

BIOINFORMATICA E METODI COMPUTAZIONALI IN BIOLOGIA

Giancarlo
Mauri

Biologia e informatica: convergenze nel XXI secolo

Se la scienza dominante nel ventesimo secolo è stata senza dubbio la fisica, con le grandi costruzioni teoriche della relatività e della meccanica quantistica e le fondamentali (nel bene e nel male) ricadute applicative (dall'energia nucleare ai viaggi spaziali), affiancata nella seconda metà del secolo dall'informatica, all'inizio del nuovo secolo l'informatica promette di continuare a giocare un ruolo centrale nel progresso dell'umanità, insieme ad una biologia in rapida crescita sia sul piano scientifico e concettuale che su quello tecnologico.

I mezzi di comunicazione di massa rispecchiano fedelmente questa situazione, occupandosi continuamente sia dell'informatica che della biologia e delle biotecnologie, di cui danno un'immagine di disciplina di frontiera estremamente dinamica con cui, nel bene e nel male, nei prossimi anni bisognerà fare i conti e che, soprattutto, potrebbe consentire di risolvere o almeno di alleviare molti dei problemi che l'umanità si trova a dover affrontare, dalle malattie alla carenza di cibo.

I successi delle biotecnologie nascono anche da una comprensione sempre più profonda delle basi molecolari dei fenomeni biologici, e quindi dalla crescente capacità di controllarli, in cui l'informatica gioca un ruolo non secondario. Negli ultimi dieci anni si è infatti assistito ad una crescente convergenza tra informatica e biologia, con importanti apporti concettuali e strumentali in ambedue le direzioni.

Da un lato, gli informatici si sono ormai resi conto che la natura utilizza sistemi di elaborazione dell'informazione estremamente efficienti e sofisticati a tutti i livelli, e stanno cercando di capirli e di imitarli per costruire macchine più potenti ed efficienti. Ad esempio, benchè il calcolatore elettronico sia incomparabilmente più veloce, soprattutto nell'eseguire calcoli, rispetto al cervello umano, ed abbia

una capacità di memorizzazione estremamente più elevata, per molti compiti che non solo gli uomini, ma tutti gli animali eseguono in modo naturale esso è ancora inadeguato: il controllo del movimento in ambiente variabile, il riconoscimento di persone, oggetti, suoni (anche in presenza di rumore), la comprensione del linguaggio, la capacità di adattarsi a nuove situazioni, di imparare dall'esperienza. Sono tutti aspetti del "comportamento intelligente" che il calcolatore non è in grado di imitare, perchè non sono riducibili in modo immediato alla esecuzione di calcoli. Per questo negli ultimi quindici anni sono emersi modelli e tecniche computazionali come il connessionismo e le reti neurali che si ispirano ai meccanismi neurofisiologici e puntano a riprodurre artificialmente la struttura e il funzionamento del cervello, e che hanno avuto un notevole successo anche applicativo.

Anche l'evoluzione delle specie può essere letta in chiave computazionale, e questa lettura sta alla base della teoria degli algoritmi genetici, che consentono di risolvere problemi di ottimizzazione altrimenti intrattabili. Il meccanismo dell'evoluzione permette di trovare buone soluzioni all'interno di uno "spazio delle soluzioni possibili" tanto numerose che un loro esame esaustivo richiederebbe tempi inconcepibili; basta prendere una piccola "popolazione" di soluzioni (ad esempio, mille all'interno di un insieme di mille miliardi di possibili soluzioni), scegliere le migliori tra esse (le meglio adattate all'ambiente, in termini biologici) sulla base dei criteri predefiniti, e farle riprodurre incrociandone le caratteristiche. Dopo un certo numero di generazioni, si ottengono così individui con caratteristiche altamente positive.

Una terza importante area di interazione tra biologia e informatica è più recente, e parte dalla constatazione che la cellula è un raffinatissimo sistema per memorizzare, duplicare e trasmettere informazione.

