Bioinformatica: nuovo termine o nuova scienza?

Alessandro Giuliani e Romualdo Benigni

Laboratorio di Tossicologia Comparata ed Ecotossicologia, ISS

Riassunto - L'apparizione sulla scena della nuova disciplina della bioinformatica ha rappresentato un importante punto di svolta nelle scienze biomediche. Dopo circa venti anni di assenza dal proscenio, il ragionamento quantitativo è tornato prepotentemente in auge nelle scienze biologiche, facendo rinascere l'interesse per tecniche ormai relegate a un uso altamente specialistico in campi secondari delle discipline biologiche. La necessità di gestire grandi masse di dati, insieme alla diffusione di massa di tecniche statistiche in ambienti a lungo digiuni di qualsiasi conoscenza matematica, ha provocato, e ancor più provocherà in futuro, un salutare scossone al mondo raffinato ma abbastanza sterile della statistica applicata e insieme un ripensamento del significato del lavoro del biologo.

***Parole chiave*:** bioinformatica, modellizzazione, teoria

Summary (*Bioinformatics: a new term or a new science?*) - An important turning point in the biomedical sciences has been the appearance of a new field: bioinformatics. For the last twenty years, quantitative methods and style of reasoning have been largely out of fashion in life sciences, but they have now come back into the limelight. Further, the need to manage huge amounts of data generated by post-genomic sciences, combined with the diffusion of refined statistical methods in fields where mathematics and statistics have long been altogether absent, will likely generate a positive change in both applied statistics and biology.

***Key words*:** bioinformatics, modelling, theory

In questi ultimi cinque-sei anni una nuova parola ha fatto una massiccia comparsa nelle riviste scientifiche e di divulgazione, nelle intestazioni di istituti scientifici e dipartimenti, come pure sui mezzi di comunicazione di massa: "bioinformatica".

Nella sua accezione etimologica, questo termine sta semplicemente a indicare l'uso del calcolatore nella ricerca biologica. Storicamente nasce 10-15 anni fa, quando si è posta l'esigenza di conservare e organizzare nei calcolatori dati sempre più voluminosi prodotti dai lavori di sequenziamento del genoma (umano e non). Col tempo ci si è accorti che la trasformazione di puri elenchi (sequenze geniche) in progressi in campo biomedico non era così immediata e la bioinformatica si è cimentata via via con compiti nuovi e più "intelligenti" della semplice catalogazione, ad esempio la localizzazione di geni lungo il genoma e, successivamente (post genomica), l'analisi dei profili di espressione genica (misure di distribuzione di mRNA, usando la cosiddetta tecnica dei DNA *microarray*). A breve distanza di tempo, è sorta una nuova area di ricerca caratterizzata dall'uso massiccio della bioinformatica: la proteomica, cioè l'analisi su larga scala delle proteine. Il suo sviluppo è stato ispirato dalla comprensione che il prodotto finale di un gene (cioè la proteina) può essere più complesso, per via di tutte le modificazioni post trascrizionali, e comunque, più vicino alla funzione di quanto non lo sia il gene stesso o la sua espressione (mRNA). Come si nota da questa breve carrellata, man mano che gli obiettivi e le tendenze nella ricerca cambiavano, anche la bioinformatica cambiava vesti e compiti e in qualche modo, nella presentazione dei media è stata talora quasi assimilata e considerata un tutt'uno con i vari campi di ricerca che terminano in "-omica"

(genomica, proteomica, funzionomica, ecc.). Inoltre, poiché gli strumenti di laboratorio attuali permettono la misurazione simultanea di molti dati (*high-throughput*), scopo della bioinformatica è quello di studiare i sistemi biologici a livello "sistemico" per comprendere nella sua globalità: a) la struttura del sistema, come rete gene/metabolismo/trasduzione dei segnali, e sua struttura fisica; b) dinamica di tali sistemi; c) metodi per controllare i sistemi; d) metodi per progettare e modificare sistemi con proprietà desiderate.

Ovviamente, lo scopo di questa breve presentazione non è quello di fare un bilancio degli enormi campi di indagine accennati sopra, ma di limitare il discorso alla bioinformatica propriamente detta, separandola dai campi a cui si applica ed esaminandola da un punto di vista metodologico. Ugualmente non si parlerà dei progressi nel campo dell'informatica, che vengono sfruttati dalla bioinformatica, perché seguono delle leggi di sviluppo completamente autonome.

