1 INTRODUZIONE E PRESENTAZIONE DEL LAVORO

1.1 Introduzione

Le tecniche di cristallografia a raggi X per la determinazione sperimentale della struttura tridimensionale delle macromolecole proteiche offrono al giorno d'oggi un potente strumento di lavoro, non essendo però esenti da problemi e limiti di utilizzo, specialmente per quel che riguarda la produzione della proteina cristallizzata.

La possibilità di predire la struttura tridimensionale di una proteina conoscendone la sequenza di aminoacidi che la compongono è oggetto di ricerca da diversi decenni, esattamente da quando Anfinsen [Anfinsen, 1973] dimostrò che nella struttura primaria è contenuta tutta l'informazione sufficiente per determinare univocamente il folding della ribonucleasi bovina. Una soluzione generale e soddisfacente del problema, tuttavia, è ancora lontana, e si sono delineati, nel corso degli anni e in seguito al progredire di tecnologie sperimentali e strumenti matematici, differenti approcci basati su modelli conoscitivi sostanzialmente differenti. Si possono individuare tre criteri base costituenti i paradigmi dei metodi predittivi :

1) Calcolo *ab initio* della conformazione di minima energia. (§ 2.2.3) In pratica si tenta di ricavare, dall'informazione fisico chimica associata alla struttura primaria e dall'applicazione diretta delle leggi della fisica, la conformazione energeticamente più stabile del sistema di aminoacidi studiato (fissate le necessarie condizioni al contorno). Risulta evidente che il livello delle difficoltà teoriche e computazionali connesse a questo approccio è assai notevole, e che il suo presupposto essenziale che la configurazione presente in natura sia quella ad energia minima, non può essere garantito.

2) Elaborazione con metodi statistici dei dati sperimentali relativi a strutture note (§ 2.2.1) La frequenza con cui, in una finestra di lunghezza definita all'interno della catena polipeptidica, gli aminoacidi assumono una certa struttura tridimensionale costituisce un dato statistico che, se il campione di proteine utilizzate è sufficientemente grande, è utilizzabile per effettuare predizioni di struttura secondaria a partire dalla sequenza aminoacidica. Uno dei limiti di questo approccio è che, per ridurre il costo computazionale, la dimensione della finestra nella quale viene studiata la *propensità* di un aminoacido per particolari configurazioni di struttura secondaria è limitata a poche decine di residui.

3) Uso di algoritmi connessionisti e di reti neurali, come strumenti di classificazione e di predizione. (§ 2.2.2) Tali algoritmi possono essere di tipo supervisionato (in cui, come nel metodo statistico, si utilizzano come *esempi* strutture note) o di tipo non supervisionato, dove l'unica informazione elaborata è quella sulla struttura primaria. Il primo è tipicamente dedicato a

individuare eventuali correlazioni fra struttura primaria e strutture superiori. Il secondo viene preferito quando si vuole ottenere una classificazione delle strutture primarie unicamente sulla base della loro similarità di sequenza.

I primi due approcci, storicamente più sviluppati, sono arrivati a livelli di prestazione difficilmente migliorabili (a meno di modifiche sostanziali). D'altra parte, l'utilizzo dei metodi connessionisti, di più recente introduzione, è in grado di evidenziare aspetti del problema precedentemente ignorati e di ottimizzare le tecniche di aggiornamento e di consultazione delle basi di dati. I risultati che si ottengono con l'utilizzo del percettrone (§ 2.2.2) come strumento supervisionato di predizione della struttura secondaria sono confrontabili con quelli raggiungibili con i metodi statistici (§ 2.2.1). L'osservazione che la maggiore o minore accuratezza della predizione dipende fortemente dalla scelta del training-set di proteine utilizzato per addestrare la rete è stata pienamente confermata utilizzando un percettrone a due strati. Ciò rende indispensabile, nella prospettiva di estendere la predizione a strutture superiori (supersecondaria, etc.), un criterio di classificazione delle strutture primarie, che permetta di costruire dei training-set di proteine a struttura nota omogenei con la proteina a struttura incognita.Per questo motivo, e per altri che verranno discussi in seguito, si è deciso di approfondire il ruolo di una codifica ottimizzata per i residui aminoacidici e di una classificazione oggettiva delle strutture primarie, preliminari alla applicazione di un algoritmo di predizione. A tal fine ci si è avvalsi intensivamente di procedure di simulazione numerica ed elaborazione automatica dei dati.

Nel paragrafo seguente sono delineate schematicamente le tappe principali che hanno segnato l'evoluzione di questo progetto. I diversi punti, seppur disposti in ordine logico, non rispettano necessariamente l'evoluzione cronologica dei problemi trattati, e in essi si fa riferimento ai capitoli della tesi nei quali tali argomenti sono approfonditi.

1.2 Presentazione del lavoro

A- Codifica numerica delle strutture primarie.

