

Una proteina nella rete:

Introduzione alla bioinformatica

L'era genomica ha assistito ad una crescita esponenziale delle informazioni biologiche rese disponibili dai progressi nel campo della biologia molecolare. In particolare, il sequenziamento del genoma umano e di altri organismi ha dato un forte impulso a quel settore della bioinformatica che si occupa dello studio del DNA e delle proteine. La grande sfida che la comunità scientifica sta ora affrontando consiste nel cercare di analizzare e capire l'enorme quantità di dati prodotta in laboratorio

La bioinformatica è una nuova disciplina che si occupa dello sviluppo e dell'integrazione delle applicazioni della scienza dell'informazione al servizio della ricerca scientifica in campo biotecnologico. Per fare ciò utilizza strumenti informatici per analizzare i dati biologici che descrivono sequenze di geni, composizione e struttura delle proteine, processi biochimici nelle cellule, etc.

Definizione di BIOINFORMATICA :

(da **Wikipedia**, l'enciclopedia libera) <http://it.wikipedia.org/>

Disciplina ultima arrivata nel campo delle bioscienze, la **bioinformatica** costituisce l'ambizioso tentativo di descrivere dal punto di vista numerico e statistico i fenomeni biologici: storicamente ed epistemologicamente la biologia ha sempre sofferto di una carenza in tal senso rispetto a discipline come la fisica e la chimica, ma oggi la bioinformatica tenta di supplire a questa lacuna fornendo ai risultati tipici della biochimica e della biologia molecolare un corredo di strumenti analitici e numerici davvero promettente.

La bioinformatica principalmente si occupa di:

- * fornire modelli statistici validi per l'interpretazione dei dati provenienti da esperimenti di biologia molecolare e biochimica al fine di identificare tendenze e leggi numeriche
- * generare nuovi modelli e strumenti matematici per l'analisi di sequenze di DNA, RNA e proteine la fine di creare un corpus di conoscenze relative alla frequenza di sequenze rilevanti
- * organizzare le conoscenze acquisite a livello globale su genoma e proteoma in basi di dati al fine di rendere tali dati accessibili a tutti, e ottimizzare gli algoritmi di ricerca dei dati stessi per migliorarne l'accessibilità.

Banche dati

Una delle attività principali dei bioinformatici consiste nella progettazione, costruzione e uso di **banche dati** di interesse biologico. Una banca dati raccoglie dati e informazioni derivati da esperimenti di laboratorio, da esperimenti *in silico* (cioè utilizzare il dato informatico come punto di partenza per gli esperimenti *in vitro*. Si dice "in silico", in quanto i processori dei calcolatori sono costituiti da silicio) e dalla letteratura scientifica. Le banche dati sono progettate come contenitori costruiti per immagazzinare dati in modo efficiente e razionale al fine di renderli facilmente accessibili a tutti gli utenti: ricercatori, medici, studenti, etc.

Una banca dati è costituita da voci (in inglese *entry*) ciascuna contenente informazioni sull'oggetto caratteristico della banca dati (ad esempio: sequenze nucleotidiche o referenze bibliografiche) insieme a tutte le altre informazioni che si riferiscono a quella entry in particolare).

Una *entry* di una banca dati di sequenze nucleotidiche potrebbe contenere, oltre alla sequenza di una molecola di DNA, il nome dell'organismo cui la sequenza appartiene, la lista degli articoli che riportano dati su quella sequenza, le caratteristiche funzionali (cioè si tratta di un gene o di una sequenza non codificante) e ogni altra informazione ritenuta di interesse.

Esempio di banca dati: la rubrica telefonica

Ognuno di noi ha esperienza di banche dati come le rubriche telefoniche. In una rubrica telefonica, una *entry* ha come oggetto principale il numero di telefono di uno dei nostri amici o parenti.

La nostra rubrica sarebbe totalmente inutile se insieme ai numeri di telefono non catalogassimo anche il nome e il cognome del possessore del numero di telefono

La nostra rubrica potrebbe essere arricchita anche con altre informazioni: l'indirizzo della persona (di casa e di lavoro), la sua occupazione (se non ci ricordassimo il nome dell'idraulico, dovremmo provare tutti i numeri della rubrica mentre la casa si allaga!!!!), il suo compleanno.

