Modelli Connessionistici non causali per l'analisi di sequenze e loro impiego nella classificazione delle proteine[1]

Gianluca Pollastri
DSI, Università di Firenze
docbazz@tin.it

Abstract
In questo articolo viene proposta una nuova famiglia di modelli adattivi per l'analisi di sequenze e ne viene discussa l'applicazione al problema della classificazione della struttura secondaria delle proteine. Un sistema di predizione basato sui nuovi modelli ha mostrato, sia in un esperimento di seven-fold cross-validation su 824 sequenze proteiche non omologhe, sia nel confronto coi modelli partecipanti alla competizione CASP '98 di avere prestazioni uguali o superiori ai migliori sistemi esistenti per la classificazione della struttura delle proteine.

1. Introduzione
Da tempo sono noti modelli di tipo connessionistico che permettono di processare informazione strutturata in sequenze. E' questo il caso delle RNN (Recurrent Neural Networks), o dal versante probabilistico degli HMM (Hidden Markov Models) e IOHMM (Input-Output HMM, per cui v.[3]). Una limitazione di questi modelli sta nell'ipotesi di causalità. Sia nel caso delle RNN che degli HMM/IOHMM l'analisi di una sequenza viene effettuata con l'assunzione che gli elementi di questa rappresentino una successione temporale causale.
La classificazione della struttura secondaria delle proteine consiste nell'individuazione della mappa tra la sequenza degli aminoacidi (=AA) di una proteina ed una sequenza (detta struttura secondaria=SS) che ne rappresenta la struttura tridimensionale locale ed è attualmente forse la più importante sfida nel campo della Biologia Computazionale. A partire dai tardi anni '80 l'impatto dei metodi connessionistici ha accresciuto in modo rilevante la qualità delle predizioni della SS, che fino ad allora si erano basate su metodi di tipo simbolico. Nel caso della classificazione della SS il concetto di causalità non ha significato (una sequenza di AA non è una sequenza temporale) ed i modelli che la assumono vera risultano inadeguati.
In questo lavoro è stata sviluppata una nuova famiglia di modelli connessionistici adattivi *non causali* per sequenze. La metodologia adottata è applicabile sia a modelli di tipo deterministico che di tipo probabilistico. Nel primo caso si parlerà di **BRNN** (Bidirectional Recurrent Neural Networks). Per la definizione degli algoritmi di inferenza ed apprendimento si è fatto ricorso allo schema generale per la processazione adattiva di dati strutturati definito in [6]. Il modello è stato quindi applicato al problema della classificazione della SS delle proteine.
Il materiale di questo lavoro è organizzato come segue. Nel **Capitolo 2** si traccerà una descrizione del problema della classificazione della SS. Nel **Capitolo 3** verrà definito il nuovo modello (BRNN). Infine nel **Capitolo 4** saranno descritti i risultati della sperimentazione effettuata applicando i modelli al problema.

2. La classificazione della SS delle proteine
Una proteina è una sequenza un numero variabile (da 20 a 40000) di molecole dette AA. I possibili AA sono 20, dunque una proteina può essere rappresentata come una stringa di simboli da un alfabeto A, con |A|=20. La rappresentazione di una proteina come sequenza di AA è detta struttura primaria ed è oggi disponibile su larghissima scala (centinaia di migliaia di sequenze). D'altra parte una proteina ha una disposizione tridimensionale, connessa alla funzione che è destinata a svolgere, che non è possibile ad oggi dedurre dalla sequenza degli AA, ma che può essere individuata soltanto attraverso la lunga e costosa cristallografia a raggi X ed è dunque oggi disponibile soltanto in alcune migliaia di casi. Una rappresentazione locale della struttura tridimensionale detta SS può essere ottenuta assegnando la posizione di una proteina corrispondente ad ogni AA ad una tra tre classi di struttura, la classe H (Helix), la classe E (Sheet) e la classe C (Coil). La SS è dunque una stringa di lunghezza pari alla sequenza degli AA, con simboli in un alfabeto S con |S|=3. Per un esempio delle due possibili rappresentazioni v. Fig.1.
Il problema della **classificazione della SS** consiste nell'individuazione di una funzione F:

$$F : A^* \rightarrow S^*$$

[1] Il materiale di queste pagine è tratto dalla tesi di laurea di G.Pollastri discussa il 16/4/99 all'Università di Firenze, con relatori il Prof. Giovanni Soda, il Prof. Romano Fantacci, il Prof. Paolo Frasconi ed il Dott. Pierre Baldi.

