La Matematica nella Società e nella Cultura

Rivista dell'Unione Matematica Italiana

Federica Vitale

Modelli matematici per la predizione statistico-computazionale della struttura nativa delle proteine

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Serie 1, Vol. **2** (2009), n.2 (Fascicolo Tesi di Dottorato), p. 311–314.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=RIUMI_2009_1_2_2_311_0>

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Unione Matematica Italiana, 2009.

La Matematica nella Società e nella Cultura
Rivista dell'Unione Matematica Italiana
Serie I, Vol. II, Agosto 2009, 311-314

Modelli matematici per la predizione statistico-computazionale della struttura nativa delle proteine

Federica Vitale

1. – Un approccio geometrico al protein folding

Il problema conosciuto nella comunità scientifica con il nome di *protein folding* costituisce oggi uno dei problemi più importanti della biologia molecolare [1-2].

È stato di fatto osservato che ogni proteina è ripiegata in un modo che le è caratteristico, e si è compreso che questo ripiegamento (*folding*) è la condizione fondamentale affinché essa possa svolgere la funzione che le è propria. La forma della proteina ripiegata quando si trova nel suo ambiente naturale viene chiamata *struttura terziaria*, o *configurazione nativa*, o anche *struttura tridimensionale naturale* della proteina. Il riconosciuto legame tra questa struttura e la funzione della proteina rende fondamentale la sua conoscenza e dunque di primaria importanza *predire* attendibilmente la struttura tridimensionale di una proteina qualora sia nota la sola struttura primaria e sia impossibile risolverla sperimantalmente.

I metodi oggi considerati più efficaci, e di conseguenza più comunemente utilizzati, per tentare di approssimare una configurazione tridimensionale plausibile si dividono fondamentalmente in tre gruppi: *homology modelling*, *fold recognition* e metodi *ab initio* [3-4]. I primi due sono a base comparativa. In particolare, il primo è applicabile solo quando la proteina a sequenza nota è *omologa* a qualche proteina con struttura nota, mentre il secondo si applica nel caso in cui la sequenza target non presenta similarità con qualche sequenza in banca dati che può dunque essere considerata un potenziale templato. I metodi *ab initio*, invece, lavorano su calcoli energetici che hanno lo scopo di scegliere la conformazione nativa, tra tutte quelle possibili, come quella a energia minima.

Ciascuno di questi metodi, considerato da solo, presenta ovvi e notevoli difetti sia operativi che concettuali. Ad esempio, i metodi di tipo comparativo mancano di un'analisi statistica preliminare completa e basata su un'opportuna codifica delle informazioni di carattere geometrico sulle quali condurre l'inferenza statistica. D'altra parte, i metodi *ab initio* oltre ad essere spesso improponibili sotto l'aspetto operativo dato il numero proibitivo delle coordinate che identificherebbero la configurazione di equilibrio di una proteina, presentano anche il difetto di limitare la descrizione delle interazione tra amminoacidi alle classiche forze gravitazionali, elettromagnetiche e di Van der Waals, concepite classicamente come binarie. In tal modo, si trascurano completamente le possibili interazioni multiple

e si rinucia a priori alla possibilità che una proteina che si ripiega sia interpretato come un *sistema complesso*.

In vista di queste osservazioni ci è parso opportuno presentare una metodologia che riproponesse gli aspetti chiaramente positivi e meglio fondati di tutte le tecniche note, sintetizzandole e cercando di eliminarne i difetti. Così, ciò che proponiamo è un procedimento di tipo comparativo basato su una rigorosa codifica delle proprietà geometriche intrinseche essenziali a descrivere la configurazione tridimensionale di qualsiasi sistema articolato la quale è deputata a dare una descrizione probabilistica degli effetti delle interazioni tra amminoacidi che solo da questi effetti vengono descritte senza alcuna ipotesi a priori sulla loro forma di leggi binarie classiche.

Punto di partenza della nostra analisi sono stati i file conservati nella *Protein Data Bank* (PDB), dai quali abbiamo estrapolato due informazioni: quella relativa alla sequenza proteica e quella relativa alle coordinate atomiche.

