4. RISULTATI

4.1 Esperimento 1 : Ottimizzazione dei criteri di codifica per gli aminoacidi.

Scopo dell'esperimento

Il primo *esperimento* è stato progettato per evidenziare le differenze, nelle classificazioni ottenute, dovute all'utilizzo di diverse codifiche della struttura primaria. L'obiettivo primario è quello di produrre una classificazione di sequenze aminoacidiche coerente con quella ricavabile dall'analisi delle rispettive conformazioni tridimensionali.

Modalità di esecuzione

E' stato scelto un gruppo di strutture primarie che rispettasse i seguenti vincoli :

a) devono essere già note e determinate le classi strutturali nelle quali è possibile raggruppare sottoinsiemi di proteine appartenenti al gruppo di test, in modo tale da verificare, a classificazione (di strutture primarie) avvenuta, il migliore o peggiore accordo tra classi strutturali note e classi di strutture primarie ottenute con l'algoritmo *SOMA* ;

b) le proteine devono avere lunghezze (numero di aminoacidi) non troppo differenti (fino a circa il 10-15% in più o in meno), in modo tale da garantire una sufficiente affidabilità dell'algoritmo *PBA* (vedi § 3.2.1) di equalizzazione dei profili numerici associati a ciascuna sequenza ;

c) tutti i gruppi devono essere sufficientemente rappresentati, anche se la presenza di *outlayers* non compromette a priori il funzionamento dell'algoritmo.

Considerate le premesse è stato selezionato un *training set* ridotto costituito da dieci strutture primarie di proteine, elencate nella Tabella 4.1 seguente.

I criteri seguiti per il raggruppamento delle proteine in classi (o famiglie) strutturali sono in parte quelli adottati da Dickerson [Dickerson, 1980] e in parte derivanti da informazioni sulla conformazione, tutti di natura empirica. Le proteine del *set* utilizzato sono state raggruppate in quattro classi strutturali : due citocromi c_8 (da *Rhodospirillum rubrum* e *Rhodospirillum photometricum*); due citocromi c_8 (low-spin) (da *Agrobacter tumefaciens Apple* e *Agrobacter tumefaciens B2a*) [Moore & Pettigrew, 1990] ; due citocromi c_2 (da *Rhodospirillum rubrum* e da *Rhodopseudomonas sphaeroides*) e uno c-550 (da *Paracoccus denitrificans*), tutti e tre di tipo **L**ong, secondo la classificazione strutturale di Dickerson; e tre globine (le catene alfa e beta della emoglobina di cavallo e la mioglobina di cavallo).

Allo scopo di verificare la classificazione con una procedura tradizionale di riferimento, le strutture primarie sono state analizzate con l'algoritmo di allineamento multiplo PILEUP della

suite GCG utilizzato con la matrice di Dayhoff [Dayhoff et al., 1978]. Dalla matrice di distanze prodotta é stato ricavato un dendrogramma riprodotto in Figura 4.1. Dal grafico si rilevano alcune incongruenze nelle due classificazioni da confrontare : quella ottenuta con l'analisi della struttura primaria lasciata nella sua codifica simbolica (con i nomi degli aminoacidi), e quella strutturale di riferimento (riportata nel grafico sotto forma di etichette in falsi colori). Una possibile spiegazione di questo risultato è nella insufficienza dell'informazione associata a una codifica esclusivamente simbolica, specialmente se l'obiettivo è quello di riprodurre una classificazione in accordo con le similitudini conformazionali.

Si è proceduto, quindi, al confronto tra le classificazioni ottenute con l'algoritmo *SOMA* e quattro differenti codifiche numeriche.

Le dieci sequenze selezionate hanno lunghezze comprese tra i **112** e i **153** residui, con una lunghezza intermedia pari a (112+153)/2 **133**, che costituirà la lunghezza comune di riferimento nell'applicazione dell'algoritmo *PBA* di equalizzazione dei profili.

Ciascuna delle sequenze è stata codificata in quattro modi differenti :

A) frequenze dipeptidiche : con un vettore a 400 componenti contenente i valori delle frequenze con le quali appaiono, nella sequenza, tutti i possibili dipeptidi *xy* ordinati ;

B) idrofobicità : con un vettore a 133 componenti ricavato dall'allungamento (o accorciamento) tramite l'algoritmo PBA fino alla lunghezza di riferimento del profilo di idrofobicità associato alla proteina e calcolato in base ai valori riportati in Tabella 3.1* ;

C) idrofobicità e volume : con un vettore a 266 componenti costituito da due vettori accodati di 133 componenti, relativi ai profili di idrofobicità e volume calcolati come in B).

D) componenti principali : con un vettore a 266 componenti ottenuto anch'esso con due vettori accodati di uguale lunghezza e relativi ai valori della prima e della seconda componente principale calcolate sull'insieme delle sette caratteristiche fisico-chimiche riportate in Tabella 3.1.

Nella Figura 4.2 sono riportati i profili di idrofobicità relativi alle dieci proteine del training set : con l'occasione si fa notare che i valori di idrofobicità (così come quelli delle altre sei proprietà fisico-chimiche) sono stati riscalati in modo tale da avere valore medio unitario. Ciò per fare in modo che, qualora si optasse per una codifica multipla (due o più profili accodati per ciascuna sequenza), tutte le proprietà fisico chimiche avessero peso equivalente, indipendente dall'unità di misura. I quattro gruppi di dieci vettori sono stati quindi utilizzati per ricavare una classificazione non supervisionata per mezzo dell'algoritmo *SOMA*, realizzato per una configurazione di 4X4 unità nello strato di uscita ** . Dopo diversi tentativi finalizzati alla

---

* In realtà, per motivi computazionali i valori delle proprietà fisico-chimiche riportati in Tab. 3.1 sono stati riscalati in modo tale da avere tutti valor medio unitario.

** La scelta del numero di unità dello strato di uscita, fatta in modo arbitrario, costituisce un limite superiore al numero di classi individuate, e per questa ragione viene fatta in base all'esperienza : tale numero sarà ovviamente superiore al numero di gruppi previsti.

ottimizzazione dei parametri variabili di regolazione dell'*apprendimento,* sono state ottenute quattro mappe finali di classificazione (Fig. 4.3), una per ogni tipo di codifica.

Tabella 4.1: **Training-set di proteine utilizzato nell' Esperimento 1.**

1.Cyt c8 from Rhodospirillum photometricum (125 AA) - NBRF:Ccqfcp /SW:CYCP_RHOPH
2.Cyt c8 from Rhodospirillum rubrum (126 AA) - NBRF:Ccqfcr/SW:CYCP_RHORU
3.Cyt c8 from Agrobacter tumefaciens Apple (125 AA) - NBRF:Ccaga6
4.Cyt c8 from Agrobacter tumefaciens B2A (122 AA) - NBRF:Ccagb6
5.Cyt c2 from Rhodospirillum rubrum (112 AA) - NBRF:Ccqf2r (prec.)
6.Cyt c2 from Rhodopseudomonas sphaeroides (124 AA) - NBRF:Ccrf2s (prec.)
7.Cyt c-550 from Paracoccus denitrificans (134 AA) - Cc550par [Dickerson, 1980]
8.Hb a-chain from Horse (141 AA) - NBRF:Haho
9.Hb b-chain from Horse (146 AA) - NBRF:Hbho
10.Mb from Horse (153 AA) - NBRF:Myho

Le strutture primarie

1.Ccqfcp (125 AA)
1234567890123456789012345678901234567890
ASPEAYVEYRKQALKASGDHMKALSAIVKG
QLPLNAEAAKHAEAIAAIMESLPAAFPEGT
AGIAKTEAKAVVWSKADEFKADAVKSADAA
KALAQAATAGDTAQMGKALAALGGTCKGCH
ETFRE

2.Ccqfcr(126 AA)
1234567890123456789012345678901234567890
ADPAAYVEYRKSVLSATSNYMKAIGITLKE
DLAVPNQTADHAKAIASIMETLPAAFPEGT
AGIAKTEAKAAIWKDFEAFKVASKKSQDAA
LELASAAETGDKAAIGAKLQALGGTCKACH
KEFKAD

3.Ccaga6 (125 AA)
1234567890123456789012345678901234567890
ADGGTHDARIALMKKIGGATGALGAIAKGE
KPYDAEIVKASLTTIAETAKAFPDQFNPKD
STDAEVNPKIWDNLDDFKAKAAKLSTDAET
ALAQLPADQAGVGNTLKTLGGNCGACHQAY
RIKKD

4.Ccagb6 (122 AA)
1234567890123456789012345678901234567890
AGEVEKREGMMKQIGGAMGSLAAISKGEKP
FDADTVKAAVTTIGTNAKAFPEQFPAGTET
GSAAAPAIWENFEDFKAKAAKLGTDADIVL
ANLPGDQAGVATAMKTLGADCGTCHQTYRL
KK

5.Ccqf2r (112 AA)
1234567890123456789012345678901234567890
EGDAAAGEKVSKKCLACHTFDQGGANKVGP
NLFGVFENTAAHKDNYAYSESYTEMKAKGL
TWTEANLAAYVKDPKAFVLEKSGDPKAKSK
MTFKLTKDDEIENVIAYLKTLK

6.Ccrf2s (124 AA)
1234567890123456789012345678901234567890
QEGDPEAGAKAFNQCQTCHVIVDDSGTTIA
GRNAKTGPNLYGVVGRTAGTQADFKGYGEG
MKEAGAKGLAWDEEHFVQYVQDPTKFLKEY
TGDAKAKGKMTFKLKKEADAHNIWAYLQQV
AVRP

7.Cc550par (134 AA)
1234567890123456789012345678901234567890
NEGDAAKGEKEFNKCKACHMIQAPDGTDIK
GGKTGPNLYGVVGRKIASEEGFKYGEGILE
VAEKNPDLTWTEANLIEYVTDPKPLVKKMT
DDKGAKTKMTFKMGKNQADDVAFLAQDDPD
AGEGEAAGAGSDSE

8.Haho (141 AA)
1234567890123456789012345678901234567890
VLSAADKTNVKAAWSKVGGHAGEYGAEALE
RMFLGFPTTKTYFPHFDLSHGSAQVKAHGK
KVGDALTLAVGHLDDLPGALSDLSNLHAHK
LRVDPVNFKLLSHCLLSTLAVHLPNDFTPA
VHASLDKFLSSVSTVLTSKYR

9.Hbho (146 AA)
1234567890123456789012345678901234567890
VQLSGEEKAAVLALWDKVNEEEVGGEALGR
LLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
KAHGKKVLHSFGEGVHHLDNLKGTFAALSE
LHCDKLHVDPENFRLLGNVLVVVLARHFGK
DFTPELQASYQKVVAGVANALAHKYH

10.Myho (153 AA)
1234567890123456789012345678901234567890
GLSDGEWQQVLNVWGKVEADIAGHGQEVLI
RLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGTVVLTALGGILKKKGHHEAELKPLA
QSHATKHKIPIKYLEFISDAIIHVLHSKHP
GNFGADAQGAMTKALELFRNDIAAKYKELG
FQG

Il codice a falsi colori associato alle proteine indica il loro raggruppamento sulla base di complessive somiglianze strutturali. Tutte le sequenze (eccetto la n. 7) sono state prelevate dalla banca dati NBRF e sono indicate con il loro nome logico.

