5. DISCUSSIONE

Una analisi dei risultati ottenuti con l'utilizzo di *SOMA* e con il riferimento esterno della classificazione strutturale nota, evidenzia l'insufficienza di una codifica basata sulle frequenze dipeptidiche (Fig. 4.2/A) : il motivo è probabilmente legato alla troppo piccola estensione della *finestra* di codifica della sequenza (solo due residui successivi, il dipeptide ordinato *xy*). Una informazione di questo tipo risulta essere quasi esclusivamente composizionale e di ben poco aiuto nella individuazione di regolarità estese sulla catena polipeptidica.

L'utilizzo, invece, di proprietà fisico-chimiche come descrittori degli aminoacidi componenti la proteina e la conservazione dell'informazione posizionale nella catena polipeptidica permettono di ottenere delle classificazioni più in accordo con quelle strutturali di riferimento, a maggior ragione se la quantità di informazione viene aumentata in modo ottimizzato con l'utilizzo delle componenti principali : l'accordo e tanto più buono quanto è maggiore la quantità di caratteristiche fisico-chimiche utilizzate nella codifica.

Il problema di classificare le strutture primarie potrebbe essere considerato, indipendentemente da quello di predire le strutture tridimensionali, come problema a sé. Se, da una parte, questo ne limita l'interesse generale, dall'altra rende più facile la validazione dei metodi adottati. Quest'ultima può, in questo caso, essere fatta con riferimento alla stima, facilmente ottenibile, di omologia di sequenza. E' questa la via seguita da Ferràn e Ferrara [Ferràn e Ferrara, 1991 ; 1992a ; 1992b], che propongono l'utilizzo del classificatore di Kohonen (basato sull'algoritmo *SOMA*) per la classificazione di strutture primarie di proteine codificate con le frequenze dei 400 possibili dipeptidi ordinati (§ 3.1.1 e § 3.2.1).

Considerata l'enorme importanza che potrebbe assumere la disponibilità di un metodo finalmente affidabile per la predizione delle strutture tridimensionali, è lecito chiedersi se una classificazione ottimizzata delle strutture primarie potrebbe facilitare il compito. La risposta è ovviamente positiva, e si giustifica in riferimento alle considerazioni fatte nell'introduzione, in base alle quali sia gli algoritmi predittivi statistici che quelli connessionisti offrono i migliori risultati quando le basi dati o - rispettivamente - i training sets sono "omogenei" alla struttura da predire.

Nasce qui il problema di una definizione, opportuna in questo contesto, di *omogeneità strutturale* : non può essere risolto in modo completamente oggettivo se non facendo riferimento alla informazione relativa alle sequenze o alle coordinate tridimensionali. Fra le due alternative citate è chiaramente la seconda quella di gran lunga preferibile.

La scarsità relativa di strutture proteiche tridimensionalmente risolte (si veda il grafico all'inizio del Cap. 2) dal punto di vista sperimentale rende tuttavia necessario, in molti casi, l'adozione di criteri ibridi, in cui trovano posto anche approcci semiempirici basati sulla grafica molecolare. Nel secondo esperimento, illustrato nel § 4.2, il test di affidabilità del nostro

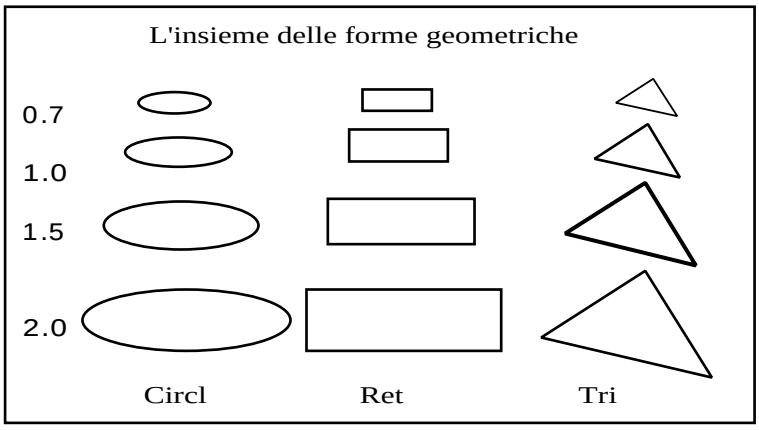
metodo ha dovuto necessariamente riferirsi ad uno standard ottenuto con un metodo del tipo suddetto [Pascarella & Argos, 1992]. Anche in questo caso i risultati si sono mostrati complessivamente soddisfacenti : la qualità della classificazione ottenuta, anche se inferiore (a parità di codifica) a quella relativa al primo esperimento (§ 4.1), va giudicata infatti in rapporto alla maggiore difficoltà oggettiva del problema. La bassa omologia di sequenza riscontrata fra gli elementi interni alle varie classi strutturali previste, considerata all'origine delle suddette difficoltà, se da una parte rende effettivamente poco utilizzabili gli standards di riferimento basati su percentuali di identità assolute (algoritmi alla Needlemann-Wunsch), dall'altra risulta trattabile - nonostante tutto - a patto di utilizzare opportune matrici di sostituzione nel calcolo della omologia di sequenza [Dayhoff et al., 1978].

