendere;**
- **usa una matrice di sostituzione in entrambe le fasi del processo di allineamento** (scansione del database e estensione della subsequenza), più preciso ha quattro opzioni fondamentali:
    - BLASTP: confronta sequenze proteiche contro un database proteico;
    - BLASTN: confronta sequenze nucleotidiche contro un database nucleotidico,
    - TBLASTN: confronta una sequenza proteica contro un database nucleotidico, traducendo ciascuna sequenza del database nucleotidico nei suoi 6 frames di lettura;
    - BLASTX: confronta una sequenza nucleotidica contro un database proteico, dopo averla tradotta nei suoi 6 frames di lettura.

BLASTP

Questo è un esempio di BLASTP (il nome deriva appunto dal fatto che serve per allineare due proteine). Gli step per usare il sito sono.

- Cliccare e aprire il sito;
- Inserire la FASTA sequence della proteina umana e di quelle specie da analizzare a confronto;
- Impostare l'algoritmo da utilizzare, questo potrebbe essere cambiato da BLASTP (default) a BLAST o delta BLAST, sono algoritmi diversi che bisogna conoscere bene per poter essere utilizzati correttamente, conviene perciò optare per il default.



Allineamento multiplo_ClustalW

ClustalW è il tool più popolare per l'allineamento multiplo di sequenze.
Dato un insieme S di n sequenze da allineare, ClustalW allinea tutte le coppie di sequenze di S separatamente e costruisce una matrice con le distanze tra ogni coppia di sequenze.

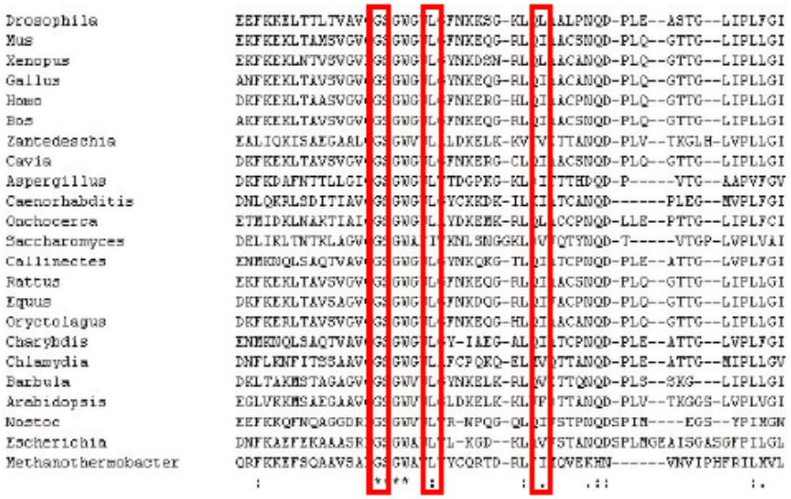
Il server ufficiale di clustalw si trova sul sito dell'EMBL: http://www.ebi.ac.uk/clustalw/index.html

Vi sono comunque molti altri server di clustalw; uno dei più popolari è quello dello Swiss Institute of Bioinformatics: http://www.ch.embnet.org/software/ClustalW.html
Questa versione di clustalw ha un'interfaccia semplificata rispetto a quella ufficiale su embl.

Quando è necessario vedere se un amminoacido o un cambiamento nucleotidico nel paziente è conservato o meno, si effettua l' allineamento ClustalW. Nella figura ne è riportato un esempio.

Vengono inserite una serie di sequenze amminoacidiche (posso scegliere se amminoacidiche o nucleotidiche) presenti in varie specie (drosophila …)
Le sequenze vengono confrontate e contraddistinte con i simboli infondo alle colonne:

- asterisco (*), indica un match del 100%, ciò significa che l'amminoacido in tutte queste sequenze è sempre presente e conservato;
- due punti (:), vuol dire che c'è un'alta similarità (>75%), probabilmente a seconda che la specie sia vertebrata o invertebrata.
- il punto (.), indica bassa o media similarità.



Il clustalW2 è stato ritirato quest'anno, una volta aperto il sito esso riconduce al nuovo Clustal Omega. Quest'ultimo guida alle procedure di allineamento e su come esso avviene. Il risultato può essere un allineamento come quello della foto sottostante. Queste procedure sono importanti perché evidenziano come si sono diversificate evolutivamente le sequenze di specie diverse, e per vedere se hanno origine comune o meno.