DOI: 10.4081/jsas.2013.102

SHELTERING FROM AVALANCHES OF BIOLOGICAL DATA: A NEW RESEARCH DIMENSION IN THE POST-GENOMICS ERA

Neri Niccolai

Dipartimento di Biotecnologie, Chimica e Farmacia, Università degli Studi di Siena, Siena, Italy

Abstract. Biomedical research in the Post-genomic Era is characterized by huge amounts of data which cannot be manually analyzed soon after their collection, but only stored in data banks for further investigations. Thus, many data banks have been created to keep some order for results obtained from high throughput techniques applied to genomics or to other "omics" studies. Some order is also needed to browse fruitfully biological data among all the available databanks. National Center for Biotechnology Information of USA offers to researchers a powerful interface to achieve this effort, but in a way that is not suitable for the public perception. A Google Earth kind of software would help to sail safely within this data ocean to specific pieces of information, as required both by researchers and by just curious people. Bioinformatics, a new interdisciplinary field that develops and improves on methods for storing, retrieving, organizing and analyzing biological data, is emerging as a powerful tool to derive novel perspective of biological processes. In our laboratory, for instance, scanning the Protein Data Bank to analyze amino acids distribution inside protein structures, new scenarios appeared which can account for long distance protein-protein interactions.

Key words: post-genomics research, Bioinformatics, biological databanks integration

Alla fine dello scorso secolo, l'impegno dei laboratori di ricerca di tutto il mondo per la decodifica del codice genetico umano produsse un impetuoso sviluppo di nuove tecniche che permisero il raggiungimento di tale obiettivo. I risultati ottenuti in quella circostanza, per certi aspetti inattesi, hanno stimolato tantissimi altri studi per la caratterizzazione del DNA di molte altre specie. Le procedure sperimentali e di analisi computazionale utilizzate a tale scopo, sono state soggette ad una continua evoluzione, in costante accelerazione, che le ha rese sempre più efficienti ed economiche.

Ad esempio, i risultati del Progetto Genoma Umano, completato nell'aprile del 2003 con la determinazione della sequenza completa del nostro DNA, sono stati ottenuti con l'investimento di 2,5 miliardi di € e dopo ricerche durate tredici anni [1].

Risultati simili sono ottenibili oggi, in pochi giorni e ad un costo che rapidamente si sta avvicinando ai 1000 € per genoma [2], grazie agli sviluppi tecnologici raggiunti ed alla estesa automazione delle procedure d'indagine.

Per questo motivo, a buon diritto, possiamo dire che siamo entrati in una nuova Era della conoscenza delle Scienze della Natura, quella post-genomica, caratterizzata dalla valanga di dati che si possono rapidamente ottenere delineando la composizione del DNA non solo di specifiche specie, ma anche di singoli individui. Le applicazioni di questa messe di dati risultano evidenti, dalla diagnosi e cura di malattie genetiche, alla ottimizzazione di strategie terapeutiche per ciascun paziente, la cosiddetta Farmaco-genomica, fino ad arrivare alla lotta razionale di specie patogene.

Allo stesso tempo, però, l'abbondanza dei dati genomici e di quelli ad essi collegati, ha determinato la necessità di una loro sistematizzazione in banche dati di pubblico accesso e di facile consultazione.

Il Centro Nazionale per l'Informazione Biotecnologica degli USA, NCBI, ha reso disponibile l'accesso in Internet a numerose delle principali banche dati che contengono informazioni sulle sequenze e sulle strutture molecolari di polimeri biologici, quali DNA, RNA e proteine [3]. Nella Figura 1 sono riportati i contenuti di alcune tra queste banche dati, aggiornati al novembre 2013, accessibili attraverso la pagina di NCBI.

Per alcune informazioni, come ad esempio il numero di proteine di cui è stata proposta la sequenza da studi genomici (oltre 100 milioni) o il numero dei riferimenti bibliografici accessibili (oltre 23 milioni), i contenuti delle banche dati sembrano talmente estesi da renderne complessa la gestione, particolarmente se ciò deve essere condotto in modo integrato.

La possibilità di rendere facilmente accessibili su

Correspondence to:
Neri Niccolai,
Department of Biotechnology, Chemistry and Pharmacy,
The University of Siena,
via Moro 2, 53100 Siena, Italy

supporti informatici tutte le informazioni disponibili relativamente agli aspetti morfologici alla base della vita, è un compito che la Biologia strutturale si sta ponendo, se pur non senza difficoltà.

Un "navigatore" che funzionasse nel modo con cui il software *Google Earth* riesce a dare una varietà d'informazioni complesse, distribuendole su base geografica, suggerisce che una procedura analoga potrebbe essere seguita, anche se gli oggetti "geograficamente" da definire hanno dimensioni che possono essere inferiori al miliardesimo di metro. Tale procedura informatica potrebbe essere sviluppata per esplorare quanto è noto a proposito del contenuto molecolare di ogni singola cellula. Infatti, una procedura di tipo *Geographic Informatic System,* o GIS [4], collegata ad aspetti morfologici di dimensioni assai diverse, da una particolare specie alle rispettive cellule, fino ad arrivare alle molecole in esse contenute, potrebbe facilitare la "navigazione" dei ricercatori nel mare di dati a disposizione. Inoltre, la stessa "navigazione" permetterebbe una rapida visualizzazione dello stato dell'arte delle conoscenze biomolecolari ad uso di una ampia utenza, comprendente anche i docenti e gli studenti dei centri di formazione universitaria e non solo.