Le mappe prodotte mostrano, sullo strato di uscita, le unità selezionate (a convergenza avvenuta) da ciascuna proteina. Per avere una prima indicazione delle similitudini di sequenza esistenti tra le proteine del training set è stata calcolata, con l'applicazione ripetuta dell'algoritmo di Needlemann e Wunsch [Needlemann & Wunsch, 1970] (utilizzando, per le *Gap penalties* , i valori di default della versione implementata nella suite GCG), una matrice (Tab. 4.2) contenente le percentuali di similarità per ciascuna coppia di proteine dell'insieme.

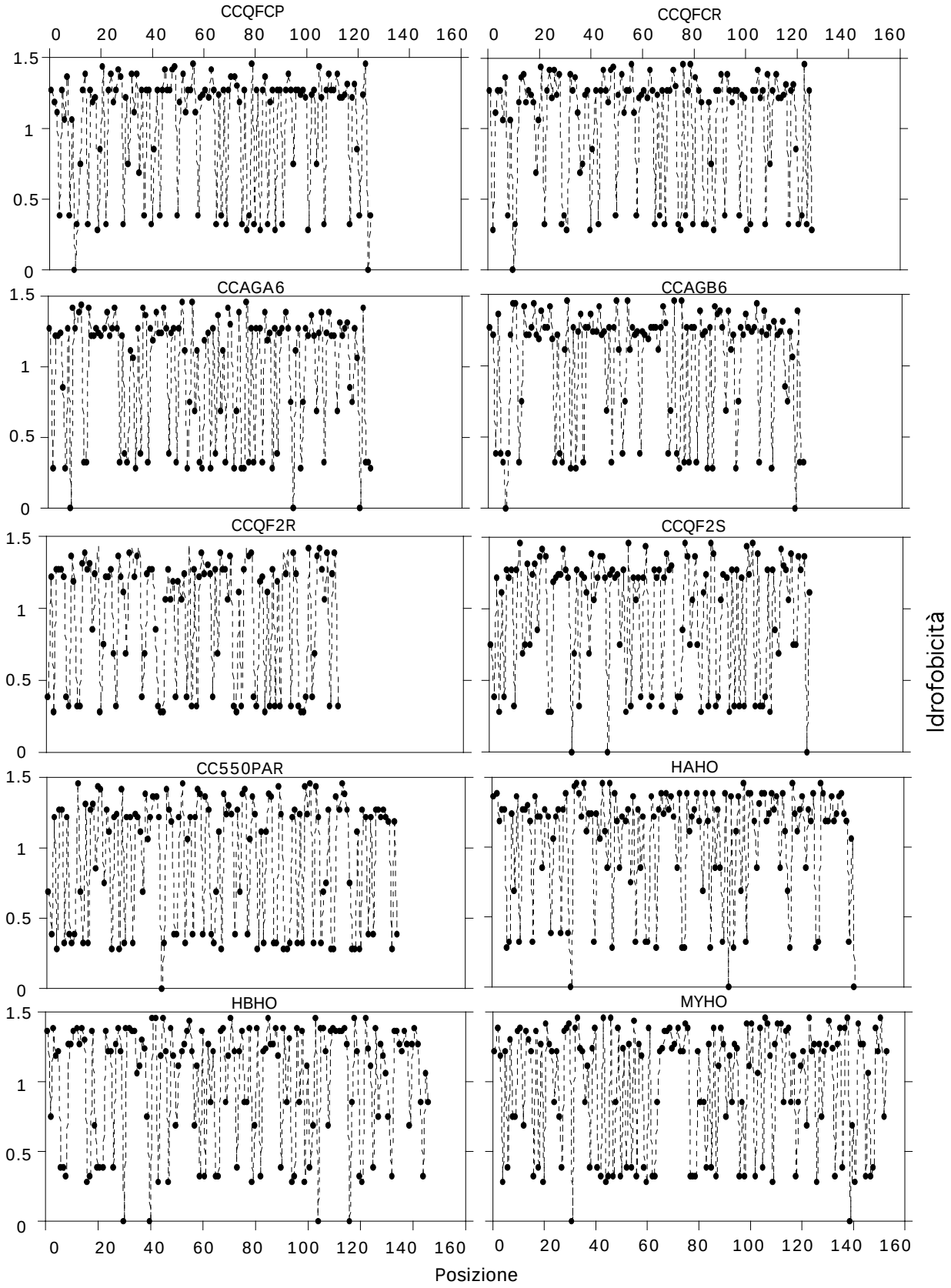


Figura 4.2 - **Profili di idrofobicità delle proteine del training set**

Nel grafico sono riportati i dieci profili di idrofobicità relativi alle proteine del training set riportato in Tabella 4.1. I valori di idrofobicità derivano da quelli riportati in Tabella 3.1, i quali sono stati riscalati in modo tale da avere media unitaria.

Tabella 4.2 : Percentuali di similarità tra le proteine elencate in Tab. 4.1.

1	100.0									
2	68.8	100.0								
3	52.5	52.1	100.0							
4	52.9	55.9	77.0	100.0						
5	35.8	40.4	46.0	48.0	100.0					
6	34.2	40.2	48.5	50.0	64.3	100.0				
7	43.2	44.6	40.8	35.6	65.4	64.5	100.0			
8	42.4	35.9	37.1	41.5	43.1	36.6	37.0	100.0		
9	39.0	43.1	45.6	38.6	38.5	40.3	37.1	67.7	100.0	
10	44.8	34.1	41.5	46.5	38.5	41.7	42.6	46.1	49.6	100.0
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Per ciascuna coppia formata dalle proteine elencate nella Tab. 4.1 (alle quali si riferiscono i numeri in grassetto) é stata calcolata la similarità per mezzo dell'algoritmo classico di Needlemann e Wunsch. Per il calcolo sono stati utilizzati per le *Gap penalties* i valori di default della versione implementata nella suite GCG (GapWeight=3.00 e LengthWeight=0.10). La matrice di sostituzione è quella di Dayhoff [Dayhoff et al., 1978].

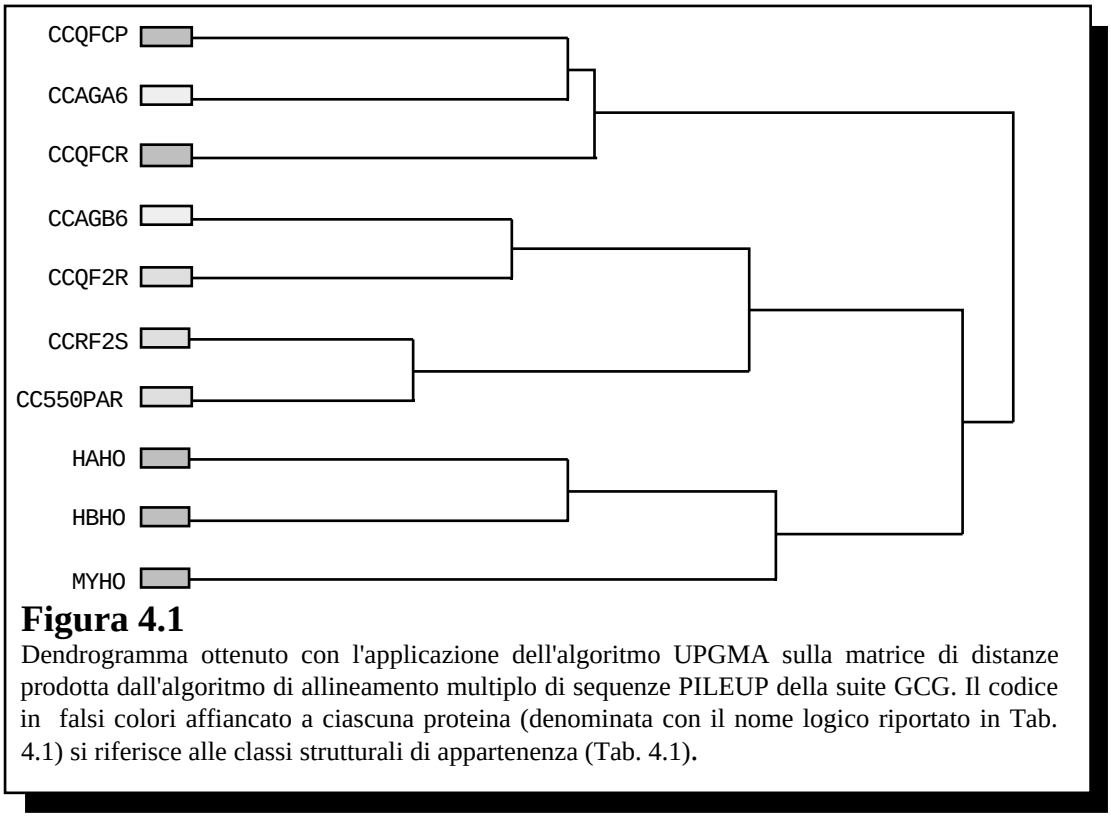


Figura 4.1
Dendrogramma ottenuto con l'applicazione dell'algoritmo UPGMA sulla matrice di distanze prodotta dall'algoritmo di allineamento multiplo di sequenze PILEUP della suite GCG. Il codice in falsi colori affiancato a ciascuna proteina (denominata con il nome logico riportato in Tab. 4.1) si riferisce alle classi strutturali di appartenenza (Tab. 4.1).

