Rendiconti
Accademia Nazionale delle Scienze detta dei XL
Memorie di Scienze Fisiche e Naturali
122° (2004), Vol. XXVIII, pp. 285-289

ANNA TRAMONTANO *

La complessità delle funzioni biochimiche: sfide computazionali

Un organismo è il risultato di un insieme complesso di funzioni svolte dai suoi componenti basilari, essenzialmente acidi nucleici e proteine, e dalle loro interazioni.

La biologia moderna ha una conoscenza quasi completa del materiale genetico che codifica per le componenti di molti organismi viventi, non solo virus e batteri ma anche organismi superiori come l'uomo [1, 2] e quindi le si presenta un'opportunità, che è anche una sfida, senza precedenti: quella di identificare tutte le componenti della vita e di comprenderne la funzione molecolare e cellulare. L'interesse a risolvere questo problema è pari solo alla sua difficoltà, dovuta alla sua estrema complessità e alla nostra limitata comprensione delle regole che governano la vita.

Un genoma è composto da una sequenza di quattro diversi nucleotidi, quello umano ne conta circa tre miliardi che dovrebbero contenere le istruzioni per almeno centomila diverse proteine. L'uomo è solo una delle tante specie di cui si conosce la sequenza genomica e quindi è facile comprendere che la quantità di dati disponibili è veramente enorme. Il problema, però, non è nella dimensione dei dati, che si può avvalere dei grandi sviluppi sia nell'hardware che nel software degli ultimi anni. Piuttosto il problema è nella tipologia dei dati biologici. Questi vanno da sequenze a strutture tridimensionali ad annotazioni funzionali. La vera difficoltà è di cercare di estrarre informazioni funzionali, che sono quelle che ci occorrono per poter comprendere ed eventualmente interferire con i meccanismi biologici, correlando tra loro dati di tipo così diverso.

La funzione di una macromolecola è il risultato dell'azione di atomi che si trovano in specifiche posizioni relative così che possano per esempio legare un substrato, riconoscere e legare una sostanza estranea all'organismo, regolare quali regioni del genoma vengono utilizzate, costruire impalcature strutturali per ottenere

* Dipartimento di Scienze Biochimiche «A. Rossi Fanelli», Università «La Sapienza» - P.le Aldo Moro, 5 - 00185 Roma.

l'organizzazione e la compartimentalizzazione cellulare, ecc. Questi meccanismi possono avere una estrema specificità e una squisita sensibilità così da rispondere a stimoli interni ed esterni, permettendo ad un organismo di sopravvivere ed evolvere.

La maggior parte di queste funzioni sono svolte dalle proteine, polimeri lineari di piccole molecole organiche, gli amino acidi, la cui sequenze specifica è determinata dalla sequenze del genoma. Le proteine assumono una struttura tridimensionale che determina la posizione relativa dei loro atomi e quindi la loro funzione.

La struttura di una proteina dipende, in linea di massima, solo ed esclusivamente dalla sua sequenza di aminoacidi, come fu dimostrato da Anfinsen in uno storico esperimento. Egli dimostrò che una proteina (nel suo caso la Ribonucleasi) una volta denaturata (cioè destrutturata) riguadagna in vitro la sua conformazione funzionale quando gli agenti chimici denaturanti sono eliminati [3]. Questo implica che l'informazione sulla struttura tridimensionale di una proteina è contenuta nella sua sequenza aminoacidica.

La funzione biologica è perciò direttamente legata al materiale genetico attraverso la struttura, e quindi la sequenza, delle proteine codificate dal genoma.

Il compito della biologia computazionale è quindi quello di identificare se una data regione genomica codifica per una proteina, come questa raggiunga la corretta struttura tridimensionale e come questa struttura sia in grado di svolgere la sua funzione da sola o attraverso le sue interazioni con altre molecole [4].

Negli organismi superiori, l'identificazione delle regioni genomiche che codificano per una proteina (i geni) non è ovvia. Nell'uomo le regioni geniche codificanti per proteine xono solo il tre per cento del genoma e perdipiù esse non sono contigue. Il problema è complesso e i metodi, generalmente statistici, che sono stati sviluppati sono certamente interessanti, ma non saranno discussi qui, dove invece focalizzeremo l'attenzione sul problema della «predizione» della struttura di una proteina data la sua sequenza.