Tornando al termine "bioinformatica" bisogna riconoscere che è poco significativo. Infatti nessuno definisce il lavoro di un commercialista (che si avvale estensivamente del computer) "econoinformatica" o "informatica amministrativa". Così come un geologo che disegna delle mappe con un sistema automatico non si definisce "geoinformatico", né uno scrittore che invece di usare la penna usa un programma di videoscrittura non può essere definito un "letterato informatico". Giustamente il nucleo fondante di queste attività (la gestione aziendale, la geologia e la letteratura) viene individuato altrove (comportamenti economici, caratteristiche della crosta terrestre, valore artistico) e non nell'uso di un mezzo informatico. Forse la necessità di coniare un nuovo termine deriva da un'imbarazzante (e per certi versi inspiegabile) assenza protrattasi per circa venti anni nella porzione maggioritaria della ricerca biologica: l'assenza del ragionamento quantitativo. Questa assenza, che ha addirittura portato taluni a mettere in dubbio la possibilità di definire la biologia molecolare come scienza propriamente detta (1), parte dagli anni '70, parallelamente alla diffusione di efficienti tecnologie sempre più automatizzate. Con gli ovvii limiti che hanno sempre le semplificazioni, un tipo di procedura che si può riconoscere in numerosi lavori di biologia molecolare è la seguente: a) separazione del materiale biologico (solitamente cellule) proveniente da due classi differenti (ad esempio, malati di una specifica patologia e soggetti sani); b) dimostrazione attraverso metodi cromatografici (ad esempio, *Southern blot*) che un gene ha un'espressione differente nei due gruppi; c) costruzione di una spiegazione plausibile del motivo per cui quel gene (o meglio la sua assenza o variazione) provochi la patologia (o la caratteristica fenotipica) in questione.

Per portare a termine questa sequenza di operazioni non c'era alcun bisogno di metodi quantitativi: la variabilità entro le popolazioni era annichilita dal fatto che il materiale biologico di partenza veniva fuso in due gruppi non ulteriormente divisibili (soggetti sani, malati) perdendo ogni riferimento individuale e quindi facendo venir meno ogni possibilità di analisi statistica. L'appiattimento sovente arbitrario del meccanismo patogenetico alla presenza o assenza di un certo gene (e quindi l'implicita identificazione del genotipo col fenotipo) rendeva inutile ogni genere di modellizzazione quantitativa che rendesse ragione della complessità del fenotipo osservato; così pure, osservare delle macchie su una lastra cromatografica è un'attività assolutamente intuitiva che non necessita di ulteriore elaborazione.

> **La bioinformatica riconcilia la biologia con il ragionamento quantitativo**

Questo tipo di approccio ha avuto luci e ombre. A notevoli avanzamenti nella conoscenza di alcuni elementi chiave dei meccanismi patogenetici (2), si sono affiancati episodi francamente disdicevoli di "cattiva scienza" in cui, ad esempio, si giungeva a identificare mirabolanti geni dell'intelligenza o della follia su basi statistiche praticamente nulle e con una metodologia scientifica decisamente puerile (3).

Durante gli anni '80 si intravede la necessità di organizzare la raccolta (fino ad allora sostanzialmente episodica e scoordinata) di informazioni genetiche all'interno di un piano sistematico di sequenziamento del DNA umano (e di altre specie interessanti per gli studi biologici). Le idee di fondo sono sempre le stesse ma trasportate su una scala di massa: il raggiungimento dell'obiettivo finale del sequenziamento dell'intero patrimonio genetico avrebbe dovuto corrispondere al completo dispiegamento di "tutto ciò che c'è da sapere". Tuttavia, il riconoscimento in corso d'opera che i sistemi biologici sono sistemi complessi, e che il rapporto genotipo/fenotipo è altamente intricato e fortemente non lineare, porta la genomica a diventare post genomica e a richiedere che la scienza qualitativa e riduzionista si trasformi in scienza quantitativa e fortemente olistica. L'emergere prepotente della bioinformatica si spiega quindi con la riscoperta della necessità della statistica. Questa era una strategia che i biologi avevano utilizzato fino agli anni '70 (la fondazione stessa della genetica moderna, da parte di Mendel, è basata sull'uso della statistica) e gruppi di scienziati lontani dai riflettori (ecologi, bio-