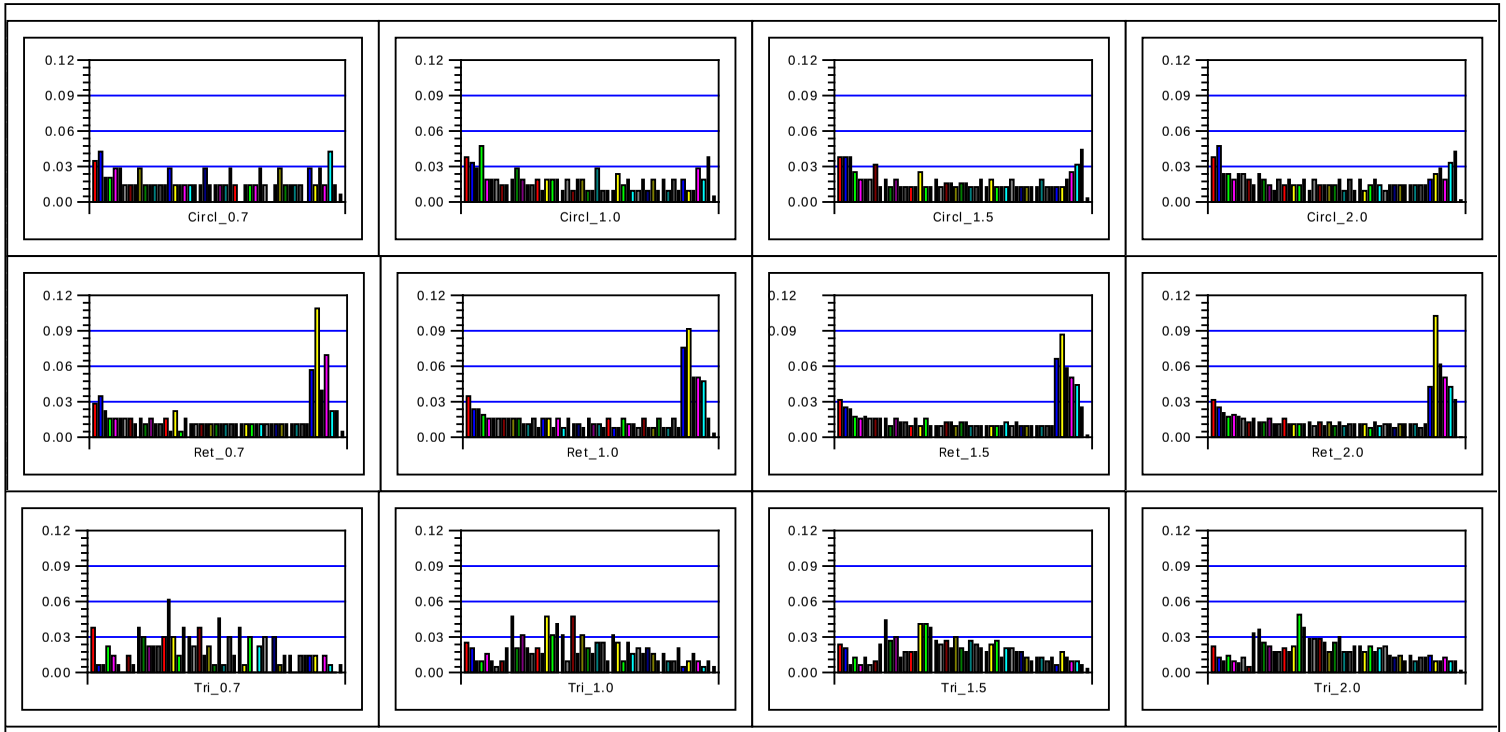
La performance del nostro metodo nei confrontio del problema suddetto, anche se soddisfacente (§ 4.2), è stata valutata utilizzando uno standard in cui concorrevano ugualmente informazioni qualitative e quantitative di carattere strutturale e funzionale.

Si è quindi reso necessario ai nostri occhi lo sviluppo di un metodo che risultasse completamente oggettivo come standard di riferimento, e l'averlo individuato e utilizzato attraverso l'algoritmo *OPA* (§ 3.2.2 e § 4.3) rende possibile una stima quantitativa e riproducibile dell'affidabilità dei risultati ottenuti.