dove A è un alfabeto di cardinalità |A|=20 (gli AA), S è un alfabeto di cardinalità |S|=3 (la SS) e A^* e S^* sono gli insiemi delle sequenze di lunghezza qualsiasi composte di simboli degli alfabeti A ed S.
Nel problema di predizione della SS la misura più comune per le prestazioni di un sistema è il Q_3 (percentuale globale di classificazione a tre stati), definito come il rapporto tra il numero di residui[2] predetti correttamente N_{corr} ed il numero totale di residui N_{tot} del database in considerazione:

$$Q_3 = 100\ N_{corr}/N_{tot} \qquad (1)$$

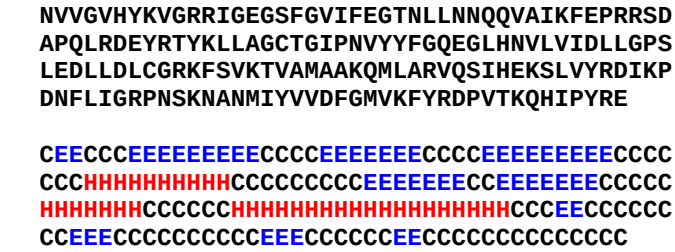

Fig.1 Una proteina: la struttura primaria e la SS.

2.1 I sistemi di predizione di seconda generazione

A partire dai tardi anni '80 il problema della classificazione della SS viene affrontato con strumenti di tipo connessionistico (reti neurali feed-forward) che, già dal primo lavoro di [8] mostrano prestazioni superiori ai metodi di tipo simbolico adottati fino ad allora. I metodi di classificazione che adottano reti neurali, detti di seconda generazione, si basano su una approssimazione di tipo locale. Si assume cioè che la struttura in posizione N possa essere predetta completamente in base al contenuto della sequenza nell'intervallo tra le posizioni (N-k) e (N+k). Le reti neurali utilizzate sono MLP e gli AA vengono passati alla rete con una codifica di tipo ortogonale. Con questo approccio, ed utilizzando come training set un database di proteine composto da circa 20000 AA, Qian e Sejnowski [8] ottengono una prestazione nella classificazione Q_3=62.7%.
Il successivo lavoro di [9] introduce una quantità considerevole di complicazioni strutturali nel design della rete neurale che deve classificare la struttura delle proteine. Queste variazioni possono comunque essere ricondotte alle due principali esigenze di ridurre il fenomeno dell'overfitting e di introdurre nella rete della conoscenza a priori relativa al problema. Con questi accorgimenti (per la descrizione si rimanda a [9]) Riis e Krogh ottengono un Q_3 pari al 66.3%, con un significativo guadagno rispetto al lavoro di Qian e Sejnowski.

