

BIOINFORMATICA

Indice

1	Che cos'è la bioinformatica	3
2	Quali sono gli obiettivi?	3
3	Le banche dati biologiche	4
4	Bioinformatica ed evoluzione	5
5	Modellizzazione proteica	6
6	Assegnare nuove funzioni ai geni	6
7	Bioinformatica e microarray a DNA	7

Introduzione

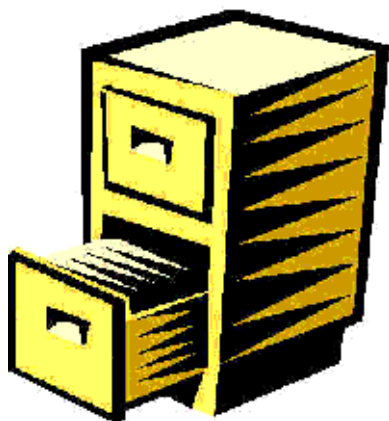
La facilità con cui è possibile "leggere" i genomi ha portato ad una crescita esplosiva delle informazioni a nostra disposizione: la bioinformatica permette di mettere un pò di ordine e di estrarre dal caos le informazioni rilevanti per capire come funzionano il nostro genoma e le nostre cellule.



Negli ultimi cinque anni il numero di informazioni a disposizione sul *genoma* è aumentato verticosamente: si parla ormai di miliardi di informazioni, con una curva di crescita che diventa sempre più verticale. Le informazioni arrivano da diversi fronti: i progetti genoma continuano a riempire le banche dati informatiche con informazioni relative alle sequenze di DNA di diversi organismi viventi, mentre i ricercatori utilizzando queste informazioni elaborano esperimenti che portano informazioni aggiuntive. Tra queste si annoverano le relazioni tra i vari geni, che hanno fatto comprendere che tutto ciò che accade all'interno delle nostre cellule è un lavoro complesso di equipe: molti geni collaborano per ogni singola funzione. Per scovare queste relazioni complesse i ricercatori incrociano le informazioni ottenute in laboratorio con quelle presenti nelle banche dati genetiche, dalle risposte che ottengono formulano nuove teorie che verificano sperimentalmente. In questo modo i tempi della ricerca accelerano: quando studia un nuovo gene il ricercatore ha già a disposizione molte

informazioni che lo indirizzano. Per esempio, facendo ricerche nelle banche dati può rispondere a domande del tipo: esistono geni simili in altri animali? quali geni si attivano quando il gene in questione si attiva? e quali si spengono? al suo interno sono presenti parti già studiate in altri geni di cui è nota la funzione? Elaborare queste informazioni richiede un enorme potenza di calcolo (solo considerando il genoma umano, tutte le volte che si fa un'analisi si devono considerare almeno tre miliardi di *basi*), che oggi per fortuna abbiamo: sono i computer, che nei laboratori di bioinformatica, spesso collegati in serie per aumentare ulteriormente il potere di calcolo, aiutano il ricercatore a ricomporre il filo del discorso nascosto nel nostro genoma. La disciplina che ha permesso di avere a disposizione strumenti e metodi per questi studi si chiama **bioinformatica**. La possibilità di analizzare grandi moli di dati e sistemi complessi ha posto le basi per la nascita di una nuova disciplina, chiamata **biologia dei sistemi**, cioè lo studio delle interazioni che avvengono tra le diverse componenti all'interno delle cellule. In pratica, invece di studiare una molecola per volta, si studia il comportamento di interi complessi (*vedi il dossier sulla biologia dei sistemi*)

