Informatica Cultura Società
Roma, 24 Gennaio 2006

BIOINFORMATICA: LA SFIDA DELLA COMPLESSITA'

Gian Antonio Danieli
Università di Padova – Centro Ricerca Interdipartimentale Biotecnologie

Il fondamentale ruolo dell'informatica nella ricerca biologica contemporanea e' testimoniato dalle diverse fasi del progetto "Genoma Umano". L'applicazione dell'informatica alla gestione delle apparecchiature per l' analisi del DNA ha consentito di realizzare nei laboratori condizioni "high-throughput" analoghe a quelle delle catene di produzione industriale automatizzata. Metodi informatici hanno consentito di ricostruire un'unica sequenza lineare di DNA per ciascun cromosoma umano, mediante collazione delle numerosissime sequenze ottenute dalle singole analisi effettuate in diversi laboratori. Sono state create banche dati di sequenze ed è stata realizzata la loro interconnessione via internet e la loro accessibilità *on-line*. Infine, sono state riunite le informazioni depositate in diverse banche dati, riferite ad aspetti diversi ma tutti attinenti alla sequenza del DNA del genoma umano, come la posizione relativa di ciascun segmento rispetto ad altri dello stesso cromosoma, la caratteristica funzionale del segmento in questione, le caratteristiche del prodotto codificato da ciascun "gene", le caratteristiche di espressione di ciascun gene in tessuti diversi dell' organismo o in diversi stadi di sviluppo, le caratteristiche variazioni di sequenza del DNA con significato patogeno, *et cetera.* Ciò fu realizzato pochi mesi prima dell' annuncio ufficiale della conclusione anticipata del progetto, per merito di un gruppo di giovani informatici guidato da Jim Kent all' Università di California, Santa Cruz (UCSC), che riuscirono a produrre "Human Genome Browser"(http://genome.ucsc.edu/); parallelamente ed in modo indipendente, un gruppo di informatici dell' European Bioinformatic Institute produsse "Ensembl" (http://www.ensembl.org/index.html). I due programmi, accessibili gratuitamente in rete, permettono oggi di navigare facilmente, punto per punto, lungo i diversi cromosomi umani, per una lunghezza complessiva di oltre 3 miliardi di nucleotidi, fornendo tutte le informazioni disponibili su ciascun segmento di sequenza selezionato dall' utente.

Un archivio elettronico di dati sul genoma umano, accessibile on line gratuitamente per i ricercatori ed a pagamento per le imprese, e' "*Gene Cards*" (http://www.genecards.org/), costruito nel 1997 da Doron Lancet e Vered Chalifa-Caspi all'Istituto Weitzman, per fornire ai ricercatori l' insieme delle informazioni relative a ciascun "gene" umano, integrando in un unico *web-accessible knowledge base* tutta l' informazione biomedica disponibile**.** Nell' archivio, oggi molto arricchito rispetto alla versione originale e continuamente aggiornato, ogni dato e' legato con uno specifico *link* al database di origine, consentendo così il rapido approfondimento di qualsiasi aspetto di ogni domanda dell'utente.

Tra i moltissimi database che raccolgono informazioni sulle sequenze del DNA umano, questi tre esempi possono dare un' idea dell' enorme mole di dati oggi disponibile sul genoma umano e della complessità della loro articolazione. Sono poi disponibili numerosi database relativi a genomi di altri organismi (molti microorganismi, lieviti, vegetali ed animali), che permettono analisi comparative e la ricostruzione delle relazioni evolutive di singoli geni e dei genomi dei diversi organismi.

Un settore molto attuale degli studi informatici sul DNA, che sfrutta anche l' analisi comparativa di genomi di organismi diversi, è l' impiego di nuovi algoritmi per individuare la presenza di brevi serie di nucleotidi con potenziale significato di regolazione sul funzionamento del DNA stesso. Un' altra recente applicazione dell' informatica alla ricerca genomica riguarda lo studio dell' espressione differenziale dei geni nei diversi tessuti dell' organismo o nello stesso tessuto in diverse condizioni (ad esempio in un tessuto sano rispetto a quello che ha subito una trasformazione neoplastica), utilizzando dati ottenuti mediante microarrays di DNA. Questi studi, basati sull'analisi di grandi matrici di dati (serie di geni diversi per serie di tumori diversi o di pazienti diversi che presentino il medesimo tipo di tumore), comportano metodi informatici sia per la gestione di tali dati che delle relative analisi statistiche. Essi hanno già generato applicazioni molto interessanti, come ad esempio la individuazione di chemioterapici adatti per specifici tipi di tumori.

L' affrontare questo tipo di problemi sta portando la bioinformatica ad occuparsi sempre di più di aspetti complessi ed inesplorati della realtà biologica.