2.2 L'informazione evoluzionistica e i sistemi di terza generazione

E' noto che la SS delle proteine tende ad essere molto meglio conservata della struttura primaria [10] e tutte le sequenze che mostrano una percentuale di AA uguali in posizione corrispondente superiore al 35% hanno SS molto simili[3]. Data la notevole disponibilità di sequenze questa conoscenza si presta ad essere sfruttata nella predizione della SS.
Data una sequenza X di cui si ignora la struttura, se è possibile trovare un insieme Ω di sequenze note che, allineate verso X mostrano con questa una identità percentuale di AA superiore ad una certa soglia [1], si può pensare che nell'insieme Ω di sequenze trovato vi sia maggiore informazione sulla struttura di X rispetto a quella che si trova nella sola sequenza degli AA di X. La pratica di generare l'insieme Ω è nota come allineamento di multiple sequenze omologhe (MA) e, a conferma della congettura che in Ω c'è più informazione che nella sola X, sull'MA si basa il salto di qualità fatto dai sistemi di classificazione della SS negli ultimi 5 anni [12].
Il sistema che ha inaugurato la terza generazione dei metodi di predizione introducendo, nel 1994, una modalità di utilizzo dell'MA è stato il PHD [11]. Nel sistema di Rost e Sander la sostanziale innovazione sta nell'utilizzo come ingresso per la rete neurale di un profilo di numeri reali che rappresentano le frequenze dei vari AA in una posizione dell'allineamento. Il sistema di predizione PHD ha dimostrato un livello di prestazioni intorno a Q_3=72% e nel 1996 è risultato il miglior predittore nella competizione CASP [5], una gara di predizione nella quale ai competitori vengono sottoposte alcune sequenze di cui ancora non è nota la struttura.
Un altro metodo per l'utilizzo dell'informazione contenuta negli MA è quello proposto da [9]. In questo caso le sequenze vengono allineate all'uscita del sistema di predizione combinando le predizioni che questo ha prodotto per ciascuna delle sequenze. Questo metodo porta ad una prestazione finale del sistema di predizione quasi identica a quella del PHD.

2.3 Problemi dell'approccio tradizionale

Il problema principale dell'approccio di [8], [11] e [9] alla classificazione della SS delle proteine sta in

[2] Per residuo si intende la posizione di una proteina corrispondente ad un AA.

[3] Mentre la maggioranza delle sequenze che hanno struttura simile ha meno del 15% di AA coincidenti, con un picco di similarità percentuale intorno all'8% nei cosiddetti omologhi remoti [10] ,[13].

quella che in 2.1 si è definita approssimazione locale. In effetti è noto (si veda, ad es. in [12] o [1]) che esiste dell'informazione che caratterizza strutture locali e che si trova a distanze anche considerevoli all'interno della sequenza rispetto al punto in cui è posta la struttura che ne dipende (dipendenze lontane). Per tentare di individuare l'entità di queste dipendenze si è così proceduto:

-Per ciascuna delle tre strutture (H,E e C) una rete neurale con due uscite (softmax) è stata addestrata a classificare la presenza/assenza della struttura nella generica posizione T, prendendo in ingresso la finestra di 3 AA centrata intorno alla posizione T+K.

-Per ogni K tra -80 e +80 le tre reti sono state addestrate sulle sequenze (483 addestramenti). Sia l'errore[4] complessivo della rete su tutto quanto il set:

$$E_l = \Sigma_n \Sigma_j t_j \log(y_j) \quad (2)$$

dove la prima sommatoria è estesa a tutti gli esempi, la seconda alle due uscite, t_j rappresenta il target j-esimo e y_j l'uscita j-esima della rete.

Se non si ha correlazione tra ingressi U e targets T della rete, cioè se:

$$P(T \mid U) = P(T) \quad (3)$$

il miglior fitting dei dati che si può ottenere è un modello che stima le probabilità incondizionate dell'appartenenza e non appartenenza ad una classe.

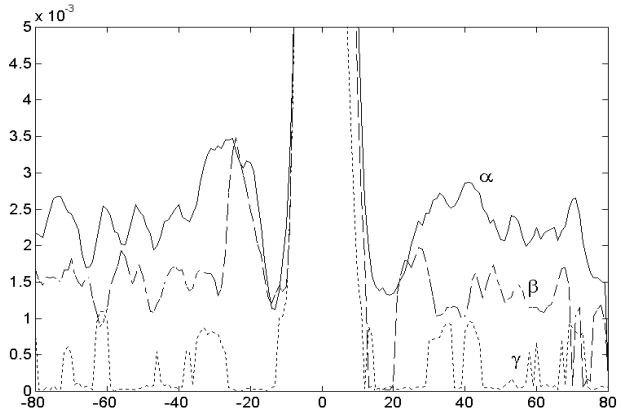

Fig.2 Inf(T-T_0).

L'informazione (2.3), al variare della distanza dalla struttura da predire. Le tre curve rappresentano le tre strutture, H (α), E (β) e C (γ).