Una delle particolarità già discusse dell'algoritmo *SOMA* utilizzato è la conservazione della topologia (si veda il paragrafo 3.1.1) : proteine simili (per ogni codifica utilizzata)risultano vicine nella mappa finale. La valutazione del risultato deve essere fatta tenendo conto dell'obiettivo : l'individuazione della codifica di struttura primaria che più *conserva*

l'informazione legata al conformazione tridimensionale della proteina minimizzando al tempo stesso la ridondanza, unitamente alla realizzazione di un appropriato criterio di classificazione. In questo senso va considerata l'utilità del codice dei colori (associato alla appartenenza, nota precedentemente, a una determinata classe strutturale) : la mappa finale è tanto migliore quanto più raggruppa in unità adiacenti proteine appartenenti alla stessa classe strutturale, e quanto più separa proteine appartenenti a classi diverse.

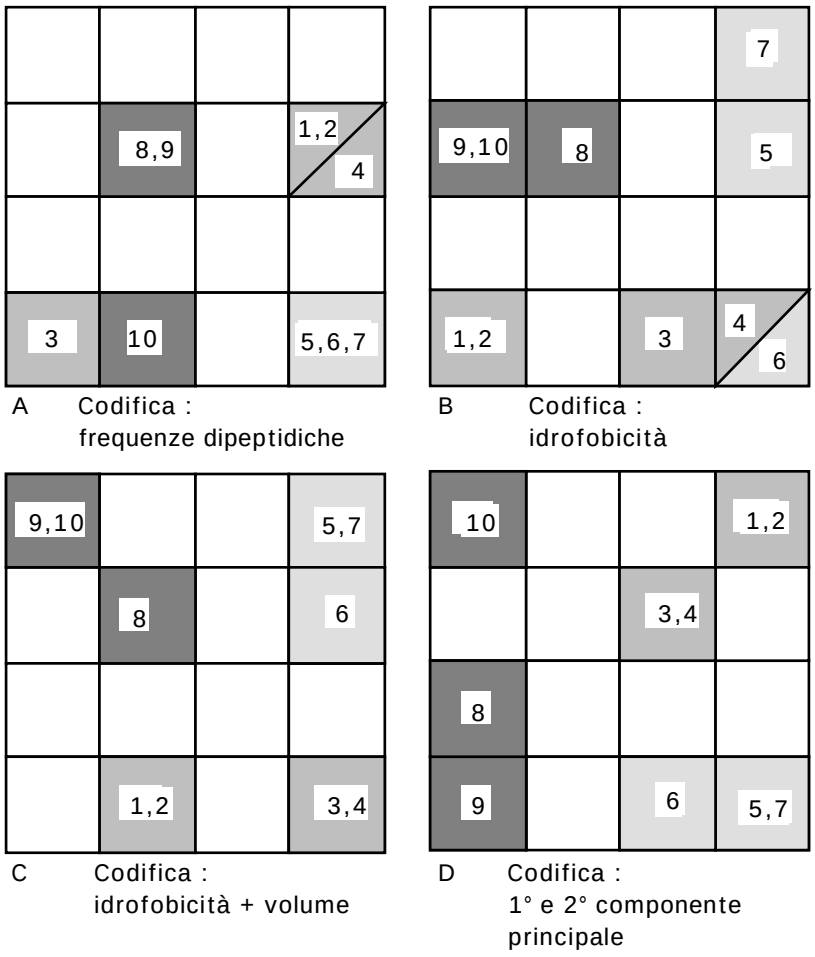


Fig.4.3 Effetti della scelta della codifica sulla classificazione ottenuta con l'algoritmo *SOMA* .

I quattro grafici mostrano la classificazione prodotta per lo stesso set di proteine (riportato in Tabella 4.1) in base a differenti schemi di codifica delle strutture primarie.

A) = frequenze dipeptidiche; **B)** e **C)** = profili di Idrofobicità e Idrofobicità + Volume del residuo, rispettivamente; **D)** = profili numerici basati sulle prime due componenti principali.

I parametri utilizzati per regolare l'evoluzione dell'algoritmo *SOMA* [Kohonen, 1984] sono gli stessi in tutte le situazioni esplorate. La riduzione di dimensionalità è, nei quattro casi, : **A)** 400 2 ; **B)** 133 2 ; **C)** e **D)** 265 2 .

Conclusioni

Nella mappa D della Fig. 4.3, ottenuta utilizzando come descrittori degli aminoacidi le prime due componenti principali, si concentra più dell'84% dell'informazione relativa alle sette grandezze fisico-chimiche della Tabella 3.1 in due sole variabili, realizzando così una ottimizzazione della codifica (e una conseguente riduzione di costo computazionale). Come si vede nel grafico (aiutati dalla presenza delle etichette in falsi colori) si ottiene una mappa in perfetto accordo con la classificazione strutturale di riferimento : le unità appartenenti alla stessa classe strutturale sono rappresentate in posizioni adiacenti sulla mappa.

Per ottenere un indice numerico della *bontà* della classificazione (in funzione della codifica) sono state calcolate le matrici di distanza tra le posizioni, per ciascuna mappa, associate a ciascuna proteina del training set. Le quattro matrici di distanza sono state quindi confrontate con la matrice di distanza ottenuta dall'applicazione dell'algoritmo di Needlemann e Wunsch (vedi Tabella 4.2) ed è stato ricavato, per ciascun confronto, un valore di correlazione.

I risultati, riportati in Tabella 4.3, confermano il fatto che la codifica che utilizza le prime due componenti principali produce la mappa che più è in accordo con la matrice di riferimento.

Tabella 4.3 : **Matrice di correlazione tra i valori di similarità (Tabella 4.2) e i risultati di classificazione.**

Fig. 4.3 A	1.00				
Fig. 4.3 B	0.43	1.00			
Fig. 4.3 C	0.35	0.58	1.00		
Fig. 4.3 D	0.47	0.62	0.52	1.00	
Tab. 4.2	0.65	0.74	0.66	**0.82**	1.00
	Fig. 4.3 A	Fig. 4.3 B	Fig. 4.3 C	Fig. 4.3 D	Tab. 4.2

I coefficienti di correlazione (Pearson) sono stati calcolati a partire dai dati riportati nella Tabella 4.2 e le distanze di Manhattan (ricavate contando il numero minimo di passi, orizzontali o verticali, per congiungere due posizioni su un riferimento *a scacchiera*) tra le proteine, ottenute dalle mappe riportate in Fig. 4.3 .

4.2 Esperimento 2 : Test di performance del metodo

Scopo dell'esperimento

Per verificare le prestazioni del metodo nei confronti di un problema di classificazione considerato "difficile" [Pascarella & Argos, 1992] è stato utilizzato un set di proteine (riportate in Tabella 4.4) più ampio del precedente (22 strutture primarie). Le proteine selezionate sono state assegnate a cinque gruppi strutturali (sempre in base a evidenze di carattere sperimentale).

Il set di 22 proteine :

##	SW_Code	AA	Struct Family
(01)	Azu1$Metj	128	Plasto/Az
(02)	Azu2$Metj	129	Plasto/Az
(03)	Cabo$Lolpe	149	CA_Bind
(04)	Calm$Dicdi	151	CA_Bind
(05)	Carp$Rhich1	147	Ac_Prot
(06)	Carp$Rhich2	178	Ac_Prot
(07)	Catr$Chlre	169	CA_Bind
(08)	H81$Heigo	183	Plasto/Az
(09)	Lca$Macrg	121	Lyz
(10)	Lyc2$Pig	128	Lyz
(11)	Lycv$Bpt4	164	Lyz
(12)	Lyc$Horse	129	Lyz
(13)	Mlen$Human	150	CA_Bind
(14)	Pa20$Notsc	145	Lipase
(15)	Pa22$Bitna	119	Lipase
(16)	Pa23$Oxysc	133	Lipase
(17)	Pa2c$Vipaa	138	Lipase
(18)	Penp$Penja1	149	Ac_Prot
(19)	Penp$Penja2	174	Ac_Prot
(20)	Plas$Arath	171	Plasto/Az
(21)	Plas$Horvu	155	Plasto/Az
(22)	Tpcs$Mouse	159	CA_Bind

Tabella 4.4 - La lista di 22 proteine
Le proteine, elencate in ordine alfabetico in base al loro codice nella nomenclatura standard SWISS_PROT sono seguite da **i)** il numero di aminoacidi di cui sono composte, **ii)** il nome della classe strutturale alla quale sono state assegnate sulla base di differenti considerazioni semiempiriche [Pascarella & Argos, 1992]

L'aspetto rilevante del problema, e, insieme, la sua difficoltà è che molto spesso queste proteine pur riconosciute come simili dal punto di vista delle strutture tridimensionali (e quindi appartenenti alla medesima classe strutturale) presentano omologie di sequenza (Tabella 4.5) particolarmente basse (nel migliore dei casi $\leq$ 56 %).

Modalità di esecuzione.

E' stata realizzata una codifica numerica basata sull' utilizzo delle prime due componenti principali calcolate sull'insieme di sette proprietà fisicochimiche. A ciò è seguita l'applicazione dell'algoritmo *PBA,* dove il valore intermedio di equalizzazione dei profili è stato fissato a **151*2** residui (il valore centrale tra la minima e la massima

lunghezza moltiplicato per il numero di descrittori utilizzati), e infine la classificazione con l'algoritmo *SOMA* su una rete di 6x6 unità.

		Penp$Penja1	Carp$Rhich1	Penp$Penja2	Carp$Rhich2	
Acidic_proteases	Penp$Penja1	1.0000	0.4375	0.1362	0.1495	
I	Carp$Rhich1		1.0000	0.1162	0.1231	
	Penp$Penja2			1.0000	0.3716	
	Carp$Rhich2				1.0000	
		Lyc$Horse	**Lyc2$Pig**	**Lca$Macrg**	**Lycv$Bpt4**	
Lysozymes	Lyc$Horse	1.0000	0.5447	0.3360	0.0683	
II	Lyc2$Pig		1.0000	0.3534	0.0890	
	Lca$Macrg			1.0000	0.0772	
	Lycv$Bpt4				1.0000	
		Calm$Dicdi	**Cabo$Lolpe**	**Tpcs$Mouse**	**Catr$Chlre**	**Mlen$Human**
Ca_binding proteins	Calm$Dicdi	1.0000	0.6333	0.4839	0.4688	0.4385
III	Cabo$Lolpe		1.0000	0.4481	0.4151	0.3344
	Tpcs$Mouse			1.0000	0.4268	0.2977
	Catr$Chlre				1.0000	0.3009
	Mlen$Human					1.0000
		Plas$Horvu	**Plas$Arath**	**Azu1$Metj**	**H81$Heigo**	**Azu2$Metj**
Plastocyanins and Azurins	Plas$Horvu	1.0000	0.5548	0.1641	0.2194	0.1628
IV	Plas$Arath		1.0000	0.1719	0.1930	0.1318
	Azu1$Metj			1.0000	0.5625	0.5156
	H81$Heigo				1.0000	0.3953
	Azu2$Metj					1.0000
		Pa23$Oxysc	**Pa20$Notsc**	**Pa22$Bitna**	**Pa2c$Vipaa**	
Lipases	Pa23$Oxysc	1.0000	0.4604	0.3730	0.3321	
V	Pa20$Notsc		1.0000	0.3409	0.3322	
	Pa22$Bitna			1.0000	0.4514	
	Pa2c$Vipaa				1.0000	

Tabella 4.5 - Set di proteine selezionato per l'esperimento.