La definizione di "sistema complesso" ben si attagliano ai sistemi biologici: un sistema complesso e' infatti formato da un alto numero di componenti indipendenti in grado di interagire; è fortemente strutturato e la sua struttura presenta variazioni; la struttura del sistema e' suscettibile di trasformazione (evoluzione) nel tempo; tale modificazione dipende dalle condizioni iniziali, da perturbazioni e dall' esistenza di molteplici soluzioni che rendano compatibile la stabilità strutturale del sistema stesso rispetto a diverse condizioni di contesto; inoltre il grado di complessità dell' oggetto non è direttamente dipendente dal numero dei componenti da cui e' formato, ma piuttosto dal numero delle loro potenziali inter-relazioni.

Le metodologie biochimiche, genetiche e di biologia molecolare hanno consentito di identificare i componenti cellulari, sub-cellulari e sub-microscopici degli oggetti biologici, fino a decifrare i processi che sottendono a fenomeni apparentemente molto complessi come la memoria. Il recente e rapidissimo progresso delle conoscenze di genomica, epressiomica, proteomica e l' applicazione dell' informatica avvicinano il momento in cui, sia pure in sistemi biologici a bassa complessità, sarà possibile definire input ed output di ogni determinato elemento del sistema ed assegnare a ciascun elemento valori di *input* ed *output* osservati sperimentalmente. E' opportuno sottolineare a questo proposito che i metodi

computazionali già oggi disponibili e la potenza degli attuali calcolatori sono in grado di affrontare simulazioni di notevole complessità. Sarà così possibile costruire dei modelli di organismi ed analizzare le conseguenze di specifiche manipolazioni prima di eseguire qualsiasi esperimento in laboratorio. Le potenzialità di questo tipo di modellizzazione sono enormi.

La decifrazione della complessità di un sistema biologico è oggi la sfida più importante per chi si occupi di Biologia. Non può sfuggire che l' informatica, ha creato le basi concettuali, gli strumenti, ma soprattutto l' ambiente adatto allo sviluppo di questo tipo di ricerca. Fino a qualche anno fa chi si occupava di bioingegneria cercava di interpretare singoli fenomeni biologici nel contesto di una cultura scientifica riferita alla meccanica e all' elettronica; l' informatica ha creato un nuovo contesto culturale, quello dell' esplorazione dei sistemi complessi. Non e' casuale che il modello di sistema "scale-free", attualmente molto studiato per l' applicazione in diversi campi ed in particolare in Biologia, sia stato concepito nel 2000 da R. Albert, H. Jeong e A.L. Barabasi analizzando la struttura della rete che connette tutti i computer del pianeta.

Oggi appare assolutamente normale l' analisi di enormi masse di dati, la loro organizzazione e classificazione e la loro condivisione indipendente dalle distanze geografiche. Molti ricercatori in ambito biologico vivono in questa dimensione senza rendersi conto delle enormi potenzialità intrinseche dell' informatica per il loro stesso campo di indagine, sfruttando la bioinformatica soltanto come utenti di specifici *software*. Sono infatti convinti che il tradizionale lavoro parcellizzato, analitico e metodico sia ancora in grado di dare risposte adeguate ad una competizione sempre più spinta tra i paesi tecnologicamente avanzati. Non si rendono conto che mediante metodi informatici non troppo sofisticati è possibile raggiungere in breve mete molto ambiziose sfruttando elementi di conoscenza già disponibili. Per esempio, tradizionalmente la ricerca di un farmaco in grado di agire su uno specifico recettore cellulare si effettua provando uno dopo l'altro l'effetto di diversi composti, ritenuti interessanti sulla base dei dati di letteratura; è tuttavia possibile un percorso alternativo: ottenere una descrizione informatica della struttura molecolare del recettore e lanciare tale informazione strutturale contro tutte le informazioni strutturali di tutti i composti chimici disponibili: viene così individuato un piccolo gruppo di molecole potenzialmente in grado di interagire con il recettore, alcune delle quali non sarebbero mai state individuate dalla ricerca bibliografica come possibili farmaci.

E’ sbagliato credere che la sfida nel settore informatico sia limitata ai Paesi fortemente industrializzati: a differenza delle classiche attività sperimentali (chimica, fisica, biologia), l’informatica ha costi di investimento in attrezzature relativamente modesti ed il successo dipende molto fortemente dal fattore umano. Investire in capitale umano e’ quindi fondamentale. Tramutare l’ entusiasmo di tanti giovani appassionati di informatica in capacità di lavoro a livelli di eccellenza e’ una sfida che il nostro Paese può raccogliere anche in un momento di limitate disponibilità economiche, perchè in Italia il livello degli insegnamenti universitari in questo settore è ottimo, se non eccellente. Sarebbe un gravissimo errore politico puntare all’ alfabetizzazione informatica a livello scolastico senza potenziare fortemente le strutture didattiche e di ricerca a livello universitario in un settore strategico tanto importante per il futuro del nostro Paese. Ci potremmo ridurre infatti in breve tempo ed in modo irrimediabile da potenziali produttori di innovazione in consumatori delle innovazioni altrui.