Detti dunque n_y e n_n i numeri delle istanze in cui rispettivamente è presente e assente la classe che la rete deve predire, e posto $p = n_y/(n_y + n_n)$, la quantità:

$$E_n = -(n_y \log(p) + n_n \log(1-p)) \quad (4)$$

è l'errore se la rete si limita a predire uscite costanti e pari a (p,1-p), cioè se stima la probablità incondizionata dell'uscita.

Si è assunta come misura dell'informazione:

$$Inf = 1 - E_l / E_n \quad (5)$$

che risulta pari a 0 se non si ha informazione in ingresso e 1 in caso di perfetto fit dei targets.

L'andamento dell'informazione così definita è riportato nella Fig.2. Da questa si può dedurre come dipendenze tra struttura e sequenza possano esistere anche a distanze di decine di AA. L'approssimazione locale, che limita l'ingresso del sistema di predizione ad un intervallo di 13-15 AA elimina completamente queste dipendenze penalizzando le prestazioni.

3. Le BRNN (Bidirectional Recurrent Neural Networks)

Una **BRNN** (*bidirectional recurrent neural network*) è un modello **non causale** di traslazione **deterministica**, definito sullo spazio delle sequenze di lunghezza finita. Il modello, come nel caso delle RNN, descrive una traslazione di tipo deterministico:

$$\mathbf{Y} = \psi(\mathbf{U}) \quad (6)$$

in cui $\mathbf{U}=\{ U_1 ,U_2, .. ,U_T \}$ è una sequenza di ingresso di lunghezza T e $\mathbf{Y}=\{ Y_1 ,Y_2, .. ,Y_T \}$ è una sequenza di uscita di pari lunghezza.

Se F_t e B_t sono due vettori rispettivamente in R^n ed R^m, U_t è un vettore in R^k che codifica l'ingresso al tempo (posizione) t-esimo della sequenza e Y_t è un vettore in R^y che rappresenta l'uscita al tempo (posizione) t-esimo della sequenza, si può definire per una BRNN la seguente dinamica vettoriale:

$$F_t = \phi(F_{t-1} , U_t) \quad (7)$$

$$B_t = \beta(B_{t+1} , U_t) \quad (8)$$

$$Y_t = \eta(F_t , B_t , U_t) \quad (9)$$

in cui $\phi()$, $\beta()$ e $\eta()$ sono funzioni non-lineari vettoriali che vengono realizzate mediante MLP.

Le tre equazioni (7)-(9) definiscono una BRNN come mappa tra una sequenza di ingresso ed una di uscita di pari lunghezza, che raccoglie l'informazione contestuale proveniente dalle due direzioni della sequenza di ingresso attraverso le due catene di stati interni $\{ F_1, F_2, .. ,F_T \}$ e $\{ B_1, B_2, .. ,B_T \}$. Nel caso della prima catena il vincolo di causalità è rispettato, mentre il fatto che la dinamica descritta è non causale risulta evidente dalla eq. (8), in cui l'evoluzione dello stato B_t dipende esplicitamente dallo stato futuro B_{t+1}.

La rappresentazione della BRNN è completata dalla condizione al contorno per gli stati $F_0 = B_{T+1} = 0$. Una BRNN è detta *stazionaria* se le tre mappe $\phi()$, $\beta()$ e $\eta()$ sono indipendenti dalla variabile t. In tal caso la BRNN è individuata da tre soli MLP (si tratta di una forma di weight sharing attraverso t). In tutto questo lavoro si è fatta l'assunzione di stazionarietà per le BRNN.

[4] Come si vede ha la forma di una entropia mutua.

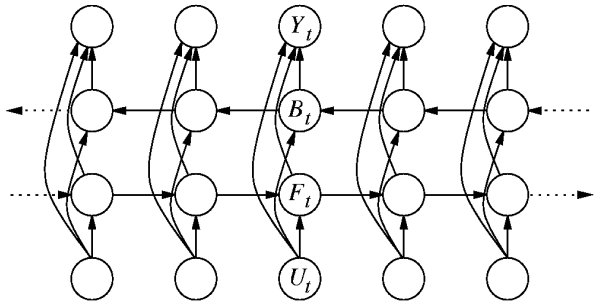

Fig.3 Modello grafico per la rete codificante di una BRNN, da [2].