Poiché la classificazione di un set di strutture primarie di proteine dipende dallaparticolare scelta adottata nella codifica dei residui aminoacidici costituenti la struttura primaria, nel corso di questa ricerca sono stati utilizzati i seguenti criteri di codifica (§ 3.2.1) numerica delle sequenze aminoacidiche, per individuare quello più rispondente allo scopo :

a) composizione aminoacidica della sequenza : per ogni proteina si ottiene un vettore a 20 componenti. Tale codifica è però troppo povera per ottenere una classificazione delle proteine che in qualche modo rispetti le relazioni di tipo strutturale;

b) matrice delle frequenze dei dipeptidi ordinati : per ogni sequenza vengono calcolati le frequenze con le quali si presentano i 20x20 dipeptidi ordinati. Per ogni proteina, di qualsiasi

lunghezza, si ottiene un vettore a 400 componenti. Viene così rispettato il vincolo sulla lunghezza uniforme dei vettori. Si perde, però, in questo modo, l'informazione sull'ordine dei diversi aminoacidi nella sequenza : tale codifica è sostanzialmente composizionale ed anche in questo caso le relazioni di tipo strutturale non sono sufficientemente conservate nella classificazione;

c) caratteristiche fisico-chimiche degli aminoacidi : la struttura primaria viene codificata numericamente assegnando ad ogni aminoacido un valore corrispondente ad una grandezza fisico-chimica rilevante nel folding della proteina (idrofobicità, volume, resistenza alla torsione, polarità, ecc.). In questo modo l'informazione sulla posizione dei differenti residui sulla catena viene conservata. Dal punto di vista della quantità di informazione associata alla codifica, è ovvio che più grandezze vengono utilizzate nella descrizione più accurata sarà la classificazione. Per massimizzare l'informazione e ridurne la ridondanza è stata effettuata una Analisi delle Componenti Principali (§ 3.1.2) su un insieme di sette caratteristiche fisico-chimiche (§ 3.2.1) degli aminoacidi e sono state ricavate le prime due componenti principali che da sole contengono più dell' 87% dell'informazione contenuta nell'intero insieme. Tale tipo di codifica non rispetta, tuttavia, il vincolo della uguale dimensione dei vettori.

B- Equalizzazione dei vettori descrittivi delle strutture primarie (algoritmo *PBA*).

Per ovviare al problema della differente lunghezza dei profili numerici associati alle sequenze aminoacidiche e utilizzati dall'algoritmo di classificazione è stato sviluppato un algoritmo (*Procust's Bed Algorithm, PBA*) (§ 3.2.1) di *allungamento* (o *accorciamento*) del profilo che ne conservasse il più possibile le caratteristiche d'insieme. Tale algoritmo è stato utilizzato, fissato un training-set di proteine le cui lunghezze non fossero troppo differenti, per equalizzare i profili ad una medesima lunghezza intermedia.

Gli algoritmi di codifica ed analisi delle strutture primarie sono stati implementati in un package software, scritto in linguaggio FutureBasic© per Macintosh, di imminente presentazione sulle riviste specializzate ed immissione nelle banche dati di software *public domain* .

C-Implementazione di metodi connessionistici di classificazione delle strutture proteiche.

Per ottenere una classificazione oggettiva delle strutture primarie è stato utilizzato l'algoritmo connessionista non supervisionato a *mappe auto organizzanti (SOMA)* , ideato da T. Kohonen [Kohonen, 1984 ; 1988] (§ 3.1.1), che gode di alcune proprietà estremamente interessanti, fra cui :

a) dato un insieme di oggetti descritti da vettori di uguale dimensione è possibile ottenere una *mappa* a dimensione ridotta (generalmente bidimensionale) nella quale gli oggetti vengono rappresentati e classificati, conservando le relazioni di similitudine esistenti nella descrizione originaria.

b) è possibile arricchire progressivamente il training-set, senza dover per questo ripetere ad ogni stadio la classificazione.

L'algoritmo di classificazione *SOMA* è stato implementato in linguaggio C sia in versione sequenziale che parallela [Sirabella, 1992; 1994] e ha permesso di ottenere i risultati riportati nel Cap. 4 .

Un vincolo essenziale imposto dall'utilizzo di tale algoritmo è che tutti i vettori descriventi gli oggetti da classificare siano della stessa lunghezza : ciò rappresenta ovviamente una grossa limitazione volendo classificare le strutture primarie di proteine.

D- La validazione del metodo (algoritmo *OPA*).

Una volta individuata la codifica opportuna, ottimizzata in funzione del rapporto informazione/ridondanza e risolto con l'algoritmo *PBA* il problema delle differenti lunghezze delle sequenze costituenti il training-set da sottoporre all'algoritmo di classificazione, sono state ottenute delle mappe bidimensionali sulle quali sono situate le diverse proteine (§ 4.1, 4.2, 4.3). Ci si è posti a questo punto il problema della validazione dei risultati e della conferma della correttezza delle scelte effettuate per ottenerli. Dato che l'obiettivo è quello di ottenere una classificazione delle strutture primarie in accordo con le reciproche relazioni strutturali, è stato appositamente sviluppato un algoritmo (algoritmo a "buccia di cipolla", *Onion peel's algorithm -OPA*) che permette di descrivere la *forma* tridimensionale della proteina (o meglio, del *backbone*), indipendentemente dal numero di residui della catena polipeptidica (§ 3.2.2). Tale algoritmo è in grado di leggere le strutture tridimensionali di proteine contenute nei files PDB, di calcolarne il centro geometrico e di valutare la distribuzione radiale di aminoacidi intorno a questo. Facendo l'analisi delle componenti principali dei vettori descriventi la distribuzione radiale di aminoacidi relativi ad un set di proteine, è possibile ottenere una mappa bidimensionale (§ 4.3) correlata con le relazioni di tipo strutturale (similitudine di forma).