Una banca dati potrà di conseguenza apparire come un elenco di *righe* o come un insieme di *tabelle*

Bruno Macchi | dentista | via Calandrino 27 | 02-72597259
Carla Cecioni | autista | piazza Crati 45 | 02-68686868
Dante Alighieri | poeta | via Monti 35 | 02-41563444

NOME	Ercole Palestri
NOME	Dante Alighieri
NOME	Carla Cecioni
NOME	Bruno Macchi
LAVORO	dentista
INDIRIZZO	via Calandrino 27
TELEFONO	02-72597259

Tipi di banche dati: primarie e specializzate

Le banche dati possono essere di due tipi: primarie o specializzate.

Le **banche dati primarie** contengono informazioni e annotazioni delle sequenze nucleotidiche e proteiche, strutture del DNA e proteine e dati sull' espressione di DNA e proteine.

Le principali banche dati primarie sono: la **EMBL** datalibrary, la **GenBank** e la **DDBJ**. La EMBL datalibrary è la banca dati europea costituita nel 1980 nel laboratorio Europeo di Biologia Molecolare di Heidelberg (Germania). La GenBank è la corrispondente banca americana costituita nel 1982 e la DDBJ è la corrispondente Giapponese. Fra le tre banche dati è stato stipulato un accordo internazionale per cui il contenuto dei dati di sequenza presenti nelle tre banche dati è quasi del tutto coincidente in quanto gli aggiornamenti quotidiani apportati in ciascuna banca dati vengono automaticamente trasmessi alle altre due.

Le **banche dati specializzate** si sono sviluppate successivamente e raccolgono insieme di dati omogenei dal punto di vista tassonomico e/o funzionale disponibili nelle Banche dati Primarie e/o in Letteratura, o derivanti da vari approcci sperimentali, rivisti e annotati con informazioni di valore aggiunto.

Una volta che i dati sono stati archiviati nelle banche dati biologiche è necessario utilizzare alcuni strumenti bioinformatici in modo tale da ricavarne informazioni. Essi si sono sviluppati in base a questi tre processi biologici fondamentali:

- la sequenza del DNA determina la sequenza aminoacidica della proteina (mediante il processo della sintesi proteica);
- la sequenza aminoacidica determina la struttura tridimensionale della proteina;
- la struttura tridimensionale della proteina ne determina la funzione.

La bioinformatica ha focalizzato la sua analisi su dati relativi a questi processi, e di conseguenza le banche dati costituiscono un potente supporto per una vasta gamma di ricerche quali, ad esempio:

- data una sequenza di acidi nucleici o proteica trovare una sequenza simile in banca dati;
- data una struttura proteica trovare, in banca dati, una struttura simile ad essa;
- data una sequenza proteica prevedere una possibile struttura tridimensionale.

I principali strumenti possono essere così organizzati:

Ricerca di sequenze simili

Sequenze omologhe sono sequenze che hanno un gene ancestrale comune. Il grado di similarità fra due sequenze può essere misurato mentre l'omologia è un dato qualitativo.

Esistono una serie di strumenti (es **BLAST**) che possono essere utilizzati per identificare similarità fra nuove sequenze con funzione e struttura sconosciuta e sequenze (archivate nelle banche dati) la cui struttura e funzione sono note.

Studio delle funzione delle proteine

Questo gruppo di programmi (es. **PROSITE**, **SMART**) permette di utilizzare una sequenza per estrarre informazioni su *motif*, domini strutturali dalle banche dati specializzate. Questo potrebbe essere di aiuto per avere informazioni sulla funzione della proteina ignota.

Analisi delle strutture

Questi strumenti permettono di comparare una struttura con una banca dati di strutture note. Molto spesso proteine con struttura simile hanno una stessa funzione, quindi determinare la struttura secondaria/terziaria è cruciale per capire la funzione. (es. **EBI-MSD**)

Analisi della sequenza primaria

Identificare/analizzare l'evoluzione, identificare mutazioni, regioni idrofobiche o altre proprietà che permettano di capire la funzione della proteina. (es. **ENSEMBL**)

Principali applicazioni della bioinformatica

Numerose possono essere le applicazioni della bioinformatica. Qui citeremo solo un aspetto della **medicina molecolare**. Si ritiene che molte malattie siano associate ad una componente genetica. La malattia, infatti, può essere ereditaria (sono note circa 3000-4000 malattie genetiche come la fibrosi cistica, alcune forme di diabete, etc) oppure essere il risultato di fattori ambientali che causano alterazioni del genoma (tumori, malattie cardiache, ecc). Una branca della bioinformatica studia quali geni siano associati a diverse malattie per capirne più chiaramente le basi molecolari con lo scopo di migliorarne la prevenzione e la cura.