3 Le banche dati biologiche

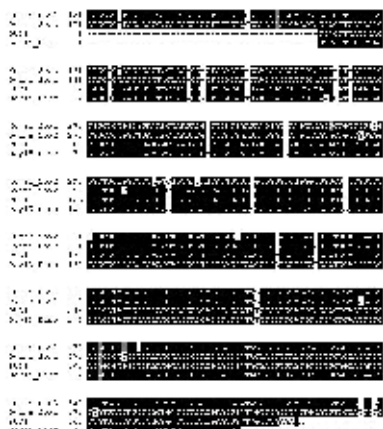


Organizzare le informazioni in modo che siano facilmente reperibili ed analizzabili è una delle principali problematiche della bioinformatica. Il secondo problema è che questo deve poter essere fatto da tutti, quindi devono essere sistemi di archiviazioni accessibili via web. Attualmente tutti i dati sui genomi sono conservati nei *server* di strutture diverse, che ogni notte sincronizzano le informazioni: l'EMBL in Europa (si trova ad Heidelberg, ed è stata la prima ad essere fondata nel 1981), la GenBank negli Stati Uniti (1982) e la DDBJ in Giappone. Infine, oltre ad essere consultabili ed analizzabili deve essere possibile implementarli continuamente, visto che i dati a disposizione aumentano ogni giorno in modo considerevole. In pratica, è come trovarsi in una biblioteca in cui arrivano centinaia di libri al giorno e doverli archiviare. E' indispensabile avere lo spazio dove metterlo e un archivio informatico su cui segnare dove sono. La prima operazione è stabilire quali parametri vogliamo raccogliere nel nostro archivio: titolo del libro, autori, editori, anno di pubblicazione, genere... oltre ovviamente alla posizione nella biblioteca. Quindi costruiamo una tabella dove iniziamo a raccogliere i dati. Visto che la tabella la costruiamo al computer, può essere ordinata ogni volta che vogliamo secondo la colonna che ci interessa: per esempio in ordine alfabetico secondo il nome dell'autore. Oppure possiamo ricercare il nostro libro inserendo la parola chiave nel motore di ricerca del software che stiamo usando. In realtà, dovremmo anche

decidere come organizzare nello spazio i libri: secondo il genere? secondo la tipologia? Prima la tipologia e poi il genere? Creare, cioè, una specie di classificazione a più livelli. Questo è esattamente il problema che hanno i ricercatori quando ottengono i dati dalle analisi genetiche, solo che invece del "titolo del libro" considerano parametri come "nome del gene", "posizione sul cromosoma", "specie in cui è stato rilevato," chi e come l'ha descritto". Li inseriscono quindi in un archivio che viene chiamato **banca dati biologica** (genetica se è composta da sequenze di DNA, proteica se è composta da sequenze di amminoacidi). In realtà c'è una complicazione in più. Finché si trattava di accumulare semplicemente le sequenze non era un problema, ma da quando a queste sono state associate funzioni diverse in ambiti diversi (un gene dentro una cellula del cervello può avere una funzione diversa da quella che ha dentro una cellula del fegato) è insorto anche il problema di catalogare per parole chiave. Ma, a differenza della nostra biblioteca, dove c'è uno spazio fisico di riferimento, sui computer il problema non si pone: ad ogni gene sono associate più parole chiave relative al loro coinvolgimento in specifici progetti: è come se potessero stare su più scaffali contemporaneamente. I ricercatori possono cercare i geni sulla base delle "parole chiave", oppure, al contrario, possono cercare di capire con quali altri geni condividono determinate parole chiave, nell'ipotesi che lavorino insieme. Con le banche dati è possibile individuare tutte le molecole che rispondono a determinati criteri, per esempio: tutti i geni che contengono la sequenza ATGGTTGATTTATATGCG-CATTCAGCGATCTATC, appartengono alla specie *Homo sapiens* e sono coinvolti nell'infiammazione. Otteniamo così un lungo elenco di geni, di cui possiamo consultare la singola scheda o usare come base di partenza per altre ricerche. Ovviamente questa è una semplificazione dell'intero processo che è molto più complesso, soprattutto per via della mole di dati considerati, ma intuitivamente rende l'idea di cos'è una banca dati genetica: non è altro che un archivio consultabile e implementabile dove sono raccolte le informazioni su singole parti dei genomi e sulle proteine. Questo archivio può

essere consultato direttamente dal ricercatore per cercare informazioni specifiche. Oppure, può essere consultato attraverso software che effettuano analisi su grande scala, come un confronto tra l'attività di 100 mila geni dopo l'esposizione ad un farmaco osservata sperimentalmente al fine di classificarli in gruppi sia sulla base dell'attività misurata che sulla catalogazione presente in banca data (per esempio: geni attivati e coinvolti in processi infiammatori, geni che non si sono attivati e coinvolti nella divisione cellulare, geni che si sono attivati molto e che sono coinvolti nella detossificazione da farmaci). Dai risultati, è possibile avere una prima idea sull'effetto del farmaco: per esempio, se ha inibito i geni coinvolti nell'attivazione, senza attivare eccessivamente i sistemi genici che indicano intossicamento cellulare.