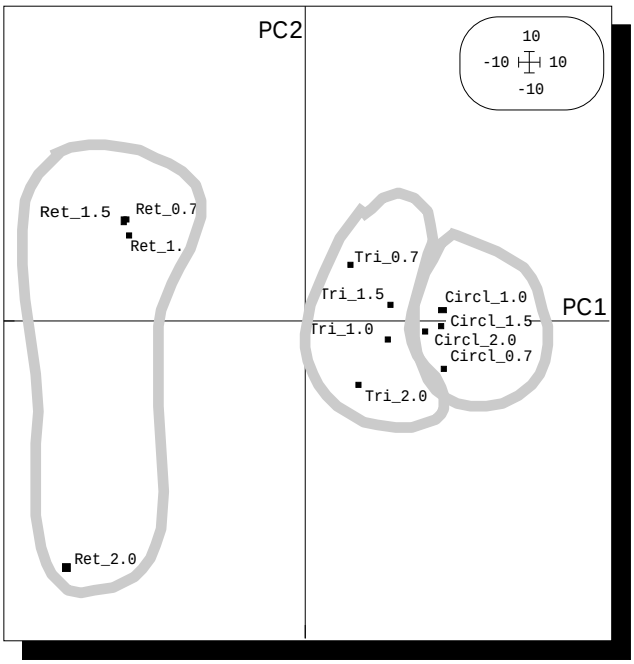
Prospettive di sviluppo

• Nell'analisi delle classificazioni di strutture terziarie prodotte con l'algoritmo *OPA* si rileva una sua prevedibile applicabilità anche in contesti più generali (i.e. pattern recognition di immagini bidimensionali, etc.). A tale fine si mostrano dei risultati, del tutto preliminari, di classificazione di semplici forme geometriche bidimensionali, ottenuti con l'algoritmo *OPA.*





Istogrammi relativi alla distribuzione radiale di punti intorno al centro per 12 figure geometriche bidimensionali.



		PC1(44%)	PC2(16%)
Circl_0.7	1	1.7238	-1.9286
Circl_1.0	2	1.3790	1.8229
Circl_1.5	3	2.7697	0.5594
Circl_2.0	4	2.6829	-0.4881
Ret_0.7	5	3.8448	-0.2536
Ret_1.0	6	4.5054	-1.4687
Ret_1.5	7	4.4224	-0.0742
Ret_2.0	8	4.4434	0.3769
Tri_0.7	9	-7.9436	-7.7535
Tri_1.0	10	-6.0796	3.1828
Tri_1.5	11	-5.9384	3.2765
Tri_2.0	12	-5.8098	2.7482

Figura 5.1 Classificazione di semplici forme geometriche bidimensionali
Nel grafico sono riportati i valori delle prime due componenti principali calcolate sull'insieme di 12 vettori a 50 componenti (rappresentati dagli istogrammmi riportati nella pagina precedente) contenenti la distribuzione radiale di punti per 12 figure geometriche bidimensionali. Dal grafico è evidente che una buona separazione delle forme avviene solo sull'asse della PC1. Il significato della componente PC2 è da spiegare.

Una ulteriore evoluzione dell'algoritmo, già in fase di realizzazione, sarà nel calcolo della distribuzione radiale di aminoacidi non più rispetto ad un solo centro (il baricentro del backbone) ma intorno ad un numero arbitrario di *punti principali* , e cioé i baricentri dei clusters di massima varianza (in analogia alla teoria delle componenti principali) [Flury, 1993]. Aumentando così il numero dei *punti di vista* del folding di una proteina sarà possibile realizzare una codifica ancora più efficiente.

• L'utilizzo dell'algoritmo *SOMA* di classificazione può essere sfruttato, insieme all'algoritmo *PBA* di equalizzazione dei profili, per la preparazione di training set omogenei con una sequenza X di cui si vuole predire la struttura secondaria. Per ottenere ciò è sufficiente classificare la sequenza X insieme a numerose altre di cui è nota la struttura tridimensionale. Le proteine note che, nella mappa finale prodotta da *SOMA*, saranno adiacenti alla posizione assegnata alla proteina X, costituiranno un set di proteine omogenee con quella X , e potranno essere utilizzate, a loro volta, come training set per un percettrone multistrato (§ 2.2). La predizione della struttura secondaria della proteina X ottenuta *addestrando* il percettrone con il training set omogeneo sarà, in questo modo, migliorata.