La forza che spinge un polipeptide ad assumere una struttura definita è la necessità di stabilizzarsi cioè di raggiungere uno stato a minima energia: una proteina assume una certa struttura perché le interazioni che ne conseguono minimizzano la sua energia libera. Ovviamente deve anche esistere un percorso cinetico praticabile che porti la proteina dallo stato destrutturato a quello nativo. Se riuscissimo a generare tutte le possibili conformazioni di una proteina e a valutare correttamente la loro energia libera, potremmo semplicemente scegliere la conformazione a energia più bassa.

Abbiamo però due problemi.

Il primo è che la stabilità di una proteina è solo di qualche Kcal/mole (cioè più o meno dell'ordine di qualche legame idrogeno) e quindi, se anche generassimo tutte le possibili conformazioni di una proteina, dovremmo poi valutare molto accuratamente la loro energia per distinguere la struttura nativa da tutte le altre.

Il secondo, ovviamente, è che non possiamo generare tutte le possibili conformazioni di una proteina. Se immaginiamo di avere una proteina di 100 aminoacidi

e assumiamo che ciascuno degli aminoacidi possa assumere solo tre possibili conformazioni della sua catena principale, il numero di possibili conformazioni è 3^{100} cioè circa 10^{48}, un numero intrattabile [5].

Questo rende il problema della predizione «ab initio», cioè esclusivamente sulla base di calcoli energetici della struttura di una proteina insolubile, almeno al momento.

Nonostante ciò un'osservazione ci permette di predire la struttura di un gran numero di proteine: esse sono il prodotto di un'evoluzione molecolare e famiglie di proteine in specie diverse simili si sono evolute da un ancestore comune. Questo implica che le variazioni in ciascun passo evolutivo sono state compatibili con la funzione della proteina, o almeno con il fatto che la proteina potesse svolgere una funzione. La funzione di una proteina è di solito svolta da pochi specifici residui, ma, affinché questi residui siano nel posto giusto, il resto della sequenza deve garantire una struttura nativa stabile. Quindi proteine con sequenza simile non possono che avere struttura simile e questo può essere sfruttato per costruire modelli tridimensionali di proteine per omologia [6].

Un'altra utile osservazione che si permette di estendere il campo di applicabilità dei metodi di predizione di struttura è che la relazione tra sequenza e struttura non è biunivoca: proteine con sequenza simile hanno struttura simile ma spesso proteine con sequenza diversa hanno strutture simili [7]. La natura sembra riutilizzare le stessa architettura più e più volte per ottenere funzioni diverse. Anche questa osservazione può essere sfruttata per costruire modelli tridimensionali di proteine con i metodi detti di riconoscimento di fold.

Occorre però valutare l'affidabilità di questi metodi per sapere se e in quali casi un modello tridimensionale di una proteina è sufficientemente accurato per specifiche applicazioni.

Ovviamente, per valutare l'efficacia di un metodo di predizione, è necessario confrontare la predizione con il risultato sperimentale, però la maggior parte dei metodi di predizione sono euristici cioè basati su statistiche e osservazioni di strutture note di proteine, quindi in qualche modo contengono già informazioni sulle strutture che sono usate per testare il metodo. Questo implica che non sempre si può essere certi che il metodo non predica meglio la struttura di proteine già note rispetto a proteine non ancora determinate sperimentalmente.

Questo problema è affrontato (e parzialmente risolto) dalla comunità scientifica internazionale istituendo un esperimento a scala mondiale chiamato CASP (Critical Assessment of Methods for Structure Prediction) che valuta i risultati di predizioni di proteine di struttura non nota la cui struttura sarà però determinata in tempi ragionevolmente brevi [8-12].

Ogni due anni viene chiesto a cristallografi e spettroscopisti NMR che stanno per risolvere la struttura di una proteina di rendere disponibile la sua sequenza.

Si annunciano poi le sequenze di queste proteine target insieme con la data in cui ci si aspetta che la struttura sia disponibile. I predittori quindi depositano i loro modelli prima che la struttura sia resa pubblica.

