2 TECNICHE DI PREDIZIONE DELLE STRUTTURE PROTEICHE

La conoscenza della struttura tridimensionale delle proteine è un requisito fondamentale per la determinazione delle loro caratteristiche funzionali, oltre a costituire un dato indispensabile per lo studio dei meccanismi stessi di folding.

D'altra parte, se si tiene conto del numero di sequenze completamente determinate, che, al giorno d'oggi, è maggiore di almeno un ordine di grandezza del numero di strutture tridimensionali risolte sperimentalmente (Figura 2.1), è decisamente auspicabile l'approfondimento ed il raffinamento di tutti quei criteri propri della biologia molecolare teorica finalizzati alla predizione delle strutture superiori di proteine a partire dalla conoscenza della sola sequenza aminoacidica.

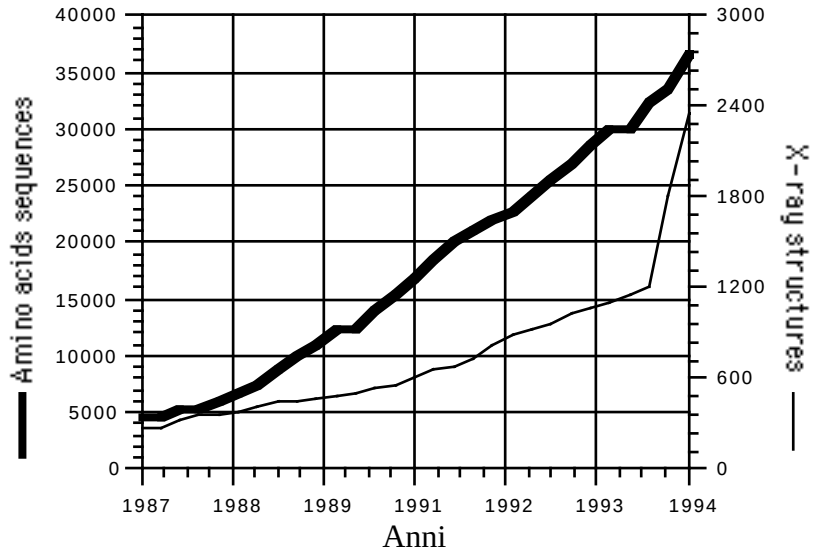

Figura 2.1
Differenza tra il numero di strutture e sequenze proteiche. Il grafico mostra (linea grossa) la accumulazione, negli anni, di sequenze aminoacidiche di proteine nella banca dati SWISS-PROT e quella di strutture terziarie di proteine (linea fine)nella Brookhaven Protein Databank (PDB), a partire dal 1987. La differenza tra strutture e sequenze diviene ancora più marcata se si eliminano i dati associati a strutture omologhe fortemente correlate : rimarrebero solo poche centinaia di strutture terziarie differenti. Da [Eisenhaber, Persson e Argos, 1995]

Le tecniche di predizione della struttura secondaria sono arrivate, nel corso degli anni, a offrire un discreto grado di affidabilità, per quanto i risultati migliori non arrivino al 70% di successi. Sono distinguibili, in base al tipo di approccio, due diverse tipologie : le tecniche che utilizzano *metodi statistici* e quelle che utilizzano *sistemi connessionisti*.

L'obiettivo di queste tecniche è l'assegnazione di elementi di struttura secondaria a segmenti più o meno estesi di sequenza aminoacidica, partendo dalla conoscenza di un

numero sufficientemente grande di strutture proteiche, utilizzate come campioni esemplari, delle quali siano note sia la sequenza che la configurazione tridimensionale. Con l'analisi di un vasto repertorio di strutture è possibile quantificare numericamente la *propensità* che determinati residui aminoacidici hanno a realizzare precise configurazioni secondarie (α−*helix*, β-*strand*, *coil*), permettendo così di utilizzare, nella predizione, delle osservabili statistiche (frequenze) direttamente ottenibili dalle basi di dati.