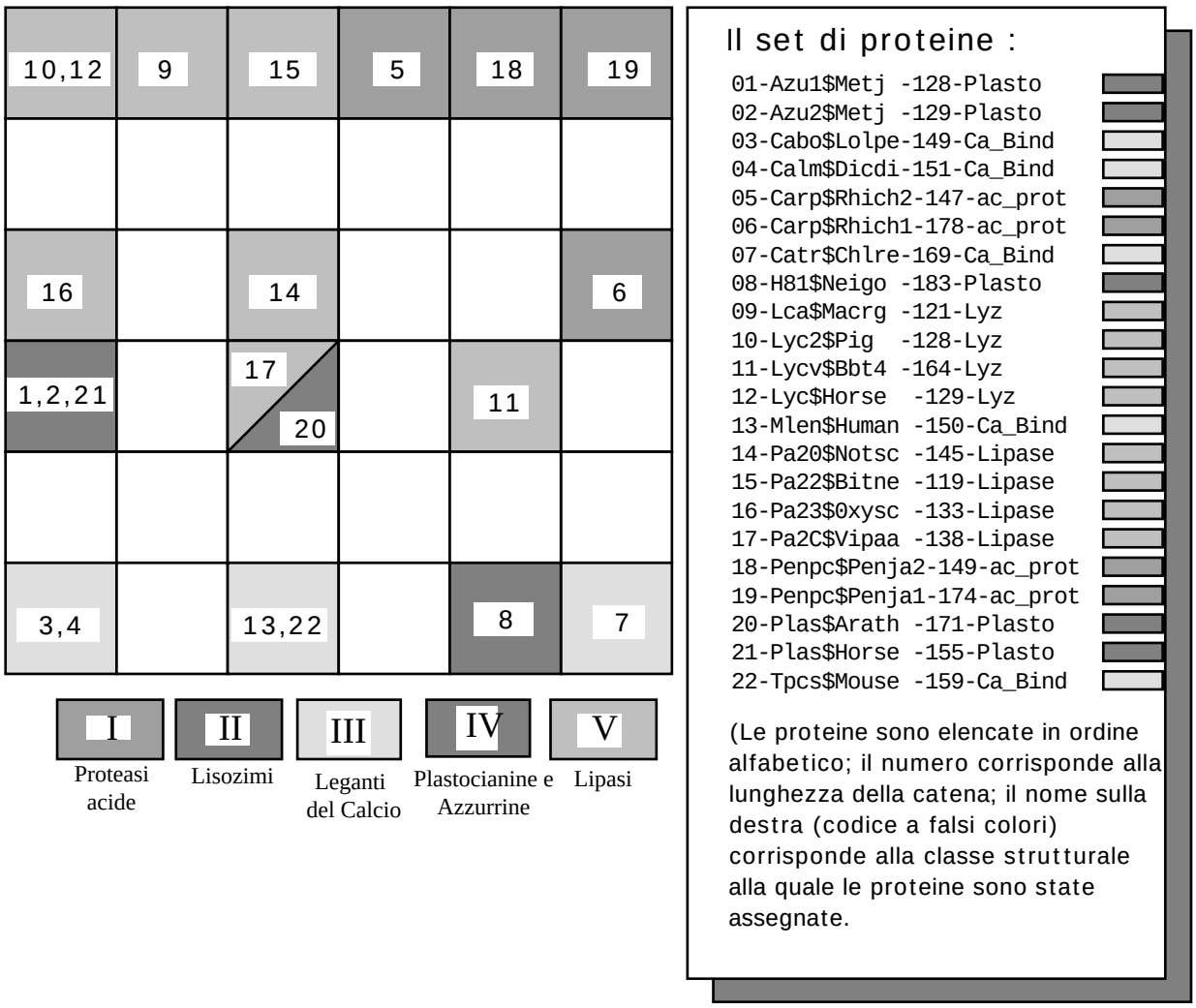
Il codice in falsi colori indica la classe strutturale alla quale appartengono le proteine ed è lo stesso utilizzato nelle Figure 4.4 e 4.5, riportanti i risultati di classificazione con l'algoritmo *SOMA* e con l'algoritmo PILEUP. Le matrici numeriche riportano invece i punteggi di omologia di sequenza calcolati sempre con l'algoritmo PILEUP e la matrice di identità (0/1 = Diverso/Uguale) per ciascuna famiglia strutturale.

Discussione

Nonostante le considerazioni fatte all'inizio del paragrafo sulla bassa omologia tra sequenze di stesse classi strutturali, i risultati riportati (Figura 4.4) ottenuti con l'utilizzo degli algoritmi *SOMA* e *PBA* possono essere considerati molto incoraggianti.

Utilizzando infatti il codice a falsi colori indicante la classificazione strutturale nella rappresentazione della mappa di classificazione prodotta si può osservare una disposizione delle unità di uscita selezionate da ciascuna struttura primaria tale da assegnare a elementi della medesima classe strutturale posizioni adiacenti sulla mappa.

Figura 4.4 - Mappa di classificazione ottenuta con l'algoritmo *SOMA*



Le incongruenze osservabili (la posizione delle proteine corrispondenti ai numeri d'ordine **11** e **8**), dipendono probabilmente da una scelta non ottimale dei parametri di controllo dell'algoritmo *SOMA* o da una codifica non ancora sufficientemente calibrata, o anche, ipotesi non trascurabile, dall'individuazione di elementi di similarità nei profili numerici non in accordo con le classi di derivazione semiempirica usate in questo caso come riferimento.

Sempre per effettuare una sorta di validazione esterna dell'algoritmo di classificazione e del metodo di codifica, è stata prodotta una classificazione delle 22 strutture primarie con gli algoritmi tradizionali della suite GCG : il programma di allineamento multiplo PILEUP é stato utilizzato per l'intero set di proteine e la matrice di similitudine prodotta è stata a sua volta utilizzata per generare un dendrogramma contenente le relazioni mutue tra tutte le strutture primarie.

Conclusioni

I risultati, riportati in Figura 4.5, rivelano un buon accordo tra i raggruppamenti che emergono dalle due analisi (*SOMA* e PILEUP), come appare dal confronto tra la mappa bidimensionale e la struttura ad albero riportate.

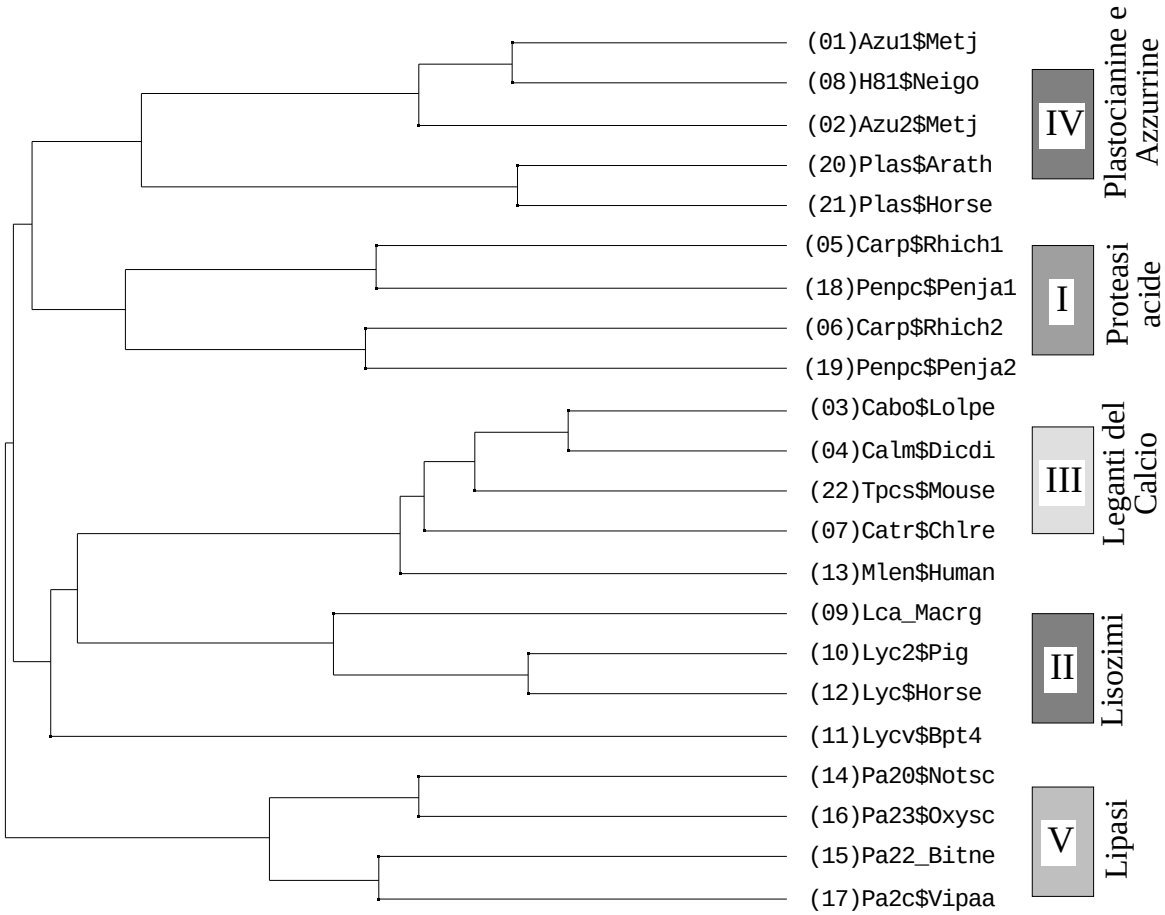


Figura 4.5 - Classificazione ottenuta con l'algoritmo di allineamento multiplo PILEUP.