A ciascuna sequenza proteica abbiamo associato il $\mathcal{P}$*-profilo*, ovvero una codifica numerica della sequenza in modo da descrivere ciascun amminoacido, le sue caratteristiche chimiche ($\mathcal{P}$*-profilo chimico*), funzionali ($\mathcal{P}$*-profilo funzionale*), idrofobiche/idrofiliche ($\mathcal{P}$*-profilo idrofobico/idrofilico*) [5-6]. Fissata poi una qualunque posizione j sulla sequenza, abbiamo introdotto le restrizioni dei profili all'insieme $\{1, 2, \ldots, j-1\}$, che nel nostro modello rappresentano i *background chimici, funzionali, idrofobici/idrofilici* di una sottosequenza, ciò che potremmo interpretare come la *storia* della sottosequenza e che vengono utilizzati come strumento di confronto tra due proteine. A tale scopo è stato necessario introdurre una funzione *distanza* che permette di strutturare a spazio pseudo-metrico l'insieme dei profili [7].

Per quanto concerne la seconda informazione estrapolata dai file PDB essa viene utilizzata per completare il confronto tra proteine sulla base di parametri geometrici opportunamente costruiti [5]. Abbiamo a tal proposito osservato che, per un confronto tra proteine è necessario eliminare la possibilità di falsarlo a causa di *errori di parallasse*. Per chiarire di che cosa si tratta, osserviamo che le coordinate degli atomi di una qualsiasi proteina variano in generale con l'apparato sperimentale usato per determinarle. Ed è possibile immaginare il caso limite in cui *una stessa* proteina, osservata in due laboratori diversi, viene rappresentata con dimensioni diverse, semplicemente perché un apparato sperimentale l'ha vista *ruotata* oppure *traslata* rispetto alla posizione nella quale l'ha vista l'altro. Tuttavia, tra le coordinate dei diversi atomi sussistono delle relazioni *intrinseche* che vengono conservate nelle rototraslazioni, e sono queste che devono confrontarsi. In altre parole, sono le *posizioni relative* dei diversi atomi, e le proprietà di simmetria che ne conseguono, ciò che caratterizza ciascuna proteina. Per condurre il confronto *soltanto* su queste proprietà geometriche, ciascuna proteina si deve rappresentare in un opportuno riferimento. Il riferimento che è stato ipotizzato nel nostro schema di lavoro è il riferimento *centrale o di simmetria geometrica left-shifted* con assi orientati secondo le dimensioni crescenti degli assi di simmetria dell'ellissoide di simmetria geometrica. In un tale riferimento la proteina è totalmente contenuta nel primo ottante con

tre facce sovrapposte nei tre piani coordinati. Ottenuto questo nuovo set di coordinate abbiamo introdotto una procedura per il calcolo dei parametri geometrici necessari per il confronto che sono stati individuati in *curvatura* e *torsione* di tutte le quaterne di amminoacidi consecutivi che si possono ottenere *camminando* su una biosequenza con passo unitario. Dunque, per ogni fissata quaterna lungo una biosequenza abbiamo due angoli di curvatura ed uno di torsione. Per analizzare la distribuzione di questi angoli abbiamo fatto ricorso a tecniche di statistica circolare ridefinendo opportunamente gli indici di posizione e di dispersione per adattarli al linguaggio da noi introdotto.

L'informazione che si ottiene dallo studio statistico dei parametri geometrici unitamente alla precedente nozione di distanza permettere di valutare, in dipendenza dei valori che si ottengono in corrispondenza di un'assegnata coppia di classi d'equivalenza, la deformazione locale subita da una sottosequenza comune di amminoacidi nel passaggio dall'una all'altra classe. Poiché la misura di tale deformazione è di tipo statistico, ciò corrisponde ad una definizione di *stabilità* (o di *instabilità*) in senso statistico. L'introduzione della nozione di stabilità statistica di una funzione (o di un sistema di funzioni) di una k-upla di amminoacidi e delle sue occorrenze nelle diverse proteine, rispetto a perturbazioni del background, che peraltro fa uso di note tecniche di raffronto tra campioni, ha lo scopo di istituire un legame statistico tra i profili e i valori dei parametri geometrici che viene integrata con una procedura di previsione altamente probabile della differenza delle medie di una stessa funzione in corrispondenza di background diversi e in condizioni di instabilità. A partire da questa, siamo in condizione di associare, con elevata probabilità, valori dei parametri geometrici locali a tutte le posizioni di una qualsiasi proteina non osservata, anche se il background corrispondente a una particolare posizione e a una particolare sottosequenza di amminoacidi che da essa inizi non troverà corrispettivi nel catalogo delle proteine osservate [7].