Alternativamente, sempre disponendo di un ricco database di strutture, è possibile *addestrare* una rete neuronale artificiale ad assegnare ad un pattern complesso (i.e., un segmento di sequenza aminoacidica) una delle tre possibili configurazioni secondarie fondamentali. Nei seguenti paragrafi saranno delineate le caratteristiche principali delle tecniche più utilizzate, nei due approcci, per la predizione della struttura secondaria delle proteine.

2.1 Metodi statistici

Una tra le tecniche di predizione su base statistica più utilizzate è quella elaborata da Chou e Fasman [Chou & Fasman, 1974], dove la propensità per un tipo di aminoacido (i.e. K) a trovarsi in una particolare struttura secondaria (i.e. α) è data dal rapporto tra la frazione di residui (K) che si trovano in quella struttura secondaria (α) e la frazione di residui (K) che si trovano in una qualsiasi delle tre strutture (α, β, coil). Il calcolo delle propensità, eseguito dagli autori citati su una base di dati di 15 strutture determinate cristallograficamente, fornisce una tabella nella quale ogni aminoacido viene classificato con un coefficiente che riflette la frequenza con la quale *forma*, *interrompe*, o è *indifferente* alla formazione di una α elica e, rispettivamente, di un foglietto β. La procedura per la quale avviene la predizione è basata sulla individuazione di serie di valori immediatamente successivi di alta propensità, per localizzare i siti di formazione di eliche o foglietti, estesi sulla sequenza fintanto che i valori di propensità media (calcolati su una finestra di 5 o 6 residui) rimangono al di sopra di una soglia prefissata. Su un set di 19 proteine, (15 delle quali utilizzate per il calcolo della tabella di propensità), sono state predette correttamente le strutture secondarie (nelle tre modalità) del 77% dei residui. Utilizzando un set più esteso di proteine, il metodo di Chou e Fasman ha permesso di ottenere solo il 50% dei risultati corretti [Kabsch & Sander, 1983]. Ciò evidenzia la dipendenza del risultato dall'estensione e dai criteri di scelta del set di strutture proteiche utilizzato per il calcolo delle propensità.

Un altro criterio statistico di predizione, elaborato da Garnier, Osguthorpe e Robson, e noto con il nome di GORIII [Garnier, 1978] associa all'informazione sulla propensità quella sulle interazioni statisticamente rilevanti tra le coppie di residui sulla catena polipeptidica. L'informazione I associata al j-esimo residuo R_j per la configurazione di struttura secondaria di tipo Z (elica, foglietto o coil) viene considerata nella sua dipendenza dal *contesto*

aminoacidico locale. Considerando una *finestra* di 8 residui prima e 8 dopo il j-esimo, si avrà :

$$I\left(S_j = Z;\, R_{j-8},\ldots,\, R_{j+8}\right) = \sum_{m=-8}^{8} I\left(S_j = Z;\, R_{j+m} | R_j\right)$$

dove

$$I\left(S_j = Z;\, R_{j+m} | R_j\right) = \log\left[\frac{P\left(Z|\left(R_{j+m}|R_j\right)\right)}{P(Z)}\right]$$

e P è la probabilità condizionale. Utilizzando un set di 68 diverse strutture cristallografiche di proteine, sono state ottenute predizioni di struttura secondaria per un valore pari al 63% di risultati corretti per proteine appartenenti al training-set.

I metodi di questo tipo, che predicono la struttura secondaria sulla base della analisi statistica della propensità non permettono di ottenere, abbiamo visto, risultati molto migliori del 60/70% di predizioni corrette.

Un'altra proposta interessante [Geourjon & Deléage, 1994] per la predizione di struttura secondaria consiste in un metodo auto-ottimizzato, sempre basato sull'allineamento multiplo. Esso prevede, prima di tutto, la selezione di un data-base limitato di proteine con struttura secondaria nota e omologhe con la proteina in esame. Quindi, vengono predette le strutture secondarie (note) delle proteine del data-base, aggiustando i parametri dell'algoritmo fino a che non si presentino più miglioramenti nel risultato. Infine, i parametri ottimizzati vengono utilizzati per la predizione della struttura secondaria della proteina in esame. Con un data-base di 239 proteine con struttura terziaria nota e con una percentuale di identità tra coppie di residui minore del 50%, si sono ottenuti dei risultati aventi una accuratezza media di predizione della struttura secondaria (sui tre stati) del 69%.