Il dendrogramma è stato prodotto a partire dalla matrice di distanze ottenuta con l'algoritmo PILEUP applicato sull'intero set di 22 proteine. I numeri accanto a ciascun nome si riferiscono ai numeri d'ordine della lista e sono quelli utilizzati nella mappa di classificazione riportata in Figura 4.4 . I codici a falsi colori si riferiscono alla classificazione strutturale.

Osservando le due codifiche è possibile fare le seguenti osservazioni :

a) i gruppi II ( ) e IV ( ), rispettivamente i lisozimi e le plastocianine/azzurrine, sono riprodotti relativamente bene, con le sole eccezioni delle proteine rispettivamente n. **11** (Lycv$Bpt4) e n. **6** (Carp$Rhich1); per quello che riguarda la n. **11** va però rilevato che anche nella classificazione ottenuta con l'allineamento multiplo di PILEUP la sua appartenza alla classe **II** è discutibile. Dal dendrogramma sembra piuttosto un *outlier*.

b) i gruppi I ( ), III ( ) e V ( ), rispettivamente proteasi acide, leganti del calcio e lipasi, appaiono poco *clusterizzate* nella mappa di classificazione, essendo distribuite su più di due gruppi di unità non contigue.

Commentare le osservazioni fatte è difficile, poiché le due classificazioni sono basate su criteri di codifica delle strutture primarie differenti, e cioè una codifica simbolica per la classificazione prodotta da PILEUP ed una fisico-chimica utilizzata dall'algoritmo *SOMA*.

Diventa quindi cruciale l'individuazione di un criterio di validazione *esterno* delle classificazioni ottenute con l'applicazione degli algoritmi *PBA* e *SOMA* che parta, più che dall'analisi delle sequenze di simboli (i nomi degli aminoacidi nella sequenza), dalla elaborazione di osservabili fisiche. A questo scopo è stato progettato l'Esperimento 3, descritto nel prossimo paragrafo.

4.3 Esperimento 3 : Validazione oggettiva del metodo.

Scopo dell'esperimento

Testare l'affidabilità della procedura di classificazione per mezzo di uno standard oggettivo e indipendente.

Modalità di esecuzione

Un criterio di validazione *esterno* delle classificazioni di strutture primarie prodotte con gli algoritmi *PBA* e *SOMA* deve partire direttamente dall'informazione disponibile con l'analisi cristallografica delle proteine, e perciò dai dati relativi alle loro conformazioni tridimensionali. A tale scopo, per questo terzo esperimento, è stato utilizzato l'algoritmo *OPA* (Onion Peel's Algorithm) descritto nel par. 3.2.2 .

La scelta del training set, questa volta, è stata fatta tra quelle proteine di cui è nota la struttura tridimensionale, e di cui è disponibile su banca dati il relativo *file* nello standard **PDB** (Protein Data Bank). L'ulteriore vincolo, dettato dall'utilizzo dell'algoritmo *SOMA*, è che le lunghezze delle catene polipeptidiche non fossero troppo differenti (nei limiti dell'applicabilità di *PBA*). Nella Tabella 4.6 sono elencate le strutture terziarie scelte per l'esperimento, e per ciascuna sono riportati, nell'ordine, il nome logico PDB, il nome logico SwissProt/NBRF, il nome della proteina, la classe strutturale di appartenenza, la sorgente, e il numero di aminoacidi.

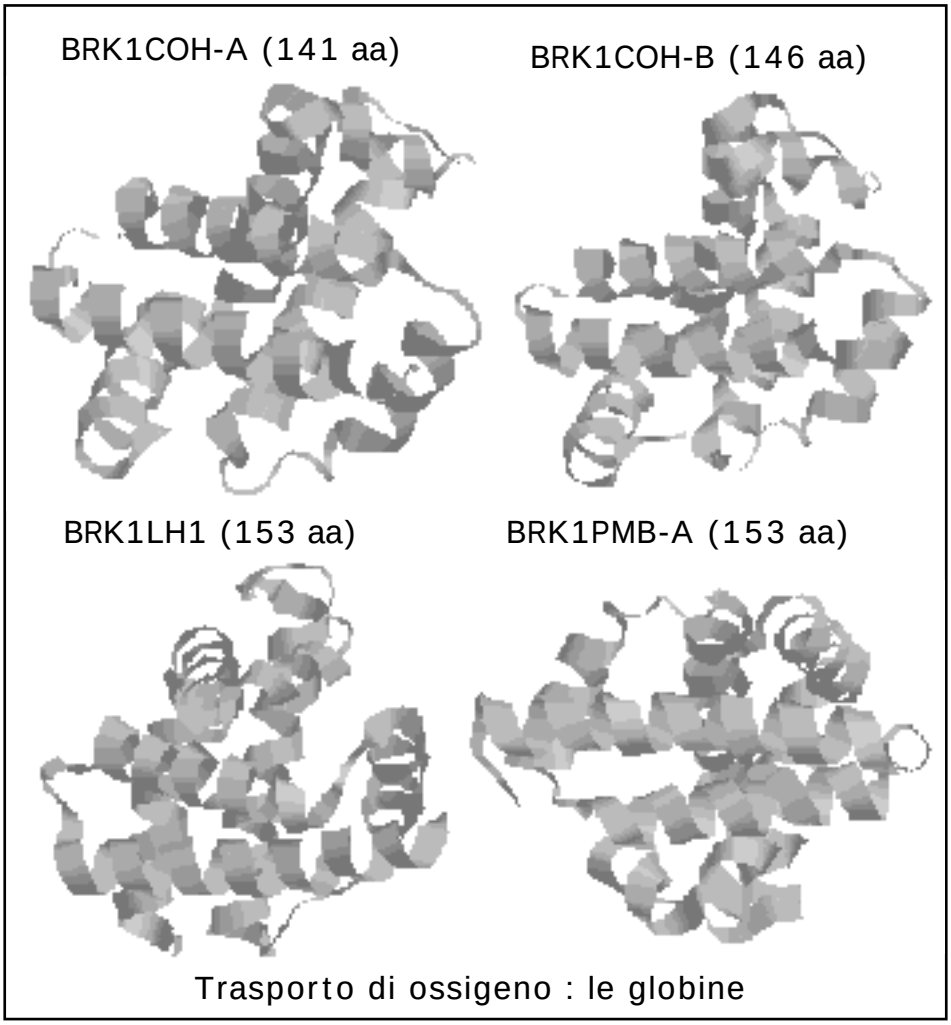
Tabella 4.6 - **Elenco delle proteine utilizzate per l'esperimento 3**

#	Nome (PDB)	Nome (SW/NBRF)	Nome	Classe strutturale	Sorgente	AA
1	BRK3CYT	Cyc_katpe.Sw	Cyt $c (Ox)	**I**-Electr trans (Heme)	Thunnus alalunga	103
2	BRK5CYT	Cyc_katpe.Sw	Cyt $c (Red)	**I**-Electr trans (Heme)	Thunnus alalunga	103
3	BRK1AZU	Azpsca.Pir1	Azurin	**II**-Electr trans (Cu)	Pseudom Aerug	128
4	BRK1PCY	Plas_popni.Sw	Plastocyan (Cu++)	**II**-Electr trans (Cu bnd)	Populus nigra it.	99
5	BRK1PMB	Myg_horse.Sw	Myoglob (Aquomet)	**III**-Oxyg storage	Sus scrofa	153
6	BRK1COH-A	Hba_horse.Sw	Hemoglob (Alpha)	**III**-Oxyg transp	Human	141
7	BRK1COH-B	Hbb_horse.Sw	Hemoglob (Beta)	**III**-Oxyg transp	Human	146
8	BRK1LH1	Lgb2_luplu.Sw	LegHemogl (Acetate)	**III**-Oxyg transp	Lupinus luteus	153
9	BRK1LZ1	Lyc_chick.Sw	Lysoz (Tricl X form)	**IV**-Hydrolase (O-Glyc)	Gallus gallus	129
10	BRK1LZT	Lyc_human.Sw	Lysoz	**IV**-Hydrolase (O-Glyc)	Human	130
11	BRK1LYM	Lyc_chick.Sw	Lysoz	**IV**-Hydrolase (O-Glyc)	Gallus gallus	129

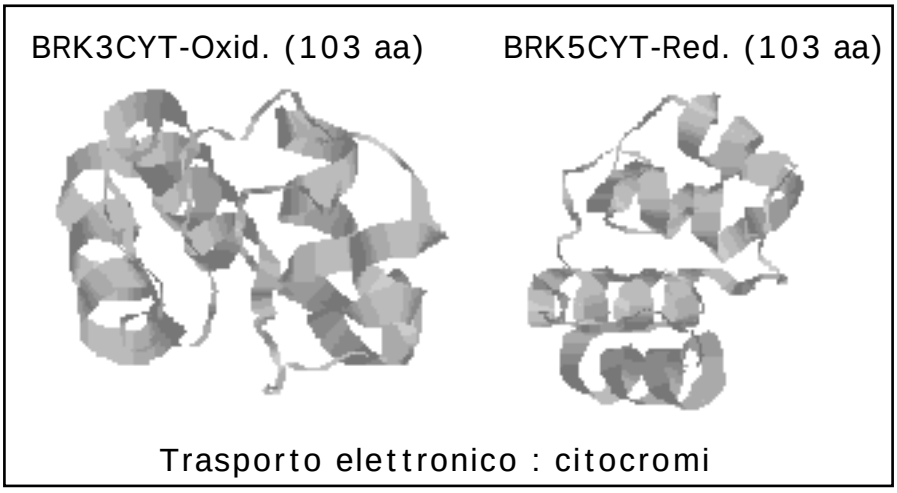
Si fa notare che due casi (le coppie 1-2 e 9-11) sono degeneri nell'associazione struttura terziaria / struttura primaria : alla stessa sequenza aminoacidica corrispondono differenti strutture terziarie.