3.1 Inferenza e apprendimento in una BRNN:

Come per il caso più generale dei modelli grafici per strutture dati [6] è opportuno, per descrivere in dettaglio i meccanismi di inferenza e di apprendimento delle BRNN, passare alla forma che il modello assume per una singola sequenza, nota come rete codificante (*encoding network*). In tal caso essenzialmente la rete ricorsiva (quindi la terna di MLP $\eta()$, $\phi()$ e $\beta()$) è uno stampo che viene replicato per ogni coppia ingresso-uscita (U_t , Y_t) della sequenza, ovvero esattamente T volte per una sequenza di lunghezza T. Un modello grafico della rete codificante è riportato in Fig.3. Per ogni t le tre connessioni entranti in Y_t sono riprodotte dalla $\eta()$, le due entranti in F_t e B_t rispettivamente da $\phi()$ e $\beta()$.

Inferenza: Partendo dallo stato F_0, vengono aggiornati tutti gli stati F_t fino al termine della catena forward (t=T), in base agli ingressi U_t ed alla mappa $\phi()$. Analogamente, a partire dallo stato B_{T+1} ed in base agli ingressi U_t ed alla mappa $\beta()$ vengono aggiornati tutti gli stati B_t della catena backward (fino a t=1). A questo punto, per ogni t della sequenza è disponibile la terna (U_t, F_t, B_t), quindi possono essere ricavate le uscite Y_t in base alla mappa $\eta()$.

Apprendimento: Si tratta di una applicazione dell'algoritmo di gradient-descent. La propagazione del gradiente avviene secondo l'algoritmo BPTS (BackPropagation Through Structure, [6]). Su tutte le uscite Y_t viene inserito il segnale di errore. Per la rete codificante di una BRNN si deve tener conto di dipendenze temporali di tipo non causale. Per propagare il segnale di errore in tutta la rete codificante è comunque sufficiente seguire un qualsiasi ordinamento topologico della rete. Ad esempio, dopo la backpropagation dell'errore attraverso le repliche della rete $\eta()$ per ogni t, si può propagare attraverso le repliche della rete $\beta()$ seguendo la catena non causale degli stati { B_1, B_2, .., B_T } per t crescente da 1 a T e attraverso le repliche della rete $\phi()$ lungo la catena causale { F_1, F_2, .. ,F_T } per t decrescente da T a 1. Si osservi che è necessario memorizzare per ogni t i segnali (vettori) di errore Fbp_t e Bbp_t provenienti dalle repliche di $\eta()$ e diretti verso le reti di transizione degli stati, per sommarli ai contributi di errore provenienti dalle catene, durante la backpropagation nelle reti di transizione $\phi()$ e $\beta()$. Come ultima osservazione, il fatto che si sia assunta la stazionarietà del modello, cioè che si utilizzino repliche della stessa rete per ciascuna delle tre mappe della BRNN significa che il gradiente complessivo per una rete è ottenuto sommando i contributi provenienti da tutti i passi temporali.

4. La sperimentazione

4.1 I dati

I dati utilizzati in questo lavoro sono stati estratti dal Brookhaven Protein Data Bank (PDB) [4]. Sono state eseguite due selezioni successive, la prima per eliminare quelle sequenze la cui struttura tridimensionale non fosse stata determinata con un alto livello di confidenza, la seconda per estrarre un subset rappresentativo eliminando le ridondanze (v [2] per dettagli). Nella fase finale dello studio si è fatto uso della release 79 (Luglio 1998) del PDB che ha portato ad un set di 824 sequenze, corrispondenti ad un totale di 184973 AA, una quantità di dati circa 7.5 volte maggiore che in [11] e [9].