L'utilizzo dell'allineamento multiplo congiuntamente agli algoritmi di predizione su singola sequenza coincide, sostanzialmente, con una operazione di classificazione delle proteine utile per aumentare la quantità di informazione correlata con la sequenza in esame, per incrementare l'omogeneità della base di dati e, quindi, per migliorare la specificità della predizione.

2.2 Metodi connessionistici

I sistemi connessionisti, noti anche come reti neurali artificiali, sono sostanzialmente degli algoritmi in grado di associare a input complessi (pattern) degli output semplici (classi). L'architettura rispecchia, in prima approssimazione, quella del sistema nervoso centrale : una serie di elementi interconnessi (da cui il termine "connessionista"), ciascuno dei quali ha molti ingressi, mediati da pesi variabili (*sinapsi*), ed una sola uscita (*assone*), il cui valore dipende in modo generalmente non lineare dagli ingressi pesati. I valori dei pesi, che determinano la risposta del sistema ad un determinato ingresso, dipendono in modo adattivo

(si utilizza frequentemente la parola *apprendimento*) dagli *esempi* che vengono utilizzati per la costruzione delle associazioni ingresso/uscita (o pattern/classe).

Le differenti tipologie dei sistemi connessionisti sono raggruppabili in due grandi famiglie: sistemi ad apprendimento supervisionato e sistemi ad apprendimento non supervisionato. Nei primi l'associazione pattern/classe viene imposta sulla base della conoscenza di un sufficiente numero di coppie note *a priori* . Nei secondi l'individuazione delle classi viene conseguita esclusivamente in base all'informazione contenuta intrinsecamente negli *esempi* .

L'utilizzo dei sistemi connessionisti nel campo della predizione della struttura secondaria delle proteine, relativamente recente, consiste nella determinazione delle relazioni complesse che associano un frammento di sequenza (pattern) a una struttura secondaria (classe). I risultati che si ottengono con l'applicazione di questo metodo vengono talvolta riportati in forma tabulare, analoga alle tavole dei valori di propensità degli aminoacidi per le varie conformazioni prodotte da alcuni dei metodi tradizionali illustrati precedentemente. Ciò mette in grado chiunque di utilizzare tali risultati ai fini della predizione della struttura secondaria di un qualunque polipeptide.

Una soluzione che ha dato risultati interessanti è quella proposta da Qian e Sejnowskij [Qian & Sejnowskij, 1988] che prevede l'utilizzo di un *percettrone multistrato* (Fig. 2.2), nel quale viene implementato l'algoritmo di apprendimento supervisionato di *back-propagation* [Rumelhart , 1986]. La rete standard usata da questi autori comprende: uno strato di output corrispondente ad una finestra di 13 residui, ognuno dei quali codificato da 20 neuroni; uno strato di output formato da 3 unità, ciascuna rappresentante una possibile conformazione secondaria da assegnare al residuo centrale della finestra; uno strato intermedio formato da 40 unità nascoste. Nello riquadro seguente è schematizzato l'algoritmo di apprendimento supervisionato di un percettrone per ottenere la predizione della struttura secondaria di proteine.

Implementazione dell'algoritmo di apprendimento supervisionato per un percettrone	
Progettazione	Si definisce la topologia della rete, ovvero si stabilisce: **A**) il numero degli strati, la loro dimensione e composizione in gruppi; **B**) la geometria e l'intensità delle connessioni (sinapsi) fra neuroni ed il valore di soglia per ciascun neurone.
Apprendimento	Si sottopongono alla rete un certo numero di casi significativi, per i quali si conosce l'esatta corrispondenza, nel caso in esame, fra struttura primaria e secondaria di una certa sequenza polipeptidica, ottimizzando in base a questa corrispondenza la geometria e l'intensità delle connessioni e il valore di soglia per i neuroni.
Interrogazione	Si sottopone allo strato di input della rete una struttura primaria, lasciando che, in base ai valori relativi alle connessioni e alle soglie, ottimizzati nella precedente fase di "Apprendimento", lo strato di output proponga una struttura secondaria corrispondente.