logi di popolazione, farmacologi) avevano continuato a usare metodicamente. Come esemplificativo di tali applicazioni quantitative alla biologia, minoritarie ma rigogliose, può essere illuminante la seguente definizione di scienza: "Per scienza si intendono descrizioni quantitative, con l'uso di un numero relativamente limitato di parametri ben conosciuti e di grafici per creare le connessioni". La definizione è di Corwin Hansch, che negli anni '60 è stato il protagonista principale della nascita della scienza delle relazioni quantitative struttura-attività, dove attraverso il formalismo fornito dalla matematica si combinano insieme due scienze apparentemente lontanissime come la chimica fisica e la biologia (4).

Bisogna chiarire che, in termini di strumenti tecnici usati, la bioinformatica può usufruire di un corredo ormai vastissimo e consolidato di approcci all'analisi dei dati, creati nel corso dell'ultimo secolo per risolvere problematiche legate a varie aree di ricerca. Queste metodologie già esistenti sono in grado di venire incontro alla stragrande maggioranza delle esigenze della bioinformatica. Per citare solo degli esempi, ricorderemo due studi compiuti di recente nel Reparto Struttura-Attività dell'Istituto Superiore di Sanità. Il primo esempio riguarda un tema squisitamente bioinformatico, cioè l'espressione genica differenziale di colture di cellule tumorali analizzata su *chip* contenenti migliaia di geni (la cosiddetta tecnologia dei DNA *microarray*) (5). I raggruppamenti in tipologie di cellule, e la definizione delle tipologie è stata compiuta con l'uso dell'analisi delle componenti principali (Figura 1). Questa tecnica fa parte delle analisi multivariate dei dati, che considerano vaste basi di dati, consistenti di molti oggetti, su ognuno dei quali vengono misurate molte caratteristiche o variabili. L'analisi delle componenti principali è stata sviluppata circa cento anni fa e usata in molti campi delle scienze biologiche (6). Il secondo esempio riguarda la modellizzazione di proteine (Figura 2). In questo campo, è stata usata la tecnica di analisi delle ricorrenze, che è un metodo nato per l'analisi delle serie temporali; tale metodo è stato usato con successo per correlare i profili di idrofobicità delle sequenze proteiche con proprietà come stabilità termica, capacità di aggregazione, effetto delle mutazioni sulla funzionalità degli enzimi (7). È importante sottolineare che le metodologie statistiche usate fanno parte della tradizione quantitativa classica delle scienze biologiche; esse sono state usate per studiare le relazioni quantitative tra struttura chimica e attività biologica delle molecole come pure gli svolgimenti temporali di segnali fisiologici e quindi erano "pronte" per essere applicate alle nuove tematiche post genomiche.

