MASTER COPY
Tel. 388/9837745

21 marzo 2011

BANCHE DATI BIOLOGICHE (Ing. Meloni)

Slide 3

Le banche dati sono un insieme di dati ed informazioni che possono derivare sia dalla letteratura scientifica (quindi da tutti gli articoli di carattere biologico), sia da analisi effettuate in laboratorio (sia in vivo sia in vitro, quindi sulle cellule), oppure sono i risultati di tutta una serie di elaborazione di dati originari ottenuti attraverso la bioinformatica.
Nelle banche dati vi è un elemento biologico centrale che rappresenta l'oggetto fondamentale, attorno al quale vengono fatte tutte le operazioni di ricerca all'interno della banca dati. Ad esempio, nelle banche dati delle sequenze di DNA l'elemento centrale sarà la sequenza stessa di DNA.

Slide 4

La struttura della banca dati è definita da 2 elementi:

1. gli *attributi*, cioè le informazioni centrali;
2. Il *formato* con cui tutti questi attributi vengono salvati.

Ultimamente si sta diffondendo molto il formato .xml, perché i dati in questo formato sono facilmente scambiabili anche via internet. Tradizionalmente i dati delle banche biologiche sono nel formato "flat-file".
Il FLAT-FILE è un formato di tipo sequenziale in cui tutte le informazioni sono riportate su più linee consecutive, per ogni informazione vi è un cMASTER COPY identifica.
Si tratta di formati semplici che possono essTel. 388/9837745 i vari programmi.

Slide 5

Viene mostrata nella slide un tipico esempio di formato flat-file in cui vediamo nella sinistra tutte le varie voci (es: locus, ecc..), si tratta quindi di una banca dati di sequenza nucleotidica, e in fondo si vede la sequenza con tutte le basi azotate.

Slide 6

Esistono 2 tipi di banche dati principali:

- ***Primarie***: sono quelle un po' più semplici, che raccolgono l'informazione minima in grado di contestualizzare la "entry" e quindi di individuarla nel giusto contesto.
- ***Specializzate***: sono più dati che derivano dall'unione di più banche dati primarie, quindi oltre alla sequenza troviamo ad esempio la funzione di un determinato gene, informazioni di tipo funzionali e così via.

Slide 7

Esistono tantissimi tipi di banche dati:

- Di letteratura scientifica: una delle più importanti è "*pubmed*" in cui possiamo accedere a tutti gli articoli che ci sono con informazioni di tipo biologico o medico; in realtà solo alcuni articoli sono free.
- Di sequenza nucleotidiche: le più importanti sono la EBL e la Genbank.
- Di sequenze di proteine.
- Esistono tanti altri tipi di banche, non è necessario saperle tutte.

Queste banche dati servono affinché la comunità scientifica possa condividere e accedere a tutte le informazioni.

MASTER COPY

Tel. 388/9837745

Slide 8

All'interno delle banche dati si fanno delle ricerche per "chiavi": si interroga la banca dati inserendo in una finestra il nome che si vuole ricercare; ad esempio nel caso di sequenze di acidi nucleici si inserisce il nome della sequenza o della specie.

Nella ricerca si utilizzano le reti logiche AND, OR e NOT, questo consente di rendere la ricerca più precisa.

Vi sono dei sistemi che consentono di ricercare l'argomento di interesse all'interno del database, i più conosciuti sono "SRS" e "entrez".

Slide 9

Nella slide viene mostrata la pagina iniziale di SRS (http://srs.ebi.ac.uk/). SRS è un sistema molto usato in quanto consente di interrogare più banche dati contemporaneamente, e di navigare attraverso varie banche dati.

Slide 10-11

ENTREZ (http://www.ncbi.nlm.nih.gov) è molto famoso in quanto è associato a PubLine, quindi a PubMed, ovvero alla ricerca della letteratura scientifica. Diversamente da SRS, è un sistema chiuso, quindi non consente di scaricare da internet oppure di duplicare su altri computer.

Nella slide viene mostrata la pagina iniziale di ENTREZ.

Slide 13

Vediamo adesso cos'è l'allineamento di sequenze. Una sequenza è un insieme di simboli dove i simboli fanno parte di un determinato alfabeto, il quale dipende dal tipo di sequenza. Ad esempio se si tratta di sequenze di DNA, i simboli saranno le 4 basi azotate ATCG, se si tratta di sequenze di RNA saranno AUCG, se si tratta di sequenze proteiche i simboli saranno i 20 amminoacidi.

Slide 14

L'operazione più semplice da fare con le sequenze è chiedersi se tra 2 sequenze esiste una qualche similarità, ovvero una somiglianza che non sia di tipo casuale ma che abbia un certo significato biologico. Questo è importante per scoprire delle informazioni funzionali delle sequenze stesse; se ad esempio troviamo una sequenza di un gene simile alla sequenze di un altro gene di cui conosciamo già la funzione, possiamo ipotizzare che la funzione del nuovo gene sia abbastanza vicina.