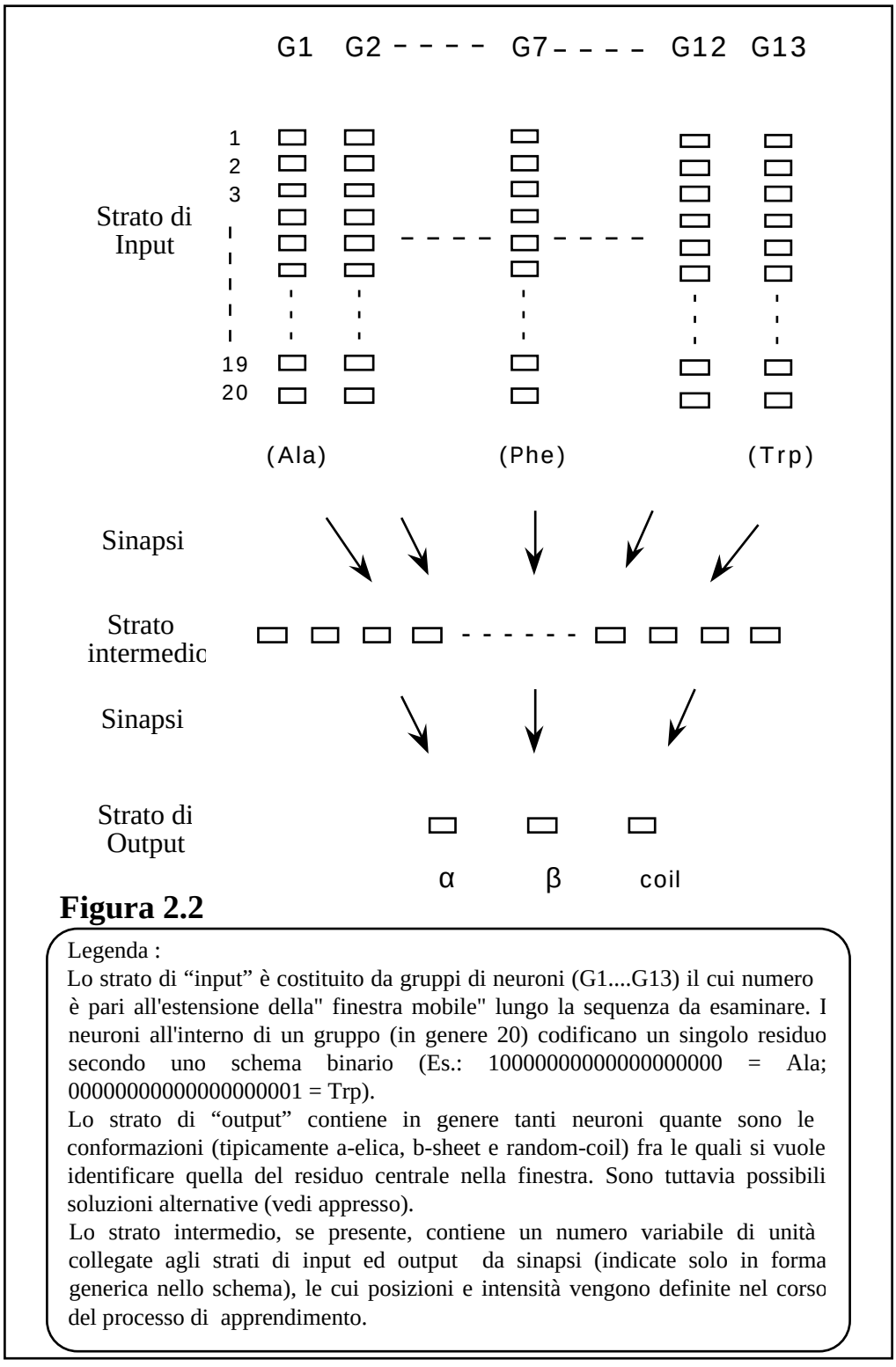

Figura 2.2

Legenda :
Lo strato di “input” è costituito da gruppi di neuroni (G1....G13) il cui numero è pari all'estensione della" finestra mobile" lungo la sequenza da esaminare. I neuroni all'interno di un gruppo (in genere 20) codificano un singolo residuo secondo uno schema binario (Es.: 10000000000000000000 = Ala; 00000000000000000001 = Trp).
Lo strato di “output” contiene in genere tanti neuroni quante sono le conformazioni (tipicamente a-elica, b-sheet e random-coil) fra le quali si vuole identificare quella del residuo centrale nella finestra. Sono tuttavia possibili soluzioni alternative (vedi appresso).
Lo strato intermedio, se presente, contiene un numero variabile di unità collegate agli strati di input ed output da sinapsi (indicate solo in forma generica nello schema), le cui posizioni e intensità vengono definite nel corso del processo di apprendimento.

Un piccolo ma significativo miglioramento nelle prestazioni è stato notato da Qian e Sejnowskij utilizzando due reti in serie, in modo che, a parità di tutte le altre condizioni, l'output della prima divenga l'input della seconda. Quest'ultimo risulta in tal modo costituito

da 13 gruppi con tre unità per gruppo, e contiene tutta l'informazione relativa alla struttura secondaria derivante dalla prima rete.

Quella appena descritta non è l'unica architettura di reti a percettroni che sia stata proposta per la predizione della struttura secondaria delle proteine.

Altre soluzioni interessanti sono state proposte da Holley e Karplus [Holley & Karplus, 1989] e da Bohr [Bohr , 1990]

Nel primo caso lo strato di input consiste in una finestra di 17 gruppi. Ogni gruppo è costituito da 21 neuroni, uno per ciascuno dei 20 aminoacidi, più uno usato quando la finestra mobile si sovrappone con l'estremità della catena polipeptidica. Lo strato intermedio contiene due soli