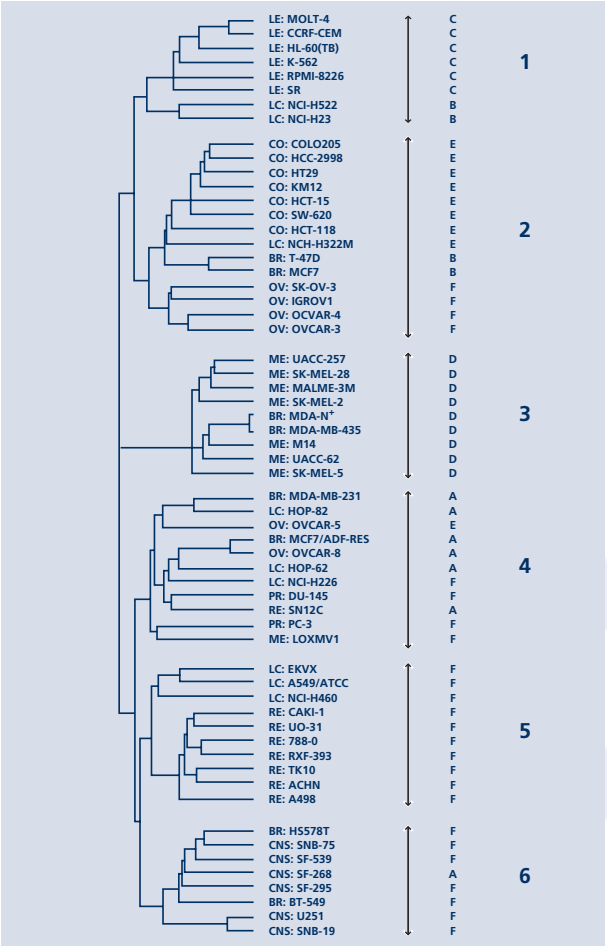

Figura 1 - Classificazione gerarchica di 60 linee tumorali sulla base dell'espressione differenziale di 1 315 geni. L'albero sulla sinistra si riferisce alla classificazione ottenuta utilizzando 1 315 variabili (i singoli geni). Sulla destra sono riportate le classi ottenute utilizzando solo 5 variabili, corrispondenti alle componenti principali dell'espressione dei geni. La forte congruenza delle due classificazioni dimostra come l'espressione genica sia fortemente strutturata in reti metaboliche altamente organizzate

> **Per scienza si intendono descrizioni quantitative, con pochi parametri e grafici per creare le connessioni**

Cosa distingue allora la bioinformatica dalla statistica applicata alla biologia? Una cosa molto importante anche se spesso poco considerata: la giovinezza. Se leggiamo una rivista di statistica applicata alla biologia (*Biometrics*, *Biometrika,* ecc.), troveremo articoli di matematica dove l'aspetto biologico è poco più che un pretesto per esercitazioni teoriche e metodologiche. Quando la statistica era giovane, pionieri come Galton, Pearson, Fisher lavoravano a stretto contatto con i biologi e i loro sforzi metodologici avevano un riconoscibile impatto sul pensiero biologico. La scelta di una tecnica era guidata e ispirata dalla soluzione di un