Le nozioni e le parametrizzazioni introdotte sono state elaborate da un'infrastruttura tecnologica, appositamente progettata e realizzata per il calcolo di tutti i parametri geometrici in tutte le condizioni prescritte. Tale infrastruttura è identificata col nome di *DWH ProGeo Structure*.

L'intero DWH-ProGeo sarà sviluppato con un approccio *bottom-up*, ovvero assemblando iterativamente più *data mart*, ciascuno dei quali sarà incentrato su una specifica categoria di PDB. In particolare, verrà sfruttato il raggruppamento in *cluster* già presente nel data base alimentante il nostro sistema e dunque i singoli data mart rappresenteranno famiglie di proteine a struttura risolta e depositate nella banca dati proteica. In tale logica, è stato inizialmente implementato DWH-ProGeo^{Myo}, il data mart relativo alla famiglia di mioglobine [8].

DWH-ProGeo^{Myo} è di rilevante importanza in quanto è stato costruito non solo come prototipo dal punto di vista computazionale, ma anche e soprattutto dal punto di vista metodologico. Infatti, dal punto di vista computazionale, viene utilizzato come set di collaudo, e ci fornisce la possibilità di testare le procedure implementate su

catene non troppo lunghe (abbiamo osservato dalla letteratura e verificato sperimentalmente che la lunghezza media di questa proteine è di 152 amminoacidi), e di ottenere statistiche consistenti, visto che il campione che si può costruire con la famiglia delle mioglobine ha un'ampiezza di $n = 229$, che si amplia ad un campione di ampiezza $n = 263$ quando si considerano le mioglobine divise per catena. Dal punto di vista metodologico, d'altra parte, i risultati ottenuti con DWH-ProGeo^{Myo} hanno lo scopo di indirizzare gli sviluppi futuri, volti a considerare ogni singola proteina come un sistema complesso di particelle o individui interagenti. In un siffatto schema gli effetti delle interazioni dovranno essere descritti in termini stocastici, ossia fornendo (per ciascuna interazione) una distribuzione di probabilità sull'insieme delle sue possibili conseguenze. In questa prospettiva, DWH-ProGeo^{Myo} prima, e DWH-ProGeo più in generale poi, avranno il compito di guidarci nella definizione dell'opportuna faccia analitica da attribuire alla distribuzione di probabilità dei valori delle variabili di stato dopo le interazioni.

BIBLIOGRAFIA

[1] G. E. Schulz, *Principles of Protein Structure* (Springer, 1978).
[2] R. H. Pain (Ed.), *Mechanisms of Protein Folding* (Oxford University Press, 2000).
[3] M. S. Waterman, *Introduction to Computational Biology*, Chapman & Hall (London, 1995).
[4] C. Hardin, T. V. Pogorelov e Z. Luthey-Schultzen, *Ab initio protein structure prediction*, Curr. Opin. Struct. Biol., **12** (2002), 176-181.
[5] B. Carbonaro, F. Vitale e C. Giordano, *On a 3D-matrix representation of the tertiary structure of a protein*, Mathematical and Computer Modelling, **43** (2006), 1434-1464.
[6] F. Vitale, *On statistically meaningful geometric properties of digital 3D-structures of proteins*, Mathematical and Computer Modelling, **48** (2008), 141-160.
[7] F. Vitale, *A topology for the space of protein chains and a notion of local statistical stability for their three-dimensional structures.*, Mathematical and Computer Modelling, **48** (2008), 610-620.
[8] M. Golfarelli e S. Rizzi, *Data Warehouse. Teoria e pratica della progettazione*, McGraw-Hill (2002).

Dipartimento di Matematica, Seconda Università di Napoli
e-mail: federica.vitale@unina2.it
Dottorato in Biologia Computazionale
con sede presso la Seconda Università di Napoli – Ciclo XIX
Direttore di ricerca: Bruno Carbonaro, Seconda Università di Napoli