neuroni. Anche lo strato di output contiene due soli neuroni, che codificano le strutture secondarie secondo lo schema : (1,0) = α ; (0,1) = β ; (0,0) = coil. I reali positivi compresi fra 0 ed 1 che costituiscono i valori effettivamente assunti dalle unità di output, vengono discretizzati in 0 o 1 con l'uso di un valore di soglia, anch'esso ottimizzato durante il processo di apprendimento. In definitiva, l' α-elica è assegnata a quei gruppi di almeno quattro residui contingui che abbiano valori della prima unità di output maggiori sia della seconda, sia della soglia; il β-sheet è assegnato ai gruppi di almeno due residui contigui per i quali i valori della seconda unità di output siano maggiori sia della prima sia della soglia; il random-coil é assegnato a tutti i rimanenti valori.

Nel secondo caso, le particolarità più significative consistono : **a**) nell'avere *finestre* molto ampie, comprendenti 25 residui per lato; **b**) lo strato di output è composto da 2 unità codificanti (il livello di confidenza per) la presenza o l'assenza di una singola configurazione secondaria. Ciò significa che ottenere il quadro completo della struttura secondaria di una proteina comporta l'uso di tante reti, ciascuna specializzata per una particolare configurazione. Per il resto, le reti usate da questi autori sono molto simili a quelle usate da Qian e Sejnowskij.

Un'osservazione di grande importanza anche pratica è che le prestazioni di una rete non dipendono in modo semplice dalle dimensioni del "training set" usato nella fase di apprendimento. In particolare:

- grande importanza riveste il grado di omologia esistente fra le proteine del “training set” e quelle del “testing set” (usate nella fase di *interrogazione*);
- tanto meglio la rete “impara a riconoscere” le proteine del training set, tanto peggiore sarà la sua abilità predittiva nei confronti di proteine “non note”.