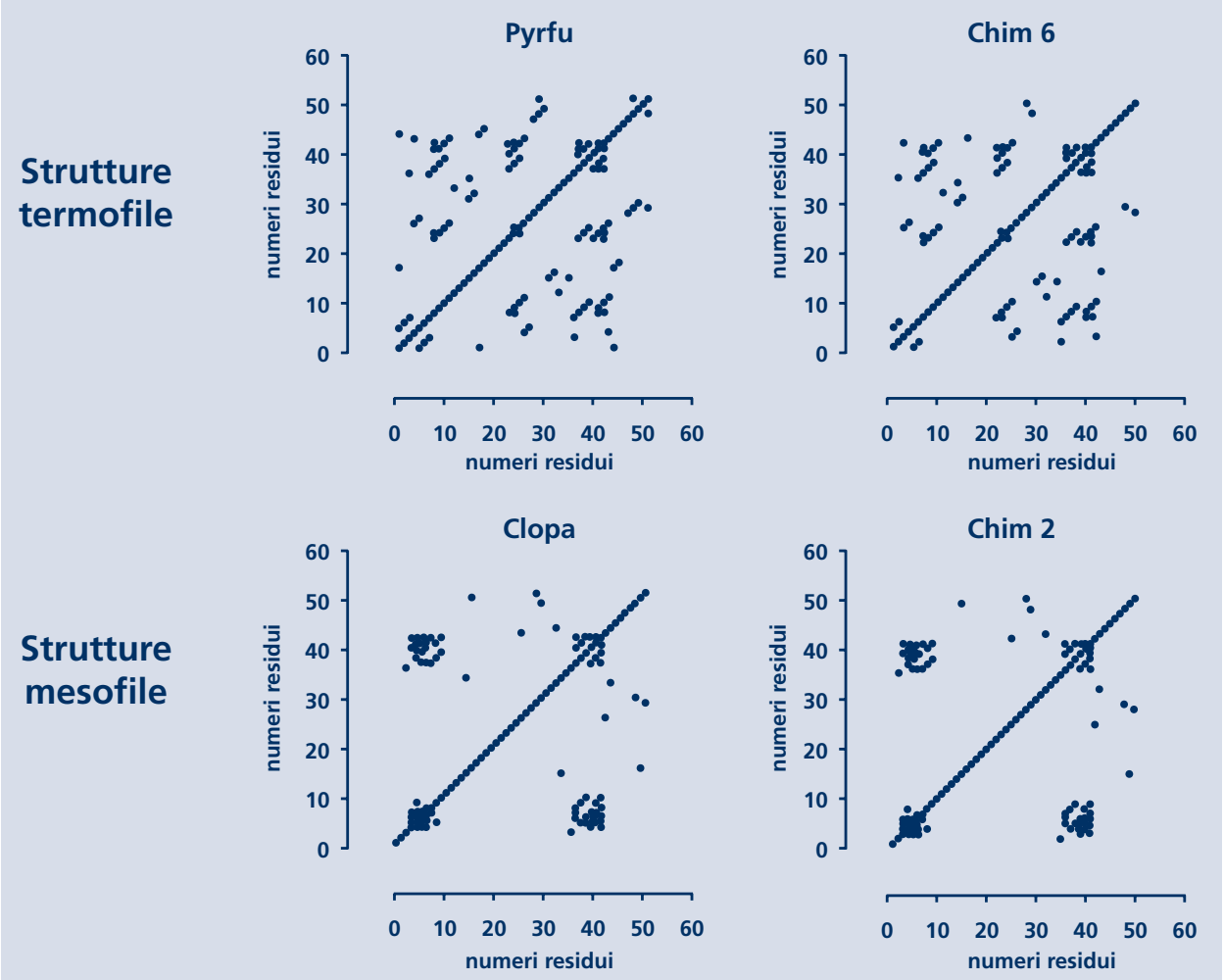

Figura 2 - Grafici di ricorrenza (RQA) delle sequenze di 4 proteine (Rubredoxine); i grafici sono rappresentazioni dei loro profili di idrofobicità. Nella parte superiore sono riportate due strutture termofile (stabili ad alte temperature), mentre nella parte inferiore due strutture mesofile (da organismi che vivono a temperatura ambiente). I grafici di ricorrenza mettono in luce con chiarezza la differenza tra i due tipi di proteine, altrimenti non rilevabile né in termini di sovrapposizione di sequenze né in termini di struttura tridimensionale

problema urgente e reale. Le metodologie erano quindi strettamente calate nella realtà e, in qualche modo, giustificate da essa. Ad esempio, un grosso impulso allo sviluppo dell'analisi delle componenti principali è derivato dal desiderio di dare valutazioni oggettive e quantitative in un campo in apparenza estremamente aleatorio come quello dei test psicologici. La statistica metodologica di punta e la biologia nel tempo si sono allontanate, mantenendo un legame di pura convenienza. Il partire da zero da parte di persone che non conoscono la statistica costringe la metodologia a eliminare tutti gli orpelli e a mantenere solo gli aspetti essenziali; d'altro canto i biologi sono costretti ad appropriarsi di una cultura quantitativa, pena l'incapacità di sfruttare appieno e in prima persona i risultati del loro lavoro.

La bioinformatica offre quindi la speranza di una fase nuova, di rapporti innovativi e reciprocamente fruttuosi tra biologia e statistica.

Riferimenti bibliografici

1. Maddox J. Is molecular biology yet a science? Nature 1992; 355: 201.
2. Gartel AJ, Tyner AL. The role of cyclin-dependent kinase inhibitor P21 in apoptosis. Mol Cancer Therap 2002; 8: 639-49.
3. Brunner HG, Nelen M, Breakfield O, et al. Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase. Science 1993; 262: 562-78.
4. Hansch C, Kurup A, Garg R, et al. Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. Chem Revs 2001; 101: 619-72.
5. Crescenzi M, Giuliani A. The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. FEBS Letters 2001; 507: 114-8.
6. Benigni R, Giuliani A. Quantitative modeling and biology: the multivariate approach. Am J Physiol 1994; 266: R1697-R1704.
7. Giuliani A, Benigni R, Zbilut JP, et al. Nonlinear signal analysis methods in the elucidation of protein sequence structure relationships. Chem Revs 2002; 102: 1471-91.