4 Bioinformatica ed evoluzione



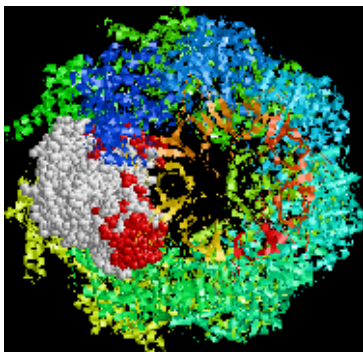
Grazie ai mezzi messi a disposizione dalla bioinformatica, oggi i ricercatori possono letteralmente leggere la storia della nostra evoluzione direttamente nel genoma. Nel paragrafo precedente abbiamo accennato al concetto di **omologia** tra i geni. In biologia, il termine omologia è utilizzato per indicare strutture anatomiche che sono simili tra animali diversi e hanno, in genere, una funzione analoga, in virtù di una loro origine comune. Per esempio, si ritrovano ossa omologhe a quelle della mano dell'uomo negli arti di altri animali, degli uccelli e nelle pinne delle balene. Lo stesso si verifica nei geni:

in organismi diversi si trovano geni simili che hanno funzioni simili. La definizione è un po' più sofisticata rispetto a quella usata nella biologia tradizionale, perché si pongono dei limiti statistici: cioè, solo se si rientra entro certi parametri percentuali di uguaglianza i geni sono considerati omologhi. E in quel caso è ragionevole pensare nella maggior parte dei casi che l'omologia sia dovuta ad un'origine comune. Per verificarlo, i ricercatori costruiscono alberi genealogici cercando gli omologhi di un certo gene in tutte le specie e confrontando le differenze. Se i geni hanno un'origine comune è probabile che siano più simili tra organismi evolutivamente più vicini. Per esempio, il gene della globina - una piccola proteina globulare - è presente in tutte le specie viventi, dai batteri all'uomo. Ma è più simile tra l'uomo e lo scimpanzé che tra l'uomo e la mosca. Questa analisi di similarità sono fatte con appositi software che calcolano la somiglianza tra sequenze diverse. Questa può essere calcolata sulla base di diversi parametri: variazioni nella sequenza che si ripercuotono nella sintesi di proteine con uno o più amminoacidi diversi, oppure variazioni silenti, cioè che non modificano la proteina; variazioni che comportano il distanziamento di alcune parti di geni e altre modifiche importanti come la duplicazione o la perdita di parti.

Un'applicazione molto importante per la salute umana in questo settore è lo studio dell'evoluzione dei ceppi virali: l'influenza, per esempio, è studiata in modo molto approfondito dal punto di vista della genetica, confrontando i vari ceppi e cercando di individuare andamenti che permettano di prevedere le future evoluzioni.

La disciplina che si occupa di questi studi si chiama **filogenetica**.

5 Modellizzazione proteica



I bioinformatici guardano le proteine come se fossero composte di porzioni componibili. Infatti, se si considera un pool abbastanza ampio di proteine si scopre che in esse sono riconoscibili parti che si ripetono, assemblate in modo diversi, in altre proteine. E' un pò come trovarsi davanti ad una scatola di mattoncini per costruzioni: ci sono quelli a forma quadrata, a forma di tegola, rettangolari, quelli con un buco per inserire un ingranaggio, con un perno per incastrare una porta e così via. Se abbiamo a disposizione una decina di mattoncini possiamo costruire molte cose diverse e a seconda di come li assembliamo otteniamo delle forme diverse. Questo è quello che accade con le proteine: sono composte da parti, chiamati **domini**, che solitamente hanno una struttura tridimensionale definita indipendentemente dalla proteina in cui sono inseriti, rintracciabili in diverse proteine. Abbiamo detto che conoscere la struttura tridimensionale di una proteina è un passo molto importante per un ricercatore, perché è la forma che stabilisce la funzione: cioè se e come interagisce con gli altri elementi all'interno della cellula. Pertanto uno dei filoni di ricerca di sempre riguarda la ricerca di tecniche che permettano di avere una rappresentazione tridimensionale della forma reale della proteina. A questo fine, le tecniche di elezione sono la **cristallografia a raggi X** e la **risonanza magnetica nucleare**. Purtroppo sono tecniche che richiedono molto tempo, molto costose e non tutte le proteine si prestano. Infatti, per poterle analizzare è necessario purificarne alte dosi, quindi immobilizzarle per lo "scatto fotografico" in strutture cristalline. E non tutte le proteine si lasciano