Nelle pagine seguenti sono state riportate le immagini, prodotte con il software RasMol®, delle strutture 3D utilizzate. Esse sono state raggruppate per classi, in modo da facilitare l'individuazione di somiglianze tra proteine della stessa classe. Ovviamente, in questo caso, la

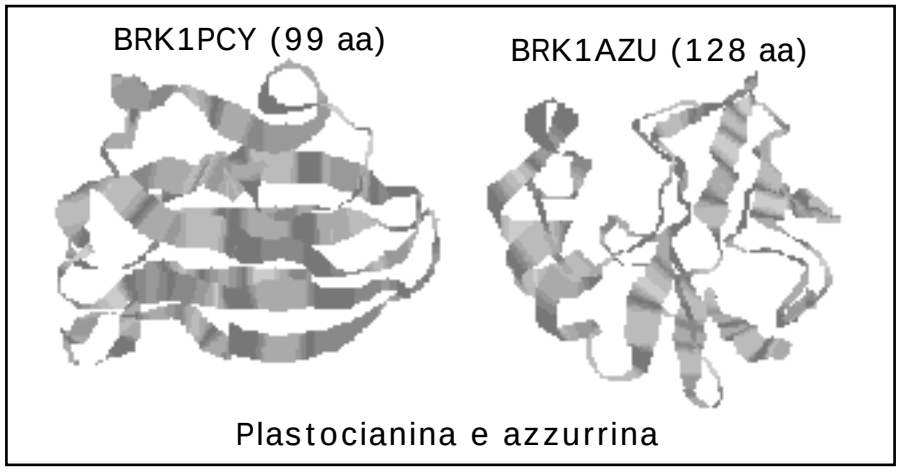
valutazione della somiglianza è fortemente dipendente dall'orientazione della molecola nello spazio e presenta degli aspetti che talvolta sono valutabili solo in modo soggettivo e dettato dall'esperienza personale.

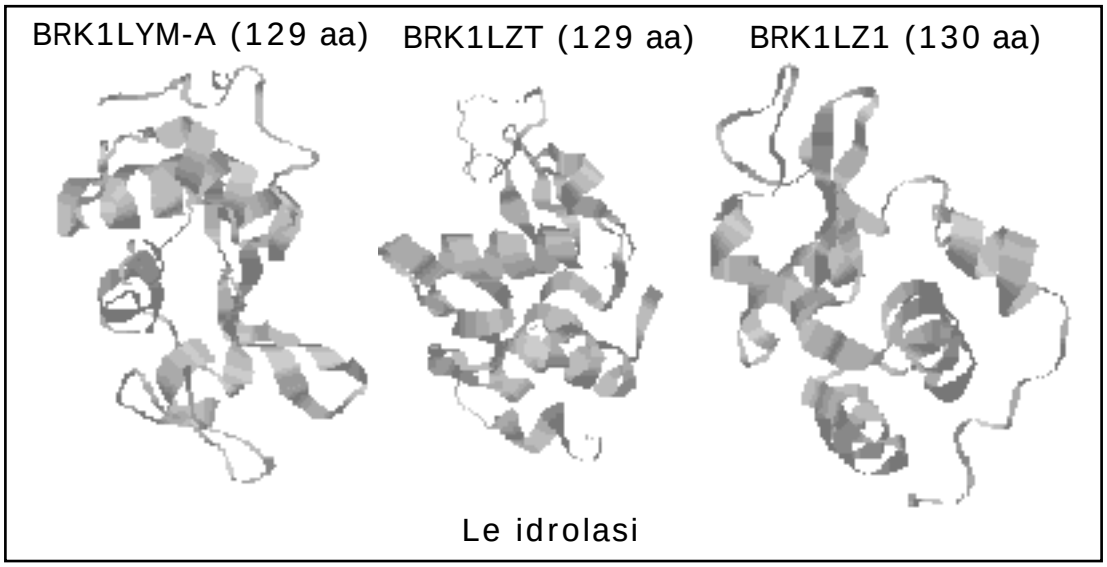


Trasporto di ossigeno : le globine



Trasporto elettronico : citocromi





I files contenenti le strutture terziarie delle proteine sotto forma di coordinate tridimensionali atomiche nel formato standard PDB sono stati analizzati da un software appositamente realizzato, e, limitandosi ai dati relativi al *backbone* della proteina, sono stati calcolati (in una prima fase), per ciascuna struttura i seguenti parametri :

1) le coordinate (X_{CM}, Y_{CM}, Z_{CM}) del baricentro del backbone ;

2) la massima e la minima distanza dei C-alpha dal baricentro ;

3) la massima distanza tra due C-alpha della catena (diametro) ;

4) la dimensione frattale del backbone [Katz, 1988].

Mentre le seconde due grandezze hanno come unico scopo una ulteriore caratterizzazione quantitativa della conformazione tridimensionale delle proteine studiate, le prime due sono state utilizzate per la codifica delle strutture terziarie, in base all'algoritmo *OPA* precedentemente descritto, secondo la seguente procedura :

5) determinazione, tra tutte le proteine del set, della massima e minima distanza dal baricentro, in modo da avere **a)** un riferimento globale sul *range* di distanze nel quale studiare la

distribuzione radiale di residui, e **b)** un *passo* di incremento normalizzato sull'intero set di strutture (vedi il punto successivo);

6) una volta scelto arbitrariamente il numero N di intervalli con il quale partizionare il *range* di distanze dal centro per ogni struttura (nell'esperimento descritto è stato posto N = 50), è stata calcolata la distribuzione radiale di residui negli N intervalli di distanza per ciascuna proteina, ottenendo così, per ciascuna di esse, un vettore descrittore di struttura a N componenti. Nella Figura 4.6 seguente sono riportati i grafici delle distribuzioni radiali di aminoacidi calcolate per l'intero set di proteine ;

7) la matrice contenente gli 11 vettori a 50 componenti viene analizzata con l'algoritmo di Analisi delle Componenti Principali (precedentemente descritto), al fine di ridurre la ridondanza, e da questa vengono estratte le prime M (con M << 50) componenti, che avranno la funzione di codificare la struttura terziaria della proteina con un numero ridotto di variabili ottimizzate. Nel caso in cui si scelga M=2 è possibile rappresentare le strutture terziarie di proteine su una mappa bidimensionale (in modo simile a quanto succede con l'algoritmo *SOMA*), rendendo immediatamente individuabile la presenza di raggruppamenti localizzati.

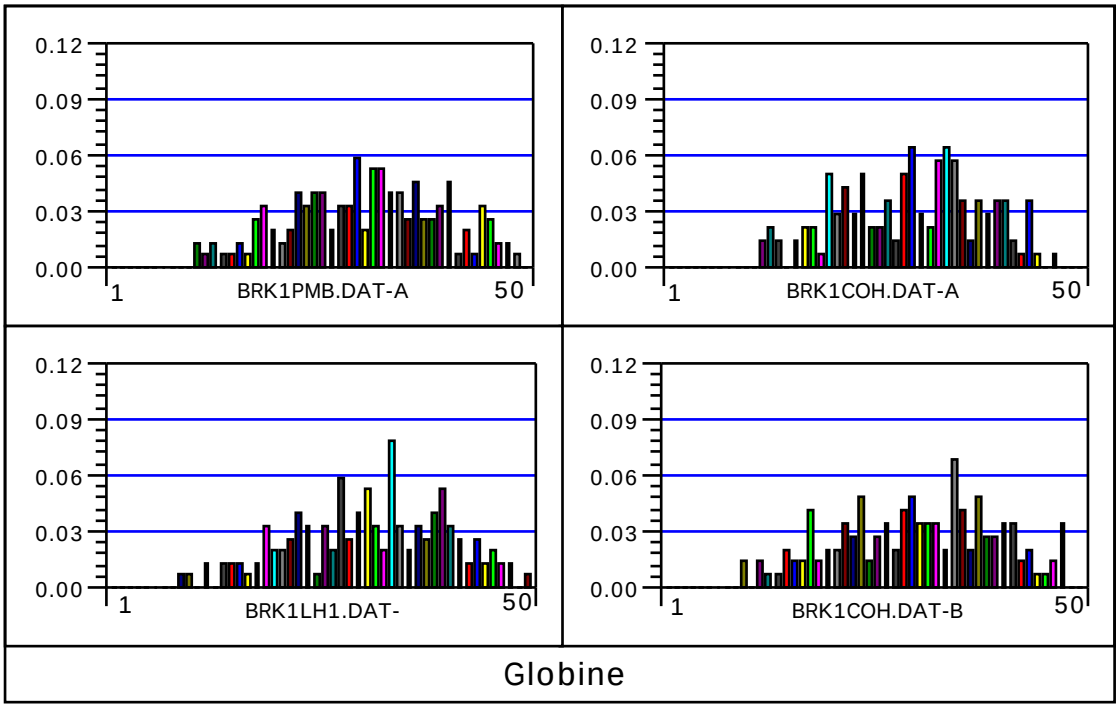
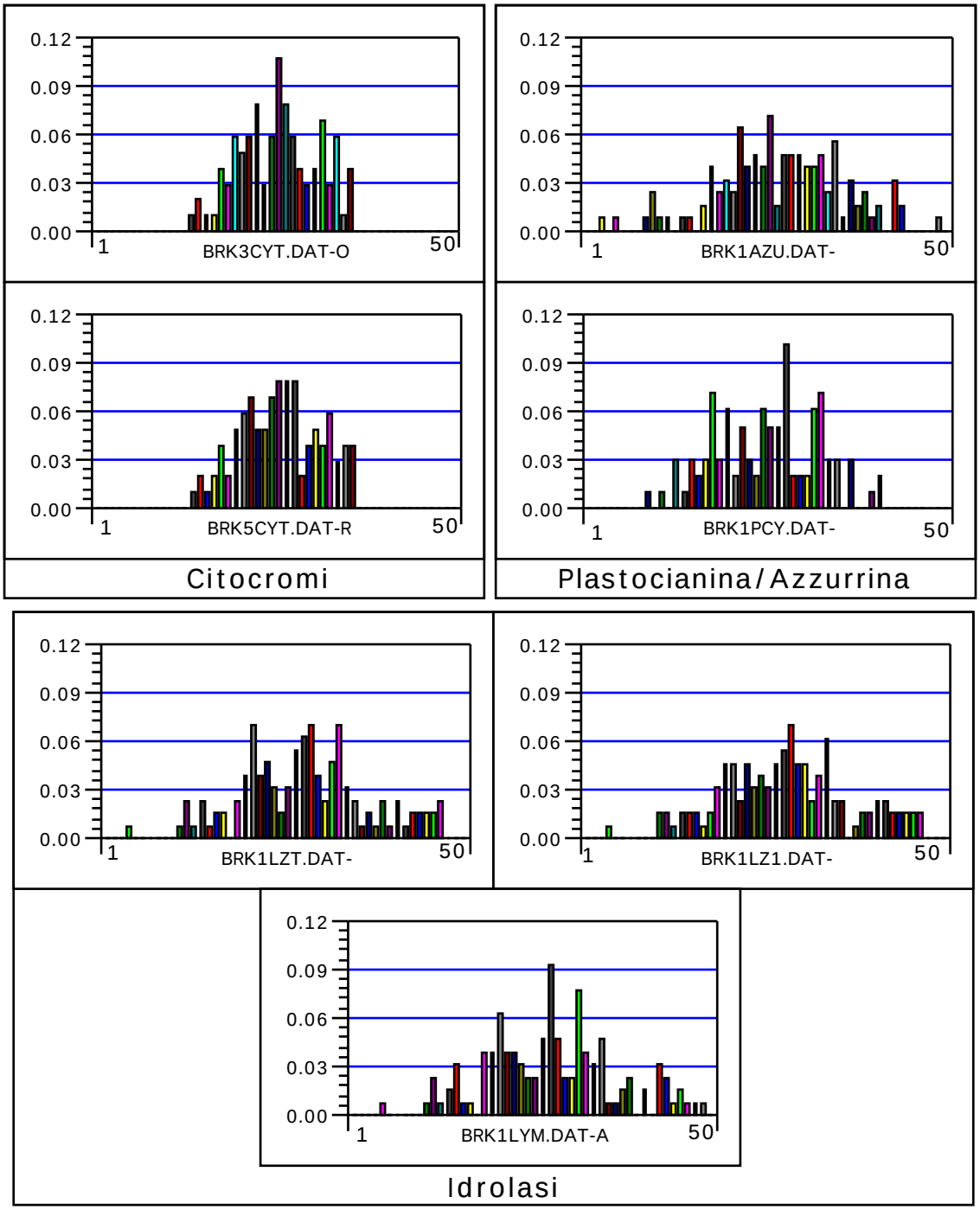