Il paragone fra reti neurali e metodi statistici tradizionali per la previsione della struttura secondaria di proteine viene effettuato utilizzando degli indici di affidabilità, alcuni dei quali sono riportati nella finestra seguente:

Indici di affidabilità usati nel confronto di metodi predittivi

Il paragone fra reti neurali e metodi statistici tradizionali per la previsione della struttura secondaria di proteine viene effettuato utilizzando i seguenti indici:

$$\mathbf{Q_3} = \text{(percentuale di predizione corretta)} = \frac{P_\alpha + P_\beta + P_{coil}}{N}$$

in cui P_i = residui previsti correttamente nella configurazione i-esima;
N = numero totale di residui.
$\mathbf{C_i}$ = (coefficiente di correlazione relativo alla configurazione i-esima) =

$$= \frac{P_i n_i - u_i o_i}{\sqrt{(n_i + u_i)(n_i + o_i)(P_i + u_i)(P_i + o_i)}}$$

in cui : *i* può essere una qualunque configurazione (α, β, coil, ...) ;
P_i = numero di residui previsti correttamente in configurazione i-esima;
n_i = " " non previsti " " " ;
o_i = " " previsti non correttamente " " ;
u_i = " " non previsti non correttamente " " .

Nella tabella seguente vengono riportati i risultati di due analisi di questo tipo apparse in letteratura che, entrambe, indicano prestazioni significativamente migliori nel caso dei percettroni.

Accuratezza nella predizione - **Q3** (%) - (Cα, Cβ, Ccoil)					
	Chou - Fasman	Robson	Lim	NN (1)	NN (2)
Qian & Sejnowskij, 1988 (#)	50.00 (.25;.19;.24)	53.00 (.31;.24;.24)	50.00 (.35;.12;.20)	62.70 (.35;.29;.28)	64.30 (.41;.31;.41)
Holley & Karplus, 1989 ($)	48.00	55.00	54.00	63.00 (.41;.32;.36)	

Note:
NN (1) e NN (2) si riferiscono rispettivamente a una e due reti neuronali (in serie). In tutti i casi l'assegnazione delle strutture secondarie è basata sull'algoritmo di Kabsch and Sander (1983).
(#) "training" = 18105 residui / 91 proteine ; "test" = 2441 residui / 15 proteine
($) "training" = 8315 residui / 48 proteine ; "test" = 2441 residui / 14 proteine

E' ovvio che l'importanza di predire correttamente la struttura tridimensionale delle macromolecole aumenta enormemente laddove le possibilità di uno studio spettroscopico diretto sono scarse o nulle, come, per esempio, per la maggior parte delle proteine integrali di membrana.

Come accennato precedentemente, risulta cruciale la scelta del set di proteine da utilizzare per l'addestramento della rete neuronale: i risultati di predizione migliori si ottengono quando la proteina incognita presenta delle omologie con quelle utilizzate per l'addestramento.

I risultati ottenuti, riportati nella tabella seguente, mostrano che in uno dei tre casi la predizione é in accordo soddisfacente con le stime ottenute dagli spettri di dicroismo circolare delle percentuali di α-elica e foglietto β esistenti nell'enzima completamente ridotto.

Influenza della composizione del 'training set' sulla predizione della struttura secondaria della nitrito reduttasi di *Pseudomonas* da parte di un percettrone a due strati.

	TRAINING SET 1 ($)	TRAINING SET 2 (§)	TRAINING SET 3 (&)	dicroismo circolare (*) (100% riduzione)
HELIX (%)	0.07	0.23	0.20	0.16 ± 0.01
BETA (%)	0.42	0.37	0.5O	0.48 ± 0.02
COIL (%)	0.50	0.41	0.29	n. d.

*) Da Tordi et al. (1984)
($) training set = Bence-Jones protein + SOD (from erythrocytes)
(§) training set = b-trypsin + ferredoxin
(&) training set = subtilisin inhibitor + plastocyanin

Il miglioramento delle prestazioni del metodo è quindi correlato con la maggiore omogeneità del *training set* con la proteina in esame (di cui si vuol predire la struttura secondaria, conoscendone solo la primaria). Ciò richiede quindi l'individuazione di un criterio di **classificazione**, che sia in grado di evidenziare le similitudini tra strutture proteiche in base alla loro sequenza aminoacidica, e, quanto più possibile, in base alla struttura tridimensionale.