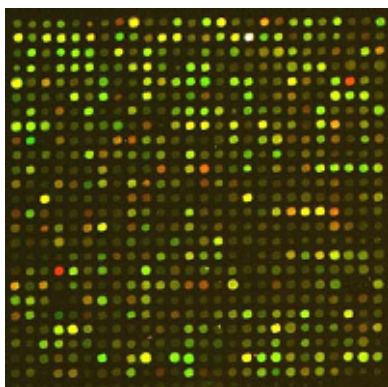
cristallizzare. Per velocizzare l'acquisizione dei dati sulla funzione delle proteine, in attesa del riscontro sperimentale, i bioinformatici hanno messo a punto dei software in grado di far tesoro delle informazioni già acquisiti su proteine di cui sono disponibili le strutture tridimensionali acquisite sperimentalmente per dedurre quelle ignote di proteine non ancora analizzate. Questi sistemi sfruttano proprio il principio dei mattoncini introdotti in questo paragrafo: analizzano la struttura tridimensionale delle proteine già studiate, individuando la forma dei mattoncini che la compongono. Quindi, analizzano la nuova proteina per individuare gli stessi mattoncini. Infine, producono alcuni modelli 3D in cui è illustrata l'ipotetica struttura della proteina dedotta dall'assemblamento dei domini diversi che la compongono. Ovviamente, la struttura è tanto più fedele quanto è stato possibile individuare in essa dei domini già noti e descritti. Sulla base della struttura, i ricercatori fanno delle ipotesi sulla funzione che verificano con **esperimenti in laboratorio**. Per esempio, se la struttura denota la presenza di una tasca che potrebbe accogliere al suo interno una seconda proteina con cui è noto interagire, i ricercatori provano a modificarla per vedere se l'interazione si interrompe. Se ciò accade, probabilmente la deduzione fatta è corretta.

6 Assegnare nuove funzioni ai geni

Modellizzare la struttura tridimensionale di una proteina è un sistema che ci permette di trarre informazioni sulla funzione del gene che la produce. Ma esistono anche altri sistemi, che ci permettono di fare ipotesi lavorando solo ed esclusivamente sulla sequenza genica. La prima cosa da fare è leggere con attenzione le informazioni già presenti nella **banca dati** in cui l'abbiamo trovato: probabilmente contiene già informazioni importanti come la presenza di domini particolari (analogamente alle proteine, i domini dei geni sono strutture ricorrenti a cui è associata una funzione specifica), la composizione, la presenza di punti che indichino il fatto che possa dar luogo a più proteine diverse

(questo è possibile perché la struttura modulare può permettere durante la produzione delle protiene di inserire o meno un dominio). In secondo luogo, si possono cercare le sequenze omologhe in banca dati. Si inserisce la sequenza in un motore di ricerca specifico (proprio come si fa su internet!) e dopo un pò otteniamo un elenco di altri geni, con indicata la somiglianza. A quel punto possiamo cercare di capire se per quelli più simili è già nota la funzione. Se la funzione è nota e i geni sono considerabili omologhi, probabilmente siamo a buon punto. Se non troviamo informazioni soddisfacenti, oppure solo una parte del nostro gene ha omologhi ben descritti in altri geni, possiamo fare un'analisi di alcune sue parti. Isoliamo le parti incognite e ricominciamo la ricerca in banca dati: può darsi che escano fuori geni molto diversi dal nostro, ma simili solo per la parte in considerazione. Questo vuol dire che non abbiamo di fronte una semplice "parte", ma un dominio, di cui magari è nota la funzione. Il nostro dominio magari non è proprio identico, perché altrimenti avremmo trovato già in prima battuta la descrizione della funzione, ma essendo simile potrebbe trattarsi di una versione evolutivamente precedente o posteriore a quella già descritta. Magari, una parte della funzione è conservata: non resta che verificare in laboratorio. Tutte queste operazioni rientrano nel termine tecnico generale di "**annotazione genomica**", cioè l'assegnazione di funzioni ai geni e loro parti.