4.2 I modelli di riferimento

Come esposto in 2.1e 2.2 i modelli che raggiungono le migliori prestazioni nel problema della classificazione della SS delle proteine sono, su singola sequenza quello proposto da [9] che ottiene un Q_3 pari al 66.3%, con l'ausilio della tecnica MA il sistema PHD di [11] con Q_3=72%. Date le differenze nell'ambiente di sperimentazione rispetto ai lavori precedenti si è scelto di riaddestrare alcuni tra i modelli di riferimento. Le prestazioni raggiunte dal miglior sistema per predizione su sequenza singola, del tutto simile a quello più efficiente di [9] è di poco inferiore al 69% (68.87%).

4.3 Le BRNN

Il modello di BRNN descritto nel cap.3 è quello che si è applicato nei test, ma la mappa $\eta()$ riportata in (9), nei modelli testati sulla classificazione della SS, è stata modificata come segue:

$$Y_t = \eta(F_{t-Cf} , .. , F_{t+Cf} , B_{t-Cb} , .. , B_{t+Cb} , U_t)$$

In tutti i test fatti si è comunque assunto Cf = Cb, quindi si userà per queste variabili l'unica denominazione di **Ct**.

Dalla Tab.1 si può dedurre come le migliori prestazioni delle BRNN siano circa uguali a quelle

del miglior modello di riferimento, nonostante in queste non si sia inserita informazione di tipo biologico.

Ct	NFB	NH	NH_T	Q_3 (SS)
2	7	11	8	68.72%
2	9	11	8	**68.79%**
3	8	11	9	68.75%
4	7	11	9	68.68%

Tab.1 Prestazioni di alcune BRNN. NFB è il numero di stati nascosti nelle catene. NH e NH_T sono il numero di unità nascoste della rete di uscita $\eta()$ e delle reti di transizione dello stato $\phi()$ e $\beta()$.

4.4 Giuria di 4 BRNN e 1 modello statico

Si è quindi composta una giuria di modelli tra le 4 migliori BRNN ed il modello di riferimento che mostrava la migliore prestazione. I risultati della giuria sono riportati in Tab.2 e si riferiscono ad una predizione ottenuta eseguendo una media non pesata delle singole predizioni. Dalla tabella si osserva che la combinazione dei 5 modelli ha una prestazione complessiva del 69.6%, con un guadagno dello 0.7-0.8% rispetto alle percentuali riportate da ciascuno dei modelli singoli.

4.5 Allineamenti multipli

La modalità che si è adottata in questo lavoro per introdurre l'informazione contenuta in un allineamento Ω è quella di predire la SS per ogni sequenza $Y_n \in \Omega$, quindi di eseguire una media pesata delle predizioni ottenute:

$$P_A(X) = w_0 P(X) + \Sigma n\, w_n P(Y_n) \qquad (10)$$

in cui $P_A()$ è la predizione con gli allineamenti, $P()$ è la predizione su singola sequenza e $\{w_n\}$ sono i pesi attribuiti alle singole predizioni. In particolare, gli allineamenti sono stati estratti dal database HSSP (v.[7]) e si sono utilizzati pesi uniformi cioè $w_n=1/(N+1)$, dove $N=|\Omega|$. Esistono, come noto, scelte più efficienti.

	MA	Single Seq.
MLP	72.61%	68.87%
BRNN1	72.58%	68.79%
BRNN2	72.74%	68.75%
BRNN3	72.64%	68.72%
BRNN4	72.62%	68.68%
Jury	**73.25%**	**69.49%**

Tab.2 Risultati di 4 BRNN e 1 modello statico + giuria tra i 5 modelli nel caso di predizione con MA e su sequenza singola.

In base alle specifiche dette si è applicato il metodo della predizione con MA alle 4 migliori BRNN, al miglior modello statico, quindi alla giuria dei 5 modelli. I risultati ottenuti sono riportati nella Tab.2. La crescita di prestazioni ottenuta coi MA è circa il 3.75%. Si ricordi che non si è fatto ricorso a tecniche più raffinate come ad es. un metodo di pesatura delle sequenze. Nonostante queste limitazioni l'insieme di predittori ottiene una prestazione finale del **73.25%**.