Le risorse sperimentali ed informatiche per iniziare questo progetto di navigazione tra banche dati biomolecolari sono già ampiamente di routine: in un prossimo futuro, si tratterà solo di perfezionarne le modalità di accesso e di integrazione, oltre che di aggiornarne progressivamente i contenuti.

Le informazioni che sono distribuite in modalità GIS da piattaforme quali *Google Earth* o simili, sono numerose e non solo di tipo morfologico come quelle che ci permettono di caratterizzare, con continuità, dalla forma dei continenti a quella di limitate porzioni di territorio e fino ai singoli edifici posizionati nelle varie parti del globo.

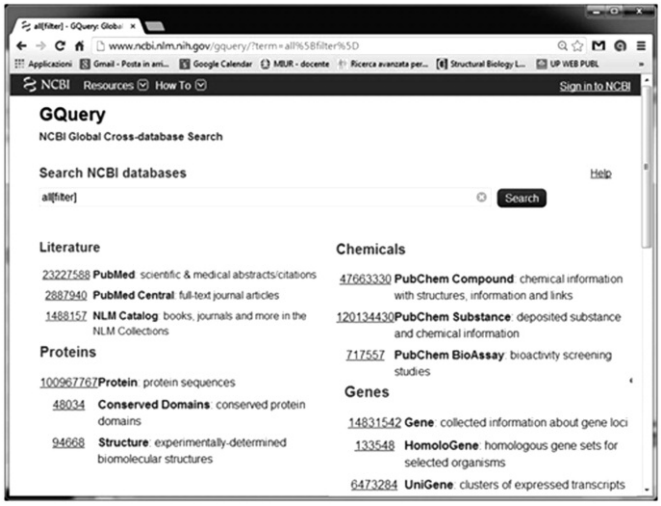

Figura 1: La pagina d'accesso alle banche dati raggiungibili attraverso lo NCBI con i relativi contenuti aggiornati al novembre 2013.

Infatti, sono accessibili per mezzo delle stesse piattaforme informatiche anche informazioni dinamiche quali, ad esempio, le condizioni del traffico stradale o la situazione meteorologica. Inoltre, mediante queste piattaforme e sulla base di modelli statistici opportuni, è possibile anche avere predizioni sull'evoluzione delle varie situazioni in archi temporali più o meno lunghi che determinano l'attendibilità della informazione offerta.

Introdurre una analoga dimensione dinamica nell'esplorazione virtuale della composizione dei vari tipi di cellule per acquisire informazioni sulla Natura nei suoi aspetti molecolari, costituisce un problema di grande complessità per il quale solo da poco tempo stanno maturando possibili risposte.

Infatti, l'abbandono di una visione riduzionistica delle conoscenze sui componenti dei vari processi alla base della Vita è assai recente e lo sviluppo di un nuovo approccio per lo studio dei fenomeni naturali, la cosiddetta "Biologia dei Sistemi", ne costituisce uno dei principali risultati [5]. Conseguentemente, le interazioni tra molecole sono state messe al centro delle indagini per individuare i meccanismi dei processi biologici alla risoluzione dei singoli atomi. Il fatto che sia stato messo in luce che le proteine, i principali strumenti molecolari che sovraintendono a tali processi, lavorino in modo concertato, ha aperto un ampio campo di ricerche. Tra queste, lo studio di come interferire con specifiche interazioni tra coppie di proteine all'interno di questa fitta rete di interazioni proteina-proteina, rappresenta una emergente prospettiva per lo sviluppo di nuovi farmaci antitumorali [6].

E dunque, la presenza di questa enorme massa di dati, se da una parte suggerisce la necessità di modalità integrate di consultazione, dall'altra sta dando energia ad una nuova area della ricerca scientifica, quella della Bioinformatica, ovvero dello sviluppo di procedure per estrarre informazioni biologiche dal contenuto di una o più banche dati [7].

E' da sottolineare che la Toscana ospita molti gruppi attivi in questo settore e l'invito della Società Italiana di Bioinformatica ad organizzare incontri regionali, qua è stato accolto con entusiasmo e le sue prime due edizioni, BIOINFORMATIHA 1 e BIOINFORMATIHA 2, si sono già svolte con notevole successo di partecipazione rispettivamente a Siena nel 2012 ed a Firenze nel 2013 [8].