2.3 Calcoli *ab initio*

L'approccio della fisica al problema del *folding* delle proteine è basato sul presupposto che la struttura nativa di una proteina corrisponde ad un sistema in equilibrio termodinamico ad un minimo di energia libera. L'osservazione di Anfinsen [Anfinsen, 1973] che tutta l'informazione necessaria per il *folding* di proteine è contenuta nella struttura primaria costituisce una solida validazione sperimentale del presupposto appena citato.

Lo stato nativo di una struttura proteica è quindi rappresentato da una unica conformazione, con la minima somma di energia potenziale intramolecolare, entropia conformazionale e energia libera del solvente.

Recentemente si è fatta avanti l'ipotesi che proteine, come ad esempio le *chaperonine* molecolari, agiscano in modo simile ad enzimi e, riducendo le barriere energetiche di transizioni conformazionali o preservando dall'aggregazione fasi intermedie di *folding*, siano in grado di controllare, come fattori esterni, le cinetiche stesse del *folding* . Tale ipotesi, confermata da evidenze sperimentali [Gething & Sambrook, 1992] [Hartl, 1994], al momento, ancora non contraddice quella classica per la quale lo stato nativo è un minimo di energia del sistema proteina-solvente.

Il calcolo del minimo di energia libera per un sistema complesso come una proteina è di difficoltà spesso insormontabile, tanto che è possibile dimostrare [Ngo & Marks, 1992 ;

Unger & Moult, 1994] che è un problema definito **NP**-completo, e cioé un problema di decisione la cui complessità non è polinomiale (**P**), o meglio non è decidibile in tempo polinomiale. Dimostrare che un problema è NP-completo equivale a togliere la speranza di risolverlo in modo efficiente [Garey & Johnson, 1979]. In questa ottica va intesa la lettura del *paradosso di Levinthal* che ritiene insufficiente il tempo di vita dell'universo affinché una proteina, nella ricerca della sua conformazione più stabile, *esplori* tutti i possibili minimi locali di energia libera nello spazio delle configurazioni. A spiegazione (e risoluzione) del paradosso va considerata l'ipotesi (supportata recentemente anche da evidenze sperimentali [Baker et al., 1992]) formulata dallo stesso Levinthal [Levinthal, 1968] che le proteine si muovano in una piccola frazione dello spazio conformazionale e si spostino in direzione del minimo locale di energia libera più accessibile.

Lo spostamento delle strutture native verso il minimo globale di energia libera avviene, in accordo con l'interpretazione evoluzionistica, grazie all'effetto congiunto di mutazioni casuali e della stessa selezione operata dall'evoluzione. Il paradosso di Levinthal può essere evitato ipotizzando la presenza di *percorsi guidati* nel processo di *folding* ("the folding tunnel") [Gulukota & Wolynes, 1994].

I metodi di calcolo della struttura tridimensionale *ab initio*, che si avvalgono cioé della ricerca dei minimi di energia conformazionale, costituiscono quindi una realistica soluzione al problema della predizione della struttura tridimensionale delle proteine, a patto che si soddisfino due requisiti fondamentali : **a)** la determinazione di una funzione "energia" che permetta di discriminare la conformazione nativa dalle altre, ad energie superiori, e **b)** un criterio efficiente ed affidabile di ricerca dei minimi energetici nello spazio delle conformazioni.

Nella valutazione dei contributi energetici coinvolti nel calcolo dell'energia conformazionale si possono individuare due gruppi fondamentali : **a)** quello relativo alle interazioni intramolecolari (*in vacuo*), sia nei termini covalenti (legami chimici, modificazioni meccaniche nella struttura chimica, etc.) che in quelli a valenza nulla (interazioni di van der Waals, legami idrogeno, interazioni coulombiane, entropia conformazionale), e **b)** quello relativo, invece, all'energia dell'interazione con il solvente, e in particolare ai termini a lungo raggio di volume (polarizzazione del mezzo), e a quelli di superficie (formazione di cavità, interazioni soluto-solvente, variazioni di struttura del solvente), a corto raggio.

