Università degli Studi di Padova

Dipartimento di Fisica e Astronomia
Laurea Triennale in Fisica

Predizione di propensitá ad aggregare di proteine a partire da strutture native

Autore:
Marin Michele,
1029107

Relatore:
Dott. Antonio Trovato

Anno accademico 2013–2014

Indice

0.1 Sommario

Diverse malattie degenerative sono legate all'aggregazione patologica di proteine in strutture fibrillari insolubili, note come fibrille amiloidi, caratterizzate dalla preseza di filamenti β perpendicolari all'asse della fibrilla. La struttura amiloide é a grandi linee comune a diverse proteine, ed é di grande interesse biomedico poter predire quanto una sequenza tenda a formare fibrille amiloidi e, nel caso, quale parte della sequenza ne stabilizzi la struttura a filamenti β. Nel gruppo di ricerca in cui ho svolto il lavoro di tesi é giá stato sviluppato un algoritmo, PASTA, per predire la propensitá ad aggregare di sequenze proteiche, basato su energie di interazione di coppie di amminoacidi coinvolti in una struttura a filamenti β.

L'algoritmo si basa sull'ipotesi che le proteine che aggregano siano caratterizzate da un elevato grado di disordine strutturale nello stato solubile (proteine nativamente non strutturate) a partire dal quale si sviluppa il processo di aggregazione. Questa condizione é soddisfatta per diverse proteine coinvolte in malattie neurodegenerative, dal peptide Aβ del morbo di Alzheimer all'α-sinucleina del morbo di Parkinson. In questo lavoro di tesi ci si propone di migliorare l'algoritmo PASTA, introducendo un termine che valuti l'energia della proteina nella struttura nativa nella maniera piú semplice possibile. A tale scopo, si propone di utilizzare un termine energetico che non dipenda dalla conoscenza della struttura nativa, ma solamente dalla composizione amminoacidica della sequenza in esame.

Approcci simili, basati su forme quadratiche nella composizione in sequenza, sono giá presenti in letteratura. La performance dell'algoritmo PASTA cosí modificato verrá valutata su di un elenco di corte catene polipeptidiche, per scoprire se possa migliorare le predizioni sulla capacitá di aggregazione di tali catene. Si valuterá poi l'effetto del nuovo termine di potenziale sui profili energetici dep peptide Aβ-40 e del prione di fungo HETs, per i quali é possibile un confronto con dati strutturali sperimentali ad alta risoluzione.

Si mostrerá infine come l'energia libera media per residuo predetta da PASTA correli con la tossicitá misurata in vivo per una serie di mutazioni dell'Aβ-42. Per fare questo ci si riferirá ad uno studio precedente, in cui la tossicitá era stata valutata su mosche della frutta ingegnerizzate in modo da esprimere tali peptidi mutati.

Capitolo 1

Introduzione

1.1 Le proteine

Le proteine sono macromolecole biologiche che svolgono un'ampia gamma di funzioni all'interno della cellula e sono formate da una o piú catene amminoacidiche. Gli amminoacidi standard, gli unici che verranno considerati in questo lavoro, sono venti. Essi legano tra loro mediante legami covalenti detti legami peptidici e sono caratterizzati da gruppi laterali, detti gruppi R, che vengono classificati in base alle loro proprietá chimiche come acidi, basici, idrofobici e idrofili. Le proteine si distinguono tra loro grazie alle diverse combinazioni, anche molto lunghe, in cui tali amminoacidi possono disporsi. Il livello piú semplice di conoscenza di una proteina é la conoscenza dell'ordine in cui gli amminoacidi sono disposti, ovvero la sua struttura primaria. Tra i gruppi CO e NH di amminoacidi spazialmente vicini si formano poi dei legami idrogeno, la cui disposizione puó portare a diverse configurazioni, che definiscono la struttra secondaria. I principali tipi di struttura secondaria sono le strutture ad α-elica, a foglietto β e a ripiegamento β.

- La struttura ad α-elica si forma quando diversi residui amminoacidici consecutivi formano angoli di legame compresi tra -60° e -45°. Lo scheletro della catena si avvolge attorno a un asse centrale, mentre i gruppi R sporgono radialmente verso l'esterno. Normalmente un giro d'elica é costituito da 3.6 amminoacidi, che coprono una distanza di 5,4 Å, e la struttura é particolarmente stabile grazie ai legami idrogeno che si instaurano all'interno dell'elica. La lunghezza media delle α-eliche é normalmente di circa 10 residui, ed esse sono di solito destrorse.