In essa l'informazione è memorizzata nelle molecole di DNA, con una densità di mille miliardi di volte superiore a quella delle memorie elettroniche tradizionali, e viene elaborata con un consumo di energia dieci miliardi di volte più basso. Se si riuscisse a controllare questi processi, sarebbe possibile realizzare calcolatori molecolari in cui l'informazione è codificata attraverso biosequenze, di nucleotidi o di aminoacidi, che vengono manipolate e trasformate con tecniche standard di laboratorio, fino ad estrarre i risultati desiderati, in grado di sostituire in molti casi con vantaggio gli attuali calcolatori elettronici. Dall'altro lato, la biologia ha sfruttato e sta sfruttando in modo sempre più massiccio gli strumenti offerti dall'informatica per risolvere i propri problemi. È così nata una nuova disciplina, la bioinformatica o biologia computazionale, che si fa carico della organizzazione e dell'analisi dei dati biologici e dello sviluppo di metodi matematico-statistici e computazionali per la caratterizzazione funzionale di biosequenze.

Naturalmente, ciò significa anche nuove figure professionali (non è raro ormai trovare richieste di bioinformatici tra le valanghe di annunci che offrono posti per web master o esperti di Internet) con le relative esigenze di formazione. Per questo, molte Università di nazioni come gli Stati Uniti, il Giappone, la Germania, la Francia, Israele offrono specifici corsi di formazione o specializzazione per bioinformatici, e finanziano in modo sostanzioso centri e programmi di ricerca in questo settore. L'Italia purtroppo sembra non aver ancora compreso fino in fondo l'importanza strategica della bioinformatica, e non riesce ad andar oltre gli sforzi isolati di alcuni volenterosi ricercatori, con il rischio di accumulare un ritardo incolmabile rispetto alle altre nazioni.

La "corsa" al genoma umano

Recentemente, il ruolo essenziale della bioinformatica è stato sottolineato dall'annuncio del completamento del sequenziamento del genoma umano.

Tutti i "piani di costruzione" di ogni essere vivente sono contenuti nel suo genoma, un "libro" scritto con sole quattro "lettere" chimiche (A, C, G, T) chiamate basi o nucleotidi e costituito, per l'uomo, da tre miliardi di caratteri (ci vorrebbero circa un milione di pagine per scriverli su carta) raccolti in lunghe catene di DNA, che spiega in dettaglio come costruire le varie parti dell'organismo con le relati-

ve caratteristiche. Il genoma umano contiene, dispersi in un mare di parole apparentemente senza senso, circa 30.000 parole "sensate", i geni, ognuno dei quali contiene le istruzioni per produrre una proteina. In certe situazioni, un gene attiva la produzione della proteina corrispondente, consentendo all'organismo di formarsi, svilupparsi e sopravvivere. È evidente che se sapessimo leggere questo libro, individuando al suo interno le pagine relative ad ogni singola parte dell'organismo, e riuscendo ad associare ad ogni caratteristica osservabile di un uomo i geni responsabili, potremmo riconoscere ed eventualmente correggere eventuali errori di progettazione, e quindi curare le malattie e le malformazioni di origine genetica, ma anche produrre medicinali "genetici" in grado di curare malattie degenerative precedentemente incurabili.

Nel 1990 un consorzio pubblico finanziato tra gli altri dal governo degli Stati Uniti e dall'inglese Wellcome Trust ha lanciato il Progetto Genoma Umano (HGP), in cui sono coinvolti più di mille biologi ed informatici di sei paesi, con l'obiettivo di arrivare entro il 2005, data successivamente anticipata al 2002, ad avere la sequenza completa del genoma umano. A sorpresa, nel mese di aprile del 2000 la società Celera Genomics, fondata solo un anno e mezzo prima, nel settembre 1998, da Craig Venter, fuoriuscito da HGP, ha annunciato di aver completato il sequenziamento del genoma umano, anticipando quindi di ben due anni lo HGP stesso. In realtà, la sequenza non era ancora completa, ed era stata ottenuta anche sfruttando dati prodotti da HGP, per cui, dopo un paio di mesi di roventi polemiche, si è arrivati ad un comunicato congiunto tra i due gruppi (Celera e HGP) nel giugno 2000, ed alla pubblicazione definitiva della sequenza nel febbraio 2001. Resta il fatto che l'obiettivo è stato raggiunto con quasi 5 anni di anticipo rispetto alla previsione iniziale.