4.6 Seven-fold cross-validation

Per ottenere una misura più affidabile delle prestazioni dei sistemi descritti è stata adottata una procedura di 7-fold cross-validation[5]. I risultati, riportati in Tab.3, confermano quanto ottenuto in precedenza, con lievi guadagni. Si può notare che l'aggiunta di una rete neurale come filtro delle predizioni dei MA porta un guadagno dello 0.4%. Tutto il sistema classifica correttamente il **73.7%** delle SS, un dato superiore rispetto ai risultati dei migliori sistemi di predizione esistenti (72%)[6].
In Fig.4 è riportata la distribuzione delle sequenze in funzione del Q_3. Oltre il 70% delle sequenze viene predetto con un $Q_3 > 70\%$.

	Q_3	Q_H	Q_E	Q_C
MLP	69.00%	69.99%	47.62%	77.82%
BRNN1	68.90%	69.06%	48.46%	77.88%
BRNN2	68.85%	69.10%	47.82%	77.94%
BRNN3	68.68%	69.08%	47.61%	77.78%
BRNN4	68.92%	69.23%	48.37%	77.89%
Jury	**69.63%**	**70.44%**	**47.69%**	**79.05%**
MA	**73.32%**			
+Filter	**73.70%**			

Tab.3 Risultati della 7-fold cross-validation.

4.7 La competizione CASP '98

Nella competizione CASP (Critical Assessment of Protein Structure Prediction) un insieme di sequenze di AA per cui la SS è in corso di determinazione viene fornita alla comunità scientifica. Decine di gruppi di ricerca da tutto il mondo hanno partecipato con le proprie predizioni SS alle passate edizioni del '94, del '96 e del '98. I risultati dell'ultima edizione, basata su 35 sequenze, sono stati resi noti nel dicembre del 1998 [5]. Il criterio di valutazione della qualità di un sistema è stato la percentuale di corretta predizione media per sequenza (Q_{3seq}).
Per testare il nostro sistema sulle sequenze CASP 5

[5] Grazie ai responsabili del **CBS** di Lyngby (DK), in particolare a Søren Brunak, che hanno messo a disposizione le loro macchine per questi esperimenti.
[6] Si ricorda comunque che la comparazione tra prestazioni ottenute su dati differenti è sempre ardua.

modelli sono stati riaddestrati su tutti i dati. Il sistema completo ha ottenuto Q_3=**71.8%**, Q_{3seq} =**74.1%**. Il vincitore della competizione JONES-2 ha sottoposto le predizioni soltanto per 23 delle 35 sequenze scelte in base al criterio della massima confidenza, riportando un Q_{3seq}=77.6%. Se $y_{j,c}$ è l'uscita del modello relativa alla classe c-esima nella posizione j-esima di una proteina e R_{tot} è il numero di residui della sequenza, per confidenza C si intende :

$$C = \sum_j \max_c y_{j,c} \; / \; R_{tot} \qquad (11)$$

In base allo stesso criterio di scelta si sono selezionate 23 sequenze ottenendo un Q_{3seq}=**77.5%**.
Inoltre in JONES-2 è stato utilizzato il nuovo database TrEMBL per la generazione degli allineamenti, molto più consistente rispetto all'HSSP [7] in base al quale sono stati generati i nostri.

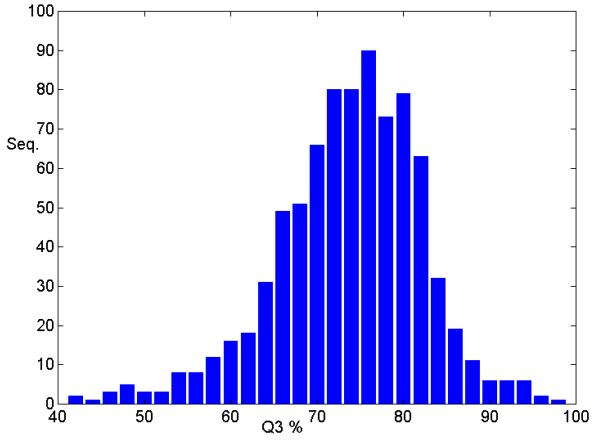

Fig.4 Risultati della 7-fold cross-validation. Numero di sequenze predette in funzione del Q_3.