- La seconda struttura piú diffusa é il foglietto β, che consiste in due o piú filamenti β disposti uno accanto all'altro e connessi da tre o piú legami idrogeno. Un filamento β é semplicemente una catena che si dispone linearmente ed é in grado di formare legami idrogeno con altri filamenti. Un foglietto beta presenta un caratteristico andamento a zig-zag che porta i gruppi R a sporgere alternativamente verso i lati. Considerando come verso positivo della catena quello in cui in ogni amminoacido il gruppo amminico precede il gruppo carbossilico, il legame é detto parallelo se le due catene a contatto sono nello stesso verso, antiparallelo altrimenti.

- Alcuni residui sono coinvolti in ripiegamenti a gomito che modificano la direzione della catena polipeptidica. Data l'abbondanza di tali ripiegamenti essi sono classificati come terzo tipo di struttura secondaria. Esistono diversi tipi di ripiegamenti β, i piú comuni dei quali sono costituiti da quattro residui che legano due segmenti β antiparalleli per formare un'ansa a forcina.

In solvente acquoso, come per esempio quello all'interno di una cellula, spesso la struttura secondaria si ripiega in una ben definita struttura tridimensionale globulare e compatta, detta stato nativo. I tipi di legame che caratterizzano questo ripiegamento sono solitamente legami ionici, interazioni idrofobiche, legami a idrogeno e legami disolfurici. É proprio la struttura

tridimensionale della proteina a determinarne il ruolo nell'organismo. É importante notare, per gli scopi di questo lavoro, che gli amminoacidi idrofobici tendono a mantenersi verso il cuore della struttura globulare.

Le macromolecole cosí formate si riuniscono infine in strutture ancora piú complicate, dette strutture quaternarie. Un esempio é l'emoglobina, che é formata da quattro sub-unitá, in cui le proteine si accoppiano dapprima in due dimeri, e infine in un tetramero. La struttura nativa é associata alle funzionalitá bilologiche della proteina ed é univocamente determinata dalla sequenza di amminoacidi. Come dimostrato dal biochimico americano Anfinsen[13], inoltre, essa é cosí stabile che le proteine tendono a riformarla anche in vitro, persino dopo essere state denaturate. Vi sono tuttavia eccezioni a questa regola.

Alcune zone delle proteine sono infatti spesso disordinate e non strutturate, e ci sono anche proteine che svolgono funzioni attive all'interno dell'organismo che hanno una struttura intrinsecamente non ordinata. É stato infatti evidenziato, sopratutto negli ultimi anni, come circa un terzo delle proteine presenti nell'uomo sia composto da proteine parzialmente o completamente disordinate, il che indica chiaramente come tali segmenti di catena abbiano un ruolo importante nel corretto funzionamento dell'organismo. Esse sono solitamente collegate a funzioni di segnalazione e regolazione cellulare e la loro attivitá é regolata da modificazioni sucessive alla traduzione. Queste proteine vengono dette anche 'flessibili', dato che spesso sono in grado di adattarsi alle superfici di differenti partner molecolari.

A bassa concentrazione queste proteine sono normalmente solubili, ma esperimenti sia in vivo che in vitro hanno dimostrato che, in alcuni casi, aumentando la concentrazione di proteine in soluzione si rende sempre piú probabile la deposizione di strutture non solubili. Queste strutture derivano appunto dall'aggregazione di piú catene proteiche.

1.2 Le fibrille amiloidi

Molte malattie derivano dal fallimento di alcune proteine a mantenere il loro stato nativo originale, eventualmente disordinato, e di svolgere quindi la giusta funzione all'interno del corpo. Una delle maggiori cause di questo fallimento, e quella di cui ci occuperemo, é l'abbandono dello stato solubile per passare a uno stato fibrillare organizzato e insolubile. Esempi importanti di questo tipo di malattie sono il morbo di Alzheimer, in cui le fibrille del peptide Aβ si accumulano nelle placche senili, e il morbo di Parkinson.

Si noti che la formazione di fibrille amiloidi non é sempre associata a patologie, in quanto esistono alcuni esempi in cui gli organismi sfruttano questo tipo di strutture per svolgere compiti specifici. Un esempio sono le fibrille usate dall'*Escherichia coli* per colonizzare superfici inerti e per mediare i contatti con le proteine. Un'altro esempio sono hypae dello *Streptomyces coelicolor*, che permettono alle sue spore di diffondersi piú efficacemente. É interessante notare come talvolta le fibrille amiloidi possano fungere da materiale trasmissibile per via ereditaria pur senza essere codificate nel DNA[1].