Il fattore che ha consentito di ridurre drasticamente i tempi previsti è stato l'utilizzo massiccio di sistemi di calcolo di enorme potenza (Celera dispone di 300 workstations del valore di 250.000 dollari l'una) che hanno fatto uso di programmi sviluppati appositamente per l'elaborazione dei dati di sequenziamento e basati su raffinate tecniche algoritmiche. In particolare, la metodologia utilizzata da Celera consiste nello spezzettare a caso alcune copie della sequenza in frammenti della lunghezza media di 350 coppie di basi che vengono sequenziati. Si cerca poi di riordi-

nare i frammenti di DNA in una sequenza coerente, cercando di individuare parti di essi che si sovrappongono: è un po' come ricostruire un puzzle da 50 milioni di pezzi, che per di più possono essere girati sulle due facce.

Le prospettive della bioinformatica

La ricostruzione della sequenza di interi genomi, tra cui il genoma umano, è un risultato di enorme importanza, ma è solo il primo passo di un lungo percorso, e i passi successivi non sono certo più semplici. Oggi i biologi dispongono di una quantità enorme di dati, che non si limitano alle sequenze genomiche. Altri dati importanti riguardano le proteine e la loro struttura. Anche le proteine possono essere viste come parole scritte usando un alfabeto formato da venti lettere (gli aminoacidi), anziché da sole quattro lettere come il genoma. Alcune proteine sono corte, costituite da una catena di pochi aminoacidi, ma altre sono molto lunghe, con parecchie migliaia di aminoacidi, e attualmente si conosce la sequenza di oltre mezzo milione di proteine.

Infine, una terza rilevante sorgente di dati è costituita dai DNA microarrays, strumenti che consentono in un certo senso di “fotografare” i geni nel momento in cui, secondo la terminologia biologica, “si esprimono”, cioè avviano il processo di costruzione della proteina corrispondente.

Tutti questi dati, sequenze genomiche, sequenze proteiche, dati di microarrays, sono raccolti in grandi banche dati accessibili a tutti i ricercatori, ma sono

dati “grezzi” che devono essere organizzati, analizzati e collegati tra loro per estrarne il “significato” biologico, e per questo è ancora necessario usare strumenti di calcolo di grande potenza, con programmi che inglobano le più avanzate tecniche informatiche per l'organizzazione dei dati e l'analisi di sequenze, adattate alle specifiche esigenze della biologia.

In sintesi, si tratta di:

- capire la struttura del DNA nel nucleo, identificando al suo interno i geni, che ne occupano una frazione trascurabile (meno del 10%)
- capire come questa struttura governa la trascrizione del DNA e l'espressione dei geni, cioè la produzione delle proteine
- riconoscere all'interno del DNA quelle particolari sottosequenze (i “segnali”) che indicano l'inizio di un gene o ne controllano l'espressione
- cercare similarità tra i genomi di specie diverse per stabilire relazioni con sequenze già caratterizzate
- assegnare ai geni la loro funzione e spiegare la funzione in termini strutturali usando strutture note o derivate da modelli
- predire la struttura spaziale delle proteine, da cui dipendono le loro proprietà, a partire dalla conoscenza della sequenza di aminoacidi di cui sono composte.

A prima vista, il compito sembra semplice, ma in realtà non lo è affatto. I tentativi di decifrare gli indizi nascosti nei dati genomici sono resi difficili dalla scarsa affidabilità dei metodi di predizione di geni



in sequenze di DNA non caratterizzate; è presuntuoso pensare di poter attribuire funzioni solo sulla base di un certo grado di similarità tra sequenze (e non sempre è chiaro cosa si intenda per “funzione”); si conosce un numero di strutture molto ridotto rispetto al numero di sequenze, e i metodi di predizione di strutture sono inaffidabili; le banche dati contengono informazioni non omogenee o imprecise. Vediamo allora, per ciascuno dei punti sopra elencati, quali sono i principali problemi aperti.