• In conclusione, vorremmo accennare al fatto che se, da una parte, l'algoritmo *PBA* ha risolto brillantemente il problema della necessaria equidimensionalità dei vettori da sottoporre all'algoritmo *SOMA* , dall'altra introduce inevitabilmente delle distorsioni nei profili numerici esprimenti le proprietà chimico-fisiche dei residui aminoacidici, i cui effetti sull'affidabilitàdelle classificazioni richiedono una ulteriore accurata investigazione. La Figura 5.2 riporta una serie di risultati che condensano gran parte del lavoro da noi svolto in questa direzione.

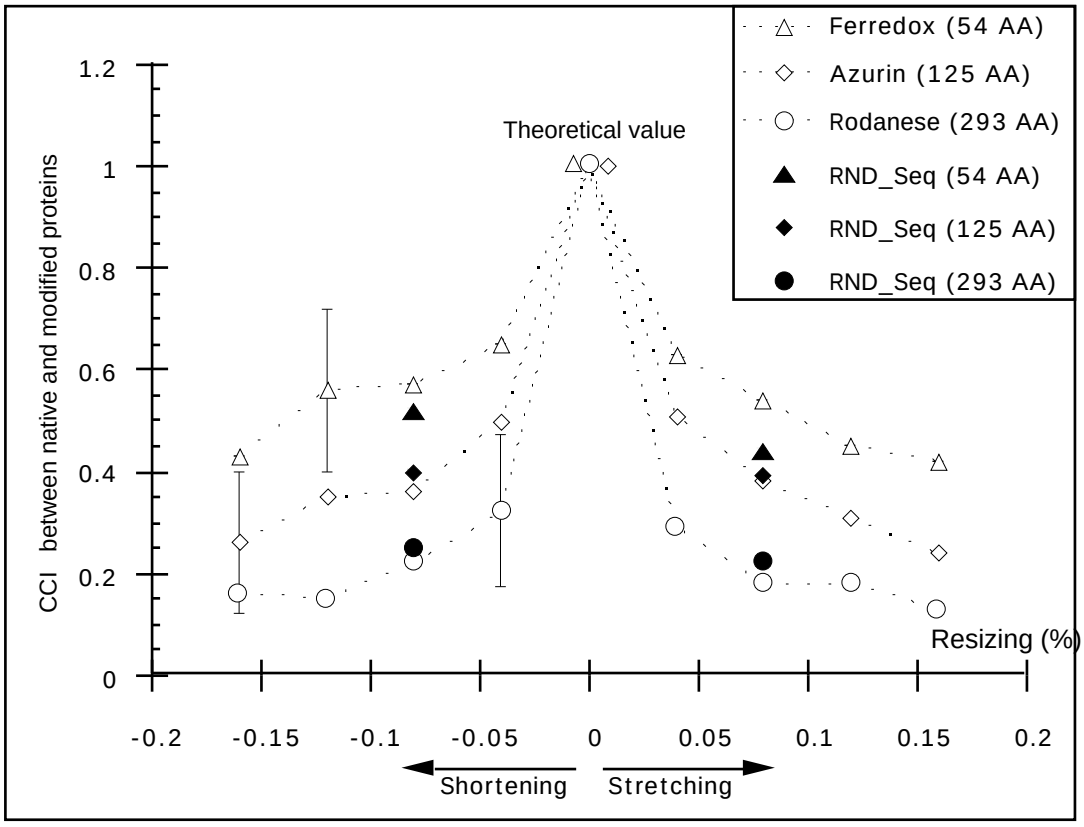


Figura 5.2 Alterazioni nei profili numerici introdotti dall'algoritmo *PBA*

Le alterazioni nei profili di idrofobicità introdotte dall'utilizzo dell'algoritmo *PBA* sono state stimate per mezzo dei valori di correlazione tra le strutture primarie native e modificate relative a tre proteine di lunghezza differente : Ferredoxina di *Peptostreptococcus asaccharolyticus* (Fepe, 54 AA), Azzurina di *Pseudomonas aeruginosa* (125 AA, from the precursor Azpsca, 128 AA) e Rhodanese bovina (Robo, 293 AA). Le strutture modificate sono in realtà la media di 50 modificazioni casuali, per ogni condizione, seguendo lo stesso criterio utilizzato per la Fig. 3.7 (§ 3.2.1). I simboli pieni si riferiscono ai risultati ottenuti per 3 sequenze generate casualmente delle stesse lunghezze. Le barre di errore indicano l'errore standard della media. Tali valori, seppur non identici, sono comunque molto simili nelle tre differenti situazioni. I dati in figura mostrano che la forma complessiva delle tre curve non cambia al variare della lunghezza del polipeptide, mentre con essa è fortemente correlata la pendenza con la quale decresce la correlazione (e quindi la conservazione della forma del profilo) all'aumentare della percentuale di resizing (in entrambe le direzioni).