La quantità di termini importanti nella determinazione dell'energia conformazionale e la necessità di calcolarla per un numero di volte pari al numero di tutti i possibili (e ammissibili) arrangiamenti, nello spazio, della catena polipeptidica rendono cruciale la messa a punto di procedure di ottimizzazione delle procedure e di riduzione dei tempi di calcolo (parallelizzazione, etc.), altrimenti inaccettabili.

La ricerca del minimo di energia conformazionale (una volta determinata la funzione energia) può avvenire con modalità differenti :

- Determinazione di tutte le possibili conformazioni.
- Utilizzo di tecniche di ricostruzione della struttura tridimensionale tramite l'aggiunta di un residuo (o di frammenti oligopeptidici) alla volta nella catena polipeptidica e il calcolo del minimo energetico per ogni passo.
- Metodi deterministici di ottimizzazione globale.
- Ricerca del minimo con procedure stocastiche (Monte Carlo).
- Integrazione delle equazioni newtoniane del moto (approccio della dinamica molecolare).

Un approccio interessante alla determinazione computazionale della struttura tridimensionale di una catena polipeptidica è quello basato sul principio di Boltzmann e sulla teoria di campo medio [Sippl, 1993] : l'obiettivo è, in questo caso, l'estrazione di una funzione energia (e di un campo di forze quale derivata dell'energia rispetto alle variabili conformazionali) a partire dall'analisi di un data-base di strutture tridimensionali note. L'interesse dell'algoritmo sta proprio nella fusione di metodi di *calcolo a priori* della struttura a partire dalle interazioni fisiche elementari tra gli atomi costituenti le proteine e l'*analisi dei dati* sperimentali relativi alle strutture proteiche già risolte in laboratorio.

L'idea di base di questo criterio parte dalla considerazione del principio di Boltzmann che descrive la distribuzione di molecole tra microstati aventi diverse energie tramite la funzione densità di probabilità p . In variabili discrete la **legge di Boltzmann** può essere scritta come

$$p_{ijk} = Z^{-1} \exp\left(-\frac{E_{ijk}}{kT}\right)$$

dove k e T sono la costante di Boltzmann e la temperatura assoluta, e gli indici i,j,k corrispondono alle diverse variabili del sistema. La quantità Z coincide con la funzione di partizione

$$Z = \sum_{ijk} \exp\left(-\frac{E_{ijk}}{kT}\right).$$

Diversamente da quanto si fa in meccanica statistica, dove si calcolano la funzione di partizione Z e la probabilità di uno stato con energia E_{ijk} nota, in questo approccio viene utilizzata l'abbondanza di dati sperimentali per ricavarsi empiricamente la distribuzione di

probabilità (e la funzione di partizione), ricavando così una **legge di Boltzmann inversa** nella quale l'energia di un sistema viene ad essere funzione della probabilità (o meglio, della *frequenza* con la quale si osserva uno stato) e della funzione di partizione Z :

$$E_{ijk} = -kT \ln(f_{ijk}) - kT \ln(Z)$$

dove E_{ijk} è detta *potenziale di forza media* . Essendo impossibile derivare la funzione di partizione Z dai dati sperimentali ed essendo l'energia definita a meno di una costante, si può porre Z=1 ottenendo così

$$E_{ijk} = -kT \ln(f_{ijk}) .$$

In tal modo, il principio di Boltzmann inverso ci permette di considerare l'effetto di tutti i molteplici contributi al calcolo di una funzione energia in una unica soluzione, a patto che si abbia un numero sufficiente di dati sperimentali relativi al sistema studiato e che si sia in grado di stabilire un appropriato sistema di riferimento nel quale inquadrare l'energia calcolata con l'applicazione della legge di Boltzmann inversa.