Figura 4.6 Le distribuzioni radiali di residui.

Negli istogrammi riportati sono rappresentate le distribuzioni radiali di C-alpha intorno al baricentro del backbone per ciascuna proteina del set selezionato. Sulle ordinate sono rappresentate le percentuali di occupazione in aminoacidi per ciascuna calotta sferica. Sulle ascisse sono rappresentati gli N (N=50) intervalli di distanza dal baricentro. Il valore massimo ed il valore minimo sono calcolati sull'intero set di proteine, sicché le scale sono normalizzate.



Nella Tabella 4.7 seguente sono riportati, per ciascuna struttura terziaria, i valori relativi al numero di aminoacidi, al diametro, alla dimensione frattale [Katz, 1988], e alle prime quattro componenti principali (con la relativa percentuale di variabilità spiegata) calcolate, come si è detto, sulle distribuzioni radiali di aminoacidi su 50 intervalli di distanza.

Nel grafico successivo (Figura 4.7) viene mostrata la mappa che si ottiene codificando le strutture terziarie delle proteine del set con la prima e la seconda componente principale.

Tabella 4.7

Per ciascuna struttura terziaria sono riportati il numero di aminoacidi, il diametro, la dimensione frattale , e le prime quattro componenti principali (con la relativa percentuale di variabilità spiegata) calcolate sull'insieme di 11 vettori a 50 componenti descriventi la distribuzione radiale di aminoacidi intorno al baricentro del backbone della proteina.

#	Name (PDB)	AA	Diam (Å)	FD	PC1 30.1 %	PC2 14.4%	PC3 12.4%	PC4 11.7%
1	BRK3CYT	103	33.64	2.12	-5.17	2.24	-0.39	-1.11
2	BRK5CYT	103	33.82	2.12	-5.00	3.66	0.55	-0.09
3	BRK1AZU	128	39.38	2.07	0.13	-0.58	-5.96	0.85
4	BRK1PCY	99	37.84	2.00	-4.01	-1.03	1.88	2.53
5	BRK1PMB	153	43.46	2.07	3.49	0.59	-0.17	1.87
6	BRKICOH-A	141	40.88	2.09	2.27	-0.38	2.63	2.60
7	BRK1COH-B	146	43.98	2.04	3.45	1.01	-0.96	2.44
8	BRK1LH1	153	46.87	2.01	5.33	3.54	1.44	-3.25
9	BRK1LZ1	129	43.51	2.00	-0.52	-4.50	1.48	-0.72
10	BRK1LZT	130	43.56	2.00	0.15	-2.42	0.18	-2.29
11	BRK1LYM	129	44.31	1.99	-0.12	-2.13	-0.68	-2.83

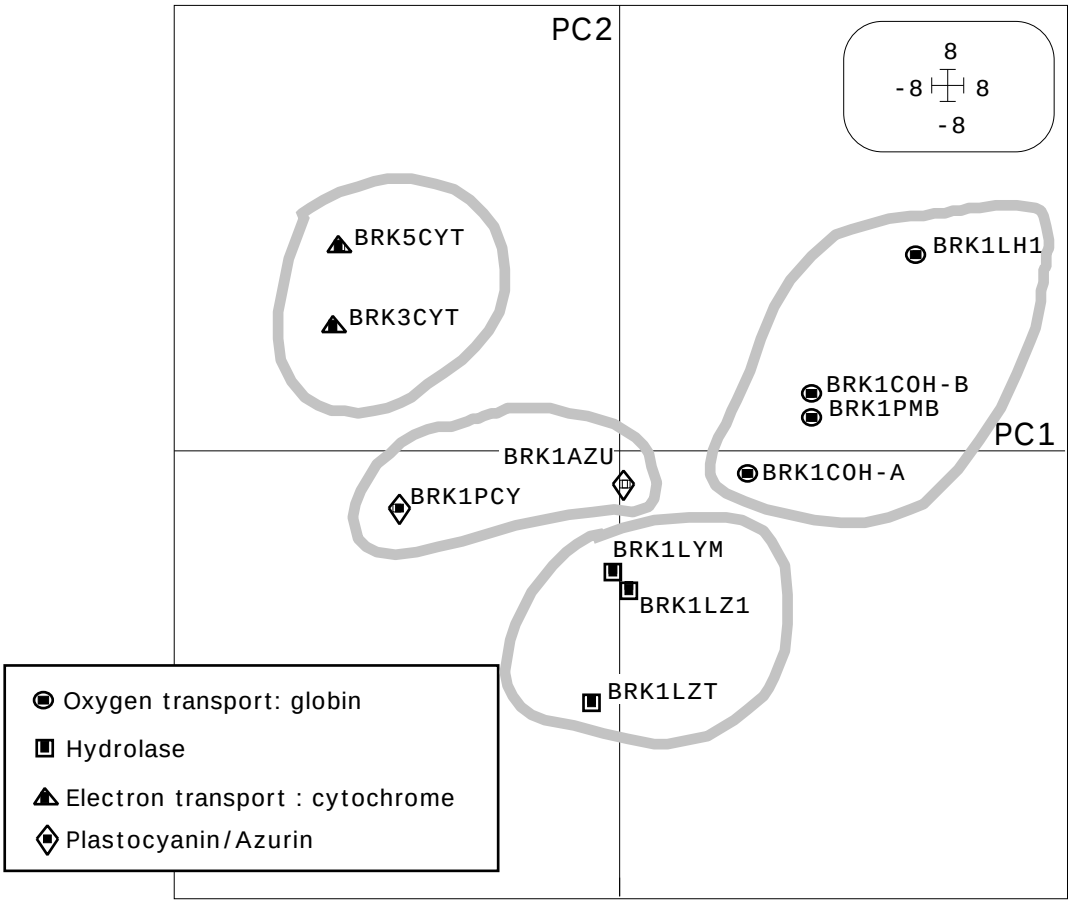


Figura 4.7 - Mappa delle strutture terziarie.

La mappa è stata ottenuta utilizzando le prime due componenti principali calcolate su un set di 11 strutture terziarie descritte da vettori a 50 elementi contenenti la distribuzione radiale di aminoacidi attorno al baricentro del backbone. Le diverse classi strutturali sono rappresentate da simboli differenti (riportati in legenda). Le proteine sono chiamate con il loro nome logico nello standard PDB. Le

delimitazioni dei clusters sono state fatte per evidenziare i raggruppamenti, che riproducono correttamente la classificazione attesa.

Per ottenere una ulteriore rappresentazione delle relazioni di similitudine ottenibili con la codifica basata sull'algoritmo *OPA* è stato prodotto un dendrogramma a partire dalla matrice completa 11x50, ed il risultato, riportato in Figura 4.8, conferma anch'esso l'accordo tra la classificazione ottenuta e quella attesa, basata su criteri tradizionali.

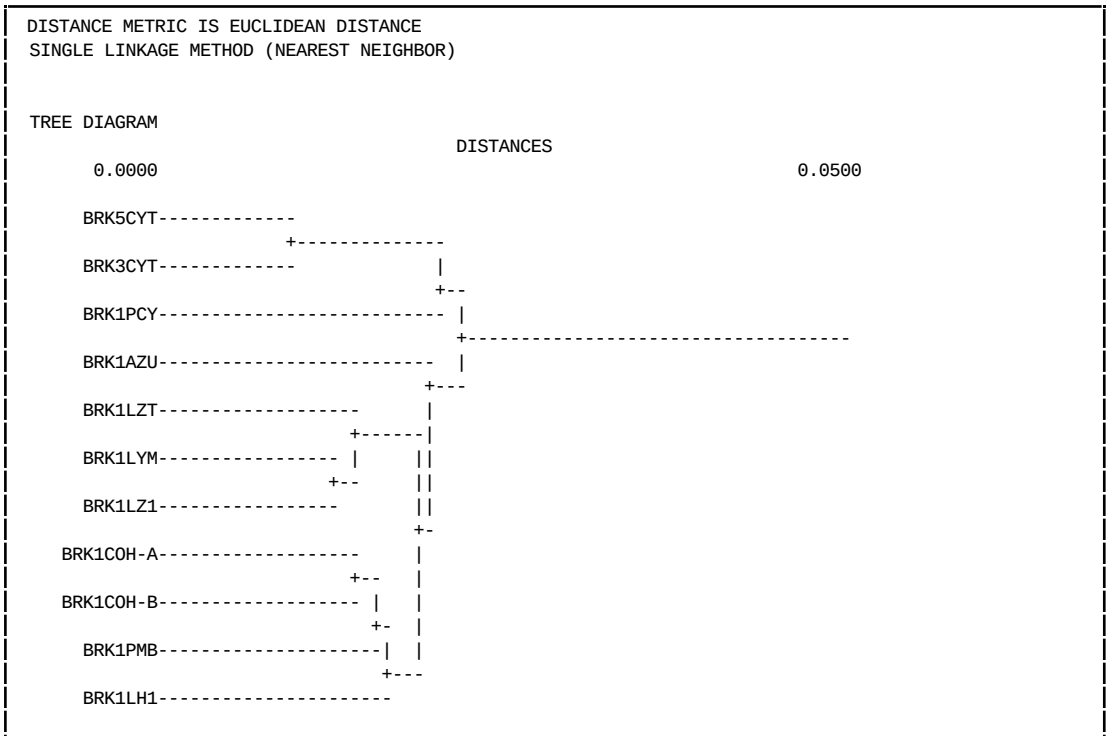


Figura 4.8 - Dendrogramma delle strutture terziarie.
Il dendrogramma è stato ottenuto a partire dalla matrice di distanze ricavata dalla matrice completa 11x50 contenente le distribuzioni di radiali di aminoacidi attorno al baricentro per tutte le proteine del set.