Per valutare la somiglianza tra 2 sequenze o tra 1 parte di essa (infatti spesso le sequenze possono essere infinitamente lunghe, quindi spesso si prendono in considerazione solo delle parti di interesse e si cercano delle parti fortemente conservate) si usa l'ALLINEAMENTO. Poi si decide se questa similarità (o somiglianza) ha qualche significato biologico oppure se è dovuto al caso.

Slide 15

Cosa vuol dire allineare 2 sequenze o 2 stringhe?

Per allineare 2 sequenze non è fondamentale che abbiano la stessa lunghezza, possiamo avere due sequenze con numeri di simboli differenti. Allineare 2 sequenze dal punto di vista algoritmico vuol dire trovare quella funzione che vada a minimizzare la distanza tra le 2 sequenze, oppure a massimizzarne la somiglianza. Questa funzione sarà una qualche trasformazione più o meno complicata che può essere applicata alla sequenza stessa.

In pratica nell'allineamento ci sono 3 fasi fondamentali:

- Scegliere una metrica di similarità da massimizzare o di distanza da minimizzare;

- Definire quali sono le trasformazioni possibili, si parla in questo caso di "search space", che è lo spazio possibile di tutte le possibili trasformazioni; si possono usare sia trasformazioni molto semplici (es: inserzione di 1 base azotata o la cancellazione) oppure operazioni + complicate;
- Bisogna definire l'algoritmo che consenta di massimizzare questa funzione di similarità, quindi l'algoritmo di ottimizzazione.

MASTER COPY

Tel. 388/9837745

Slide 16

Esistono 2 tipi generali di allineamento:

- *Globale*: si cerca di allineare il max numero di caratteri delle 2 sequenze, ed è molto più semplice se le 2 sequenze hanno lunghezze abbastanza simili;
- *Locale*: si cercano di allineare delle parti di sequenze, si cercano all'interno delle 2 sequenze quelle sottosequenze particolarmente conservate in entrambi, e non ha importanza se le sequenze hanno lunghezze diverse oppure no.

Si distingue anche tra:

- *Allineamento a coppie*: si hanno solo 2 sequenze, per cui è molto più semplice, o altrimenti si effettuano prima degli allineamenti a coppie e poi pian piano si estende. Esistono sia degli algoritmi di programmazione dinamica sia delle tecniche euristiche (es. algoritmi FASTA e BLAST che permettono di trovare le parti conservate nelle 2 sequenze).
- *Allineamento multiplo*: si hanno varie sequenze.

Slide 17

La DISTANZA tra le 2 sequenze deve essere simmetrica e sarà:

- Pari a 0, nel caso di sequenze fortemente simili tra di loro;
- Tanto maggiore quanto più le sequenze sono diverse tra di loro.

A prescindere dalla metrica usata, le distanze vengono normalizzate per la lunghezza al fine di limitare la dipendenza dalla lunghezza delle sequenze stesse.
La distanza sarà SEMPRE non negativa e simmetrica, questo significa che se io ho 2 sequenze A e B, è indipendente se valuto la distanza A-B oppure B-A.
Un tipo di distanza usata è la "DISTANZA DI HAMMING": si calcola il numero di posizioni in cui le 2 stringhe differiscono tra di loro. Viceversa, se considero la similarità di Hamming conterò il numero di coppie uguali, per cui metterò le due sequenze vicine (una sotto l'altra) e vado a vedere quante volte la A è presente nella sequenza sotto e in quella sopra, e viceversa.

Slide 18

Ad **esempio** se ho le 2 sequenze A e B, devo contare il numero di simboli che si appaia esattamente, quindi considero la similarità di Hamming:

A = A A K K Q W **B = A A K Q W**

io metto le sequenze vicine e conto quante volte la A (quella dentro la sequenza) è presente nella sequenza sotto e in quella sopra e viceversa.
In pratica faccio scorrere la 2° sequenza sulla prima ogni volta di una posizione:

AAKKQW AAKQW	AAKKQW AAKQW	AAKKQ W AA KQW	AAKKQW AAK QW
AAKKQW AAKQW	AAKKQW AAKQW	AAKKQW AAKQW	AAKKQW AAKQW
AAKKQW AAKQW	AAKKQW AAKQW		

La 1° riga in alto a sinistra è il punto di partenza; poi faccio scorrere ogni volta di un punto e ogni volta che scorro conto le coppie che sono uguali; l'allineamento ottimo si ottiene qui perché ho 4 simboli uguali; per ottenere l'allineamento ottimo ho fatto 4 shift della 2° sequenza sulla prima, quindi verrà dato un punteggio pari a 4 che poi verrà normalizzato, per cui quello che vado a cercare è sempre la similarità massima.