Un pannello di valutatori ('assessors') confronta modelli e strutture appena queste ultime sono disponibili e cerca di valutarne la qualità e di trarre conclusioni generali. C'è un assessor per ciascuno dei metodi che abbiamo brevemente decritto: modelling per omologia, riconoscimento di fold e *'ab initio'*, che in questo contesto è intesa come la predizione di proteine la cui struttura non somiglia a quella di nessuna proteina nota.

I risultati sono discussi in un convegno dove si danno appuntamento gli assessor e i predittori e i cui risultati sono poi resi pubblici sia via Web sia attraverso pubblicazioni scientifiche.

Ci sono stati esperimenti CASP nel 1994 (CASP1), nel 1996 (CASP2), nel 1998 (CASP3) e nel 2000 (CASP4) e i loro risultati sono stati di enorme utilità per comprendere sia quale è la capacità attuale dei metodi di predizione sia in che direzione vanno concentrati gli sforzi futuri.

Va però notato che la valutazione dei risultati ha un elemento di soggettività dovuta al ruolo dell'assessor, in quanto essa dipende da cosa viene considerato di buona qualità. Per esempio, a seconda dei casi, può essere preferibile predire l'intera struttura di una proteina target oppure limitarsi alle regioni per cui si ritiene di poter produrre un buon modello.

Ci sono valori numerici che si possono calcolare come la radice quadratica media della deviazione (r.m.s.d.) tra gli atomi del modello e della proteina dopo che le due strutture siano state sovrapposte nel modo migliore possibile ma la soluzione del problema di sovrapporre in modo ottimale due strutture non è unica.

Nonostante ciò gli esperimenti CASP si sono dimostrati utilissimi, soprattutto per dare indicazioni su quali aspetti dei vari processi di modelling rappresentano il collo di bottiglia per ottenere predizioni corrette e il campo si sta evolvendo rapidamente con risultati di tutto rispetto.

Questo è testimoniato anche dalla nascita di progetti di cooperazione internazionale noti con il nome di «Genomica strutturale» che mirano a risolvere sperimentalmente il maggior numero possibile di strutture di molecole biologiche nel più breve tempo possibile, selezionandole in modo mirato, cioè scegliendole razionalmente tra quelle che appartengono a famiglie evolutive su cui non vi sono informazioni strutturali così che esse possono essere usate per modellare tutte le proteine della famiglia.

La corretta identificazione di quali proteine sono omologhe e la capacità di costruire modelli affidabili sono ovviamente gli elementi centrali perché lo sforzo abbia successo ed in entrambi la biologia computazionale gioca un ruolo fondamentale.

La sfida futura è la simulazione accurata e la comprensione di tutte le funzioni cellulari, e, vista la velocità con cui nuovi traguardi vengono raggiunti ogni giorno, ci si può ragionevolmente aspettare di ottenere risultati preliminari, ma ugualmente interessantissimi, in tempi brevi.

REFERENZE

[1] Lander, E.S. et al. (2001), *Nature* 409, 860-921.
[2] Venter, J.C. et al. (2001), *Science* 291, 1304-1351.
[3] Anfinsen, C.B., Harrington, W.F., Hvidt, A. and Lindstrom-Lang, K. (1955), *Biochim. Biophys. Acta* 17, 141-142.
[4] Tramontano, A. (2001), Zanichelli, Bologna.
[5] Levinthal, C. (1968), *Journal de Chimie Physique et de Physico-Chimie Biologique* 65, 44-45.
[6] Chothia, C. and Lesk, A. (1986), *EMBO J.* 5, 823-826.
[7] Chothia, C. (1992), *Nature* 357, 543-544.
[8] Moult, J., Pedersen, J., Judson, R. and Fidelis, K. (1995), *Proteins* 23, ii-v.
[9] Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J. (1999), *Proteins Suppl.* 3, 2-6.
[10] Moult, J., Hubbard, T., Bryant, S., Fidelis, K. and Pedersen, J. (1997), *Proteins Suppl.* 1, 2-6.
[11] Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001), *Proteins Suppl.* 5, 2-6.
[12] Moult, J., A., Z., Fidelis, K. and Hubbard, T. (2003), *Proteins Suppl.* 6, 334-339.