L'utilizzo delle componenti principali per la descrizione di una grandezza multivariata ha l'effetto di far passare da una codifica con variabili delle quali si conosce il significato fisico a una nella quale alle nuove variabili, combinazioni lineari delle prime, spesso non corrisponde una osservabile fisica univocamente determinata.

A tale scopo vengono calcolati i valori di correlazione delle componenti principali con altri descrittori, questa volta corrispondenti a grandezze conosciute. Nella Tabella 4.8 seguente sono riportati, appunto, i valori di correlazione delle prime quattro componenti principali con il numero di aminoacidi (lunghezza), la dimensione frattale e il diametro, per ciascuna proteina del set.

Dalla lettura della tabella si verifica innanzitutto l'attesa ortogonalità tra le tre componenti principali (a correlazione nulla); inoltre emerge il fatto, la cui interpretazione va però fatta solo in forma qualitativa, che la prima componente principale è fortemente correlata con il numero di

aminoacidi e il diametro (rispettivamente, 0.98 e 0.87), mentre la seconda lo è, in maniera leggermente più debole, con la dimensione frattale (0.61). Ciò potrebbe significare, in prima approssimazione, che sull'asse della prima componente principale le strutture terziarie sono disposte con un ordinamento legato alla dimensione (diametro e numero di aminoacidi), mentre sull'asse della seconda componente principale sono disposte in modo da ordinarsi in base alla loro dimensione frattale. A tale proposito si ricorda che questa grandezza è in qualche modo associata alla grado di *aggrovigliamento* di una catena polipeptidica : per fare una analogia, un filo di lana completamente disteso ha una dimensione frattale molto prossima (da valori maggiori) a 1; lo stesso filo, disposto in modo tale da formare una spirale compatta dalla forma di un disco piano senza buchi, ha una dimensione frattale prossima a 2 ; se poi lo stesso filo viene aggomitolato fino ad assumere la forma di una sfera compatta, la sua dimensione frattale sarà prossima (da valori minori) a 3. E' in questa ottica che va interpretata tale misura. Inoltre, i valori di dimensione frattale trovati per le strutture terziarie esaminate sono abbastanza vicini a 2: ciò non vuol dire che le proteine studiate abbiano tutte una forma piatta! Non va dimenticato che tali valori sono relativi al solo backbone della proteina, e la dimensione frattale va intesa come indice del grado di ricoprimento dello spazio. Se fosse stata considerata la struttura terziaria nella sua completezza (tenendo conto delle catene laterali, e dei volumi occupati da ciascun residuo), avremmo trovato dei valori di dimensione frattale molto più prossimi a 3.

Tabella 4.7
Matrice di correlazione tra le prime quattro componenti principali, il numero di aminoacidi, il 'diametro' del backbone e la dimensione frattale.

	# AAs	Diamet	FracDm	PCA_1	PCA_2	PCA_3	PCA_4
# AAs	1.00						
Diamet	0.85	1.00					
FracDm	-0.25	-0.72	1.00				
PCA_1	0.98	0.87	-0.32	1.00			
PCA_2	-0.03	-0.34	0.61	0.00	1.00		
PCA_3	-0.02	0.09	-0.17	0.00	0.00	1.00	
PCA_4	-0.02	-0.26	0.34	0.00	0.00	0.00	1.00

Una volta ottenuta una classificazione delle strutture terziarie con l'algoritmo *OPA*, sono state utilizzate le strutture primarie del medesimo set di proteine per ottenerne una classificazione con gli algoritmi *PBA* e *SOMA* . Anche questa volta codificate con le prime due componenti principali calcolate su un insieme di sette proprietà fisico-chimiche, le 9 strutture primarie (si ricorda che 2 delle 11 proteine sono degeneri nella associazione struttura terziaria/strutturaprimaria) del training set sono state tradotte in profili numerici di **126 x 2** elementi*. Utilizzando anche questa volta, dopo diversi tentativi di esplorazione, una rete di 4x4 elementi di uscita (sufficienti per classificare con efficienza le quattro classi presenti), è stata

* La lunghezza intermedia è data da (MaxAA+MinAA)/2 = (153+99)/2 = 126. Si moltiplica per 2 poiché sono stati utilizzati due descrittori (PC1 e PC2).

ottenuta la mappa delle strutture primarie riportata in Figura 4.9 (insieme a quella delle corrispondenti strutture terziarie già riportata in Figura 4.8).

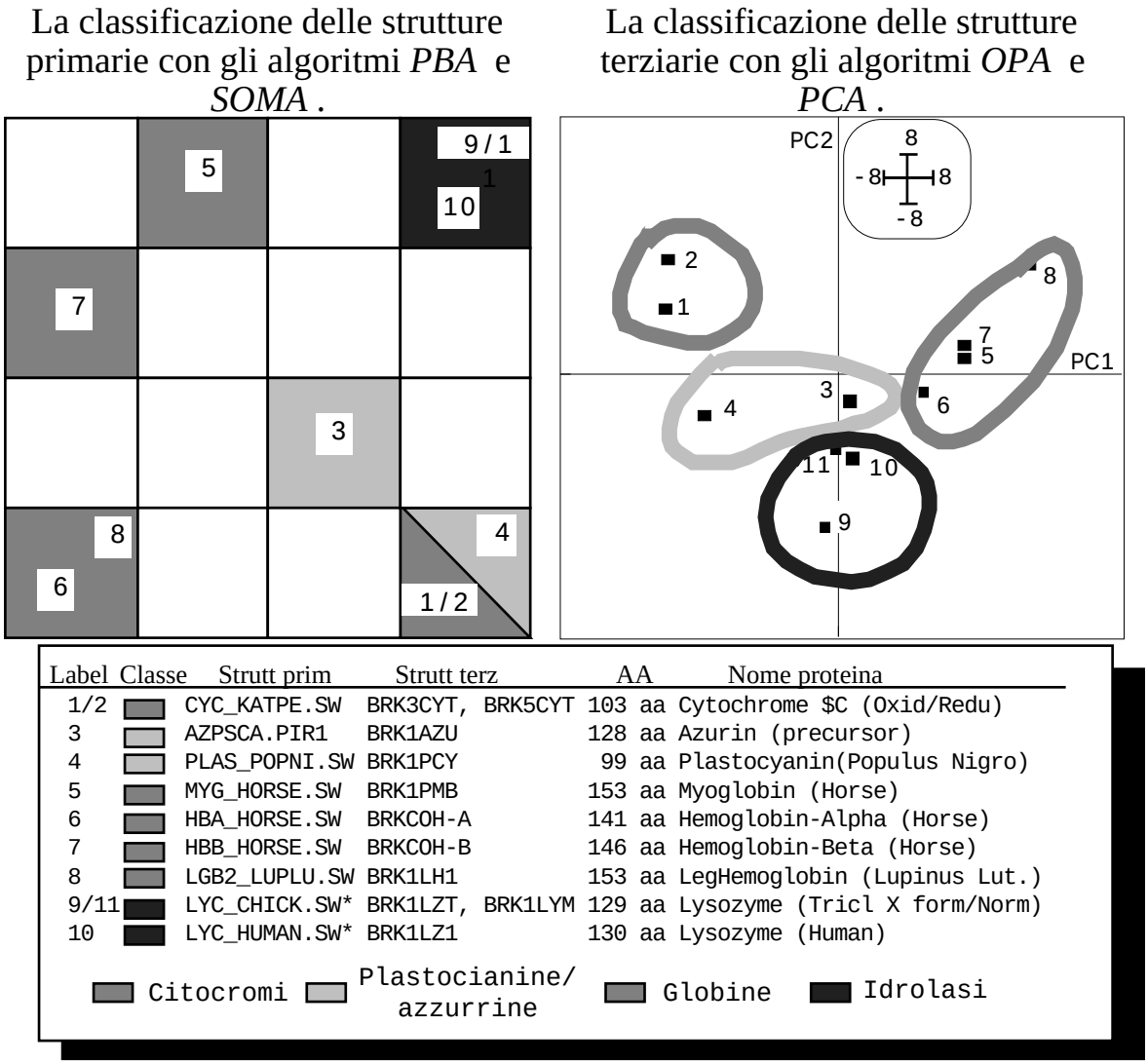


Figura 4.9 - Le mappe di classificazione delle strutture primarie e terziarie.

Le due mappe, ottenute a partire dalle strutture primarie (a sinistra) e terziarie (a destra) del set di proteine riportato in Tabella 4.6, usando rispettivamente gli algoritmi *PBA* e *SOMA* , e *OPA* e *PCA* , mostrano i cluster individuabili nelle rispettive strutture opportunamente codificate (rispettivamente, profilo numerico di componenti principali e distribuzione radiale di residui intorno al baricentro). La mappa delle strutture terziarie è stata riprodotta utilizzando esclusivamente i numeri d'ordine delle strutture (elencati in tabella), al fine di rendere più agevole il confronto con la mappa delle strutture primarie. I codici in falsi colori utilizzati nelle due mappe rappresentano le diverse classi strutturali.

Conclusioni

Dall'analisi delle due mappe emerge il buon accordo tra le due classificazioni, che conferma la correttezza delle scelte fatte per la codifica e la classificazione delle strutture primarie, e quindi l'individuazione di un criterio di analisi delle sequenze aminoacidiche idoneo per assegnare proteine, delle quali si conosce solo la struttura primaria, alla corretta classe di

folding (a patto di avere una matrice di pesi dell'algoritmo *SOMA* sufficientemente addestrata con sequenze delle quali si conosce la classe di struttura terziaria).