Slide 19

Posso rappresentare tutto ciò anche graficamente con la "DOT MATRIX"(vedi slide). In pratica costruisco una matrice NxM, dove N e M sono il numero di simboli della stringa, e metto una sequenza lungo le righe della matrice e l'altra lungo la colonna, poi do il valore di 1 per le coppie nelle caselle che corrispondono ai simboli uguali nelle due sequenze.

Poi posso anche graficare questa dot matrix e ciò che mi appare sono delle righe nere in corrispondenza di coppie di basi uguali. Quindi si avrà una diagonale abbastanza marcata e dei trattini che magari sono dovuti al caso, perché c'è qualche base spaiata uguale tra le 2 sequenze.

Slide 20

Raramente una persona usa la distanza di Hamming per una ricerca di tipo esaustiva, perché è possibile solo nel caso in cui si ha un numero basso di stringhe, cosa che nella realtà non capita mai, in quanto si lavora con stringhe molto lunghe.

Di solito, si usano delle ricerche di tipo NON ESAUSTIVO con algoritmi con costo computazionale meno complesso. Questa tecniche hanno il vantaggio di velocizzare l'operazione ma non è detto che si trova la soluzione ottima. Di solito quindi si usa un buon compromesso tra velocità e complessità computazionale.

(La prossima lezione faremo una esercitazione sulla distanza di Hamming e sulla dot matrix)

Adesso vedremo come accedere da internet ad 'SRS' ed 'entrez. In questo lezione faremo fino all'esercizio 5, gli altri esercizi li faremo la prossima lezione. Per andare sul sito di 'entrez' possiamo direttamente scrivere il nome su google, e lui ci dà direttamente il sito!

ESERCIZIO 1

1)

Ricerca di una sequenza di nucleotidi: selezionare Nucleotide dal menu a tendina (in alto a sinistra - Search) e digitare nella finestra di testo:"proinsulin" e cliccare su GO. Troviamo N entry.

- Per restringere la ricerca, attiviamo il tasto limits. Settiamo "limits to:" a ORGANISM, e cerchiamo "homo". Individuando così le entry realtive al genoma umano. Troviamo M entry.

- Combiniamo le due query. Andiamo su history e identifichiamo le due precedenti. Supponendo che siano la #1 e la #2 cerchiamo #1 AND #2.

- I risultati ottenuti sono L.

Come faccio a capire che 2 risultati ottenuti con strade diverse sono uguali?
Parti dal presupposto che il nome del gene può cambiare, stessi geni hanno nomi diversi e uno stesso nome può essere assegnato a 2 geni; il modo + facile per farlo è usare l' accession number, che è quello che identifica sempre in maniera univoca.

RISULTATI
Nucleotide:proinsulin->N entry=223;
proinsulin+homo->L=42.

2)
Sito SRS (http://srs.ebi.ac.uk)
- Cliccare su Start project per accedere a SRS6 ed inserire un nome qualsiasi.
- Selezionare dalla lista dei database (database banks o libraries) EMBL e cliccare sul tasto "Query form" in alto a sinistra.
- Inserire nelle caselle di testo le parole chiave per la ricerca:

- *nel primo campo selezionare dalla lista "Description" e digitare "proinsulin" (notare la differenza nell'operatore booleano rispetto all'NCBI & al posto di AND);*
- *nel secondo campo selezionare dal menu "Organism" e scrivere "homo". - Impostata la query inviarla cliccando sul tasto "submit query" in alto a sinistra*

Si otterranno 2 risultati.

3)
Confrontare le informazioni ottenute dalla query all'NCBI e quella all'EBI. Notiamo che la entry E00191 e AY899304 sono presenti nei risultati di Entrez. Il motivo per cui Entrez fornisce più risultati è che cerca automaticamente anche in campi non strettamente collegati alla entry fondamentale, ad esempio nella entry NM_000602 il termine proinsulin viene trovato nel titolo di alcune pubblicazioni citate. Si può selezionare refSeq che individua le sequenze di riferimento.

RISULTATO: http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz

ESERCIZIO 2

Su Entrez, andare in "all databases" e cercare il gene HIF1A. Il gene verrà trovato in numerosi database. Premendo "?" è possibile avere informazioni sui vari database.
Andare su "nucleotide" e cercare la entry NM_001530.
Nel campo display selezionare FASTA (formato compatto letto da MATLAB e da altri programmi).
Salvare la sequenza in formato FASTA come file. Notare anche la presenza del formato XML.
Utilizzare il comando "Graph" per visualizzare graficamente la sequenza.
Tornare al formato originario (GenBank). Il tasto Link->related sequences mostra le sequenze collegate. Il tasto Link->map viewer mostra la posizione del gene sul cromosoma.

Per salvare in formato FASTA bisogna andare sul MENU→ SEND→ FILE.

ESERCIZIO 3