5. Conclusioni

Si sono applicate al problema della classificazione della SS delle proteine le BRNN, un nuovo modello adattivo non causale per la processazione delle sequenze. Gli addestramenti sono stati condotti su un insieme di proteine non omologhe circa 7.5 volte più esteso rispetto a quelli riportati in letteratura ([11], [9]). Dai test le BRNN ottengono una percentuale di corretta classificazione misurata con 7-fold cross-validation leggermente inferiore al 69%, cioè significativamente migliore rispetto a quella di una qualsiasi rete neurale feed-forward e circa uguale al miglior modello di riferimento, contenente una rilevante iniezione di conoscenza strutturale. Con l'ausilio delle tecniche degli ensembles e degli MA si ottiene una prestazione del 73.7%, calcolata con 7-fold cross-validation. Le prestazioni sono state quindi comparate col sistema di predizione che ha vinto la terza edizione della competizione CASP, risultando praticamente identiche.
Si ha ragione di pensare che l'utilizzo di tecniche più evolute e di un maggiore quantitativo di dati per la gestione della tecnica MA lasci spazio ad ulteriori miglioramenti delle prestazioni.

6. Bibliografia

[1] P.Baldi, S.Brunak."Bioinformatics: The Machine Learning Approach", MIT Press, Cambridge, MA, 1998

[2] P.Baldi, S.Brunak, P.Frasconi, G.Pollastri, G.Soda."Bidirectional Dynamics for Protein Secondary Structure Prediction", IJCAI99 Workshop on "Neural, Symbolic and Reinforcement Methods for Sequence Learning" (accepted paper), Stockholm 1999.

[3] Y.Bengio, P.Frasconi. "Input/Output HMMs for Sequence Processing", IEEE Trans. on Neural Networks, 7(5), pp 1231-1249, 1996

[4] F.C.Bernstein, T.F.Koetzle, G.J.B.Williams, E.F.Meyer, M.D.Brice et al. "The Protein Data Bank: a computer based archival file for macromolecular structures.", J. Mol. Biol., 112, pp 535-542, 1977

[5] CASP3, http://PredictionCenter.llnl.gov/casp3/

[6] P.Frasconi, M.Gori, A.Sperduti. "A General Framework for Adaptive Processing of Data Structures" IEEE Trans. on Neural Networks, 9, 5 pp 768-786, 1998

[7] C.Sander, R.Schneider. Database of homology-derived secondary structure of proteins
http://www.sander.embl-heidelberg.de/hssp/

[8] N. Qian, T.J. Sejnowski. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models", J. Mol. Biol. 202, pp 865-884, 1988

[9] S.K.Riis, A.Krogh. "Improving Prediction of Protein Secondary Structure using Structured Neural Networks and Multiple Sequence Alignments", J.Comput.Biol., 3, pp 163-183, 1996

[10] B.Rost. "PHD: predicting 1D protein structure by profile based neural networks", Meth. in Enzym., 266, pp 525-539, 1996 http:// www.embl-heidelberg.de/ ~rost/ Papers/ MethEnz96.html

[11] B.Rost, C.Sander. "PHD - An automatic mail server for protein secondary structure prediction", Comput. Appl. Biosci., 10(1), pp 53-60, 1994

[12] B.Rost, C.Sander."3rd Generation Prediction Of Secondary Structure",Running title: "Prediction of Secondary Structure", in: Webster D. M. (ed.): 'Predicting protein structure'. Humana Press, 1998, in press. http:// www.embl-heidelberg.de/ ~rost/ Papers/ 98revSecStr.html

[13] B.Rost, R.Schneider. "Pedestrian guide to analysing sequence databases", in: Ashman K. (ed.): 'Core techniques in Biochemistry'. Heidelberg: Springer, 1997, in press. http:// www.embl-heidelberg.de/ ~rost/ Papers/ Springer96.html