7 Bioinformatica e microarray a DNA



I **microarray** sono piccole superfici, generalmente poste su di un vetrino da microscopia, su cui sono legate chimicamente piccole molecole di DNA. Sono utilizzati per saggiare in un campione la presenza e la quantità di determinate sequenze di DNA o *RNA*, grazie alla capacità di legarle in modo specifico e di emettere una luce proporzionale alla quantità di molecole legate. Le diverse sequenze che si possono indagare sono disposte nel vetrino secondo una **griglia**: analizzando la luminosità in ciascun quadratino si può dedurre se la sequenza corrispondente è presente, e in quale quantità, nel nostro campione. Questo tipo di analisi normalmente si effettua su diverse migliaia di sequenze in contemporanea. L'immagine che si ottiene fotografando un vetrino di microarray al termine dell'esperimento assomiglia molto a quello dell'immagine in alto a sinistra. Il computer, dopo aver registrato le luminosità, ci restituisce una tabella in cui associa ad ogni quadratino il riferimento alla sequenza che contiene e la luminosità rilevata. A questo punto, la bioinformatica ci aiuta ad interpretare i dati. Le operazioni che si possono effettuare con l'ausilio del computer e di software appositi sono moltissime, ne segnaliamo solo alcune per avere un'idea di quello che è possibile fare. Il primo punto è capire quali differenze di luminosità sono da ritenersi significative: è possibile che sul vetrino sia rappresentata tutta la gamma di luminosità possibile, tocca a noi stabilire i limiti di significatività. Normalmente si procede in parallelo anche con un microarray di controllo (oltre a dedicare alcune "cellette" dello stesso microarray a controlli interni) che sono usati come parametro di confronto: il primo passo, quindi, è "normalizzare" i dati con quelli di controllo. Quindi si applicano diversi strumenti statistici per validare i dati. A questo punto possiamo pensare veramente al nostro esperimento. Consideriamo di voler analizzare l'espressione genica (cioè quali geni sono attivi in una determinata cellula) di una cellula cancerosa, per capire quali sono le differenze a livello di funzionamento del DNA. Realizziamo l'esperimento usando due microarray diversi, uno per le cellule normali e uno per quelle tumorali. A questo punto, confrontiamo i risultati e chiediamo al computer di isolarci le

differenze: otterremo una lunga lista di sequenze che si comportano in modo diverso. Ripetiamo alcune volte l'esperimento per ripulire la nostra lista da differenze non significative: alcune di queste potrebbero essere oscillazioni biologiche che non dipendono strettamente dalla genetica. Ottenuta la nostra lista ripulita, ci accorgiamo che è ancora molto lunga: centinaia di nomi incomprensibili si susseguono. A questo punto quello che fa un bioinformatico è applicare un algoritmo statistico che ci fornisce una catalogazione ragionata della nostra lista sulla base di informazioni presenti nelle banche dati. Questa si chiama **classificazione ontologica**: le sequenze sono classificate in base alla loro funzione in insiemi e sottoinsiemi. A questo punto sarà molto più semplice leggere i nostri dati: sapremo, per esempio, che nella nostra cellula tumorale sono iperattivi i geni che appartengono ad una certa classe di molecole che controllano la proliferazione. Se siamo fortunati, una di queste è una molecola chiave (cioè, se inibita blocca l'intero sistema in cui agisce) che può essere interessante usare come base per progettare un nuovo farmaco antitumorale. In realtà la bioinformatica è utile in molti altri passaggi dell'analisi di microarray: per esempio, è fondamentale per progettare in modo adeguato le sequenze che devono essere messe nella griglia (tecnicamente definite "sonde").