Strumenti e problemi aperti

Per quanto riguarda la identificazione dei geni all'interno del genoma, le tecniche usate si basano sulla ricerca di “segnali” (cioè particolari sottosequenze) nella sequenza, sulle statistiche di contenuto (cioè sulla percentuale di ciascuna base e sulla distribuzione delle basi all'interno della sequenza) e sulla similarità con geni noti. In teoria, il compito di trovare particolari sottosequenze all'interno di una sequenza data è considerato un compito semplice, per cui sono noti parecchi algoritmi efficienti. In pratica, quando si deve trattare con sequenze di DNA le cose si complicano sia per la lunghezza della sequenza, sia perchè la sequenza può comparire con piccole varianti, cioè con alcune basi aggiunte, eliminate o sostituite.

In un recente test di valutazione di strumenti per l'identificazione di geni eseguito su parte del genoma della *Drosophila*, la maggior parte di questi strumenti ha identificato il 95% dei nucleotidi codificanti, ma le strutture di introni/exoni sono state predette correttamente solo per il 40% circa dei geni. I diversi metodi non sono riusciti a trovare tra il 5% e il 95% dei geni, e ne hanno identificati in modo errato fino al 55%. Una ulteriore prova della debolezza dei metodi di predizione dei geni è data dalla incertezza tuttora esistente sul numero di geni nel genoma umano: anche se le stime attualmente più accreditate danno un numero di circa 30.000, fino a poche mesi fa si parlava di oltre 100.000 geni. Probabilmente, l'ostacolo maggiore per un conteggio esatto dei geni sta nel fatto che la definizione stessa di gene non è chiara: è una unità ereditabile corrispondente ad un fenotipo osservabile (come il colore degli occhi) ? o un pacchetto di informazioni che codifica una o più proteine ? o che codifica RNA ? Ancora, i geni sono geni se non sono espressi ?

Alcune delle tecniche di analisi di sequenze utilizzate per “decifrare” il genoma sono applicabili

anche alle sequenze proteiche. Tuttavia, nell'analisi delle proteine è necessario tener conto anche di altri aspetti, ad esempio capire come interagiscono tra di loro e con altre molecole, ma per questo le informazioni essenziali sono contenute nella struttura spaziale delle proteine. Infatti le sequenze proteiche si avvolgono e si ripiegano su se stesse in una precisa struttura tridimensionale, ma delle oltre 500.000 proteine note solo di circa 2.000 si conosce la struttura. I metodi di predizione delle strutture proteiche vanno da strategie che simulano le forze chimiche e fisiche che determinano la struttura, e richiedono l'esecuzione di enormi quantità di calcolo, ad approcci che cercano di usare informazioni estratte dai database di strutture, o somiglianze od omologie con sequenze di struttura già nota. Tuttavia, i programmi relativi sono costosi, richiedono tempi di calcolo lunghissimi e solo in una piccola percentuale di casi danno risultati realistici; il problema rimane quindi sostanzialmente insoluto, tranne che per proteine di piccole dimensioni.

L'identificazione dei geni e delle strutture delle proteine non basta ancora per farci capire fino in fondo, ed eventualmente controllare, i meccanismi che regolano la vita. Infatti si tratta di informazioni separate, isolate, mentre è sempre più evidente che la funzione di un gene, o di una proteina, dipende da una complessa rete di influenze e interrelazioni reciproche, e può essere diversa in contesti diversi. Se vogliamo ottenere il massimo dai dati genomici, dobbiamo tener conto delle informazioni sulla regolazione dell'espressione genica, sui cammini metabolici e sulle cascate di segnali. Le proteine non lavorano isolatamente, ma fanno parte di reti complesse. Scoprire queste reti e le loro interazioni è di vitale importanza per comprendere lo sviluppo normale e patologico delle cellule. Per questo, la bioinformatica dovrà predisporre database con un forte livello di integrazione e interoperabilità che permettano all'utente di ragionare su sorgenti di dati diversificate e di estrarre da esse conoscenze nuove e significative.