Uno di questi gruppi è quello che opera nel laboratorio di Biologia strutturale del Dipartimento di Biotecnologie, Chimica e Farmacia dell'Università di Siena che io coordino. Nel corso dei nostri studi, abbiamo messo a punto uno strumento di calcolo che fornisce una misura quantitativa della profondità degli atomi all'interno di strutture molecolari complesse [9]. Una scansione bioinformatica con tale strumento di calcolo su tutto il contenuto della Protein Data Bank [10], la banca delle strutture molecolari che sono state determinate con tecniche sperimentali, ha messo in evidenza che la superficie delle proteine è ricca di cariche elet-

triche di segno opposto che molto spesso si trovano a breve distanza. Questa recente osservazione [11] ha suggerito una possibile risposta, che la Biologia dei sistemi al momento non ha, per spiegare i meccanismi di comunicazione a distanza che regolano la rete di interazioni tra molecole all'interno della cellula ed, in generale, nei fluidi biologici. A questo proposito, si deve sottolineare il fatto che le interazioni tra le diverse molecole che sono impegnate a sostenere i processi vitali, avviene in una situazione di grande "affollamento molecolare" di difficile valutazione sia dal punto di vista sperimentale che teorico [12].

In ogni caso, più si ampliano le conoscenze molecolari sulla complessità della Natura, tanto più possono essere escluse le precedenti ipotesi che attribuivano ad incontri casuali un qualsivoglia ruolo nella formazione degli aggregati tra molecole che sono necessari per l'istaurarsi dei vari processi biologici. La presenza, dunque, sulle superfici delle proteine di coppie di cariche elettriche di segno opposto in continuo movimento a causa dei moti che avvengono all'interno delle molecole proteiche (vedi l'animazione realizzata durante un nostro studio e depositata su *YouTube* [13]), potrebbe generare segnali elettromagnetici di efficacia sufficiente a rimuovere la casualità nel movimento reciproco delle molecole all'interno del loro naturale ambiente biologico. Questo meccanismo di comunicazione intermolecolare a distanza lo stiamo valutando con grande attenzione, anche negli aspetti quantitativi che non risultano di facile analisi. In ogni caso, il fatto che le cariche elettriche in movimento sulla superficie delle proteine possano generare onde elettromagnetiche che ne determinano una loro impronta caratteristica, eliminando così ogni tipo di casualità per gli incontri tra molecole diverse, sembra sempre più un "uovo di Coulomb", del tipo iconizzato nella Figura 2. Questi effetti elettrodinamici saranno, finalmente, presi in considerazione per avere una descrizione accurata dei modi con cui le biomolecole riescono a comunicare a distanza, nel mezzo del grande traffico molecolare che è alla base della Vita.

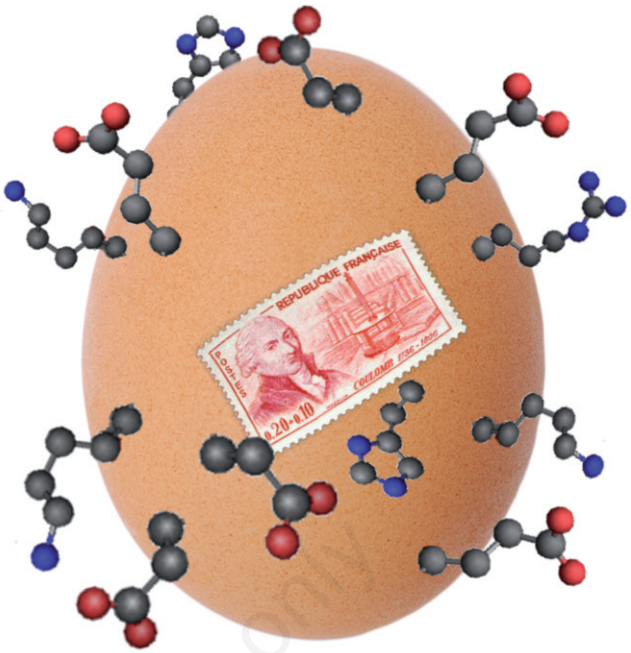

Figura 2: La vicinanza di catene laterali di amminoacidi in rapido movimento reciproco con carica elettrica opposta, potrebbe generare segnali elettromagnetici caratteristici per ciascuna proteina e necessari per un riconoscimento molecolare a lunga distanza.

BIBLIOGRAFIA

1. *http://web.ornl.gov/sci/techresources/Human_Genome/project/timeline.shtml*
2. *http://www.genome.gov/sequencingcosts*
3. *http://www.ncbi.nlm.nih.gov*
4. *http://en.wikipedia.org/wiki/Geographic_information_system*
5. *http://it.wikipedia.org/wiki/Biologia_dei_sistemi*
6. *Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 2007;450:1001-9.*
7. *http://it.wikipedia.org/wiki/Bioinformatica*
8. *http://www.bioinformatiha.it*
9. *Varrazzo D, Bernini A, Spiga O, et al. Three-dimensional computation of atom depth in complex molecular structures. Bioinformatics 2005;21:2856-60.*
10. *http://www.pdb.org*
11. *Alocci D, Bernini A, Niccolai N. Atom depth analysis delineates mechanisms of protein intermolecular interactions. Biochem Biophys Res Commun. 436:725-729; 2013.*
12. *McGuffee SR, Elcock AH. Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput Biol 2010;6:e1000694; .*
13. *http://www.youtube.com/watch?v=AeYAHOCaRbk*