Analisi effettuate su fibrille Ex vivo estratte dai pazienti e su fibrille prodotte in vitro hanno mostrato la struttura di tali fibrille. Esse sono tipicamente formate da 2 a 6 protofilamenti, ciascuno di circa 2-5 nm di diametro, che si incastrano tra loro per formare fibrille solitamente a forma di *corde* o *nastri*. Le corde hanno una larghezza da 7 a 13 nm, mentre i nastri sono spessi da 2 a 5 nm e possono arrivare ad essere larghi anche 30 nm.

Solo negli ultimi anni, grazie a tecniche quali la diffrazione a raggi X e la solid-state NMR, si é riusciti ad indagare piú nel dettaglio la struttura di queste fibrille. Attraverso lo studio di vari casi si sono potuti notare una serie di elementi in comune, tra cui una quasi sempre presente struttura a croce-β, caratterizzata da filamenti beta ortogonali e legami idrogeno paralleli all'asse della fibrilla (vedi rappresentazione pittorica in Figura 1.1. in basso a destra). La struttura a croce-β consiste in un doppio foglietto β, con entrambi i foglietti formati molto spesso, ma non sempre, da filamenti β paralleli in registro. Con in registro si intende che ogni residuo é legato all'analogo residuo corrispondente di un'altra catena. Le catene laterali che sporgono dai due foglietti β formano una cerniera impermeabile, che lega i foglietti tra loro. La presenza cosí diffusa di questo

tipo di struttura supporta l'interpretazione secondo cui sono le proprietá fisico-chimiche delle catene polipeptidiche a determinare la struttura delle fibrille. Un'altra caratteristica comune delle fibrille amiloidi, che risulta importante in quanto ne permette l'individuazione e ne facilita lo studio, é il fatto che esse legano ad alcuni coloranti specifici, cioé il Congo-Red e la Tioflavina T.

All'interno delle strutture amiloidi ci sono peró anche alcune differenze, che si manifestano non solo tra catene diverse, ma anche in conformazioni diverse possibili per la stessa catena. Sebbene la fase fibrillare insolubile di tali proteine sia molto stabile, infatti, le condizioni in cui avviene la loro formazione puó portare a risultati diversi tra loro. Si parla quindi di polimorfismo conformazionale, che é stato riscontrato in diversi esempi e che consiste solitamente in variazioni sul tema comune delle strutture a croce β.

Il processo di aggregazione di proteine nativamente disordinate in fibrille amiloidi é piuttosto complesso ed é solitamente preceduto dall'aggregarsi di nuclei critici, da cui la crescita della fibrilla prosegue piú rapidamente. I nuclei sono a loro volta costituiti a partire da proto-fibrille, che sembra si formino grazie alla riorganizzazione e all'assemblamento di oligomeri relativamente disorganizzati e non strutturati.

Le fibrille si possono peró formare anche a partire da proteine globulari, nonostante in generale si creda che debba prima avvenire un processo di totale o parziale perdita della struttura tridimensionale. É importante notare che spesso basta una piccola percentuale di materiale fibrillare in equilibrio con la proteina nel suo stato nativo per cominciare una reazione a catena che porta alla formazione delle fibrille amiloidi e alla scomparsa delle proteine nello stato nativo. Il cambiamento delle condizioni della soluzione, comunque, puó portare all'inversione di questo processo, e in generale al passaggio conformazionale della proteina in uno dei molti stati che puó assumere.

Un'analisi effettuata su un grande numero di proteine non strutturate nel loro stato nativo ha evidenziato come l'evoluzione ha in molti casi cercato di contrastare il formarsi delle fibrille amiloidi. Dato che l'aggregazione é molto influenzata da fattori quali l'idrofobicitá e la carica delle catene, sono state condotte delle ricerche per scoprire se la natura avesse favorito o meno la propensione ad aggregare. É stato trovato che le catene tendono a non avere piú di tre o quattro residui idrofobici consecutivi e che schemi di residui alternativamente idrofobici e idrofilici, che favoriscono la formazione di foglietti β, sono meno frequenti di quanto ci si aspetterebbe da una distribuzione casuale. Solitamente le zone piú significative per l'aggregazione sono costituite da pochi residui, spesso non legati nello stato nativo. Poter predire la propensitá ad aggregare di una data proteina e le zone della catena in cui avviene il processo di aggregazione é un problema di grande interesse in campo biomedico. Nel gruppo di ricerca in cui ho svolto il lavoro di tesi é giá da tempo stato sviluppato un algoritmo che permette di calcolare la propensione all'aggregazione di catene polipeptidiche. Scopo della tesi é quello di valutare gli effetti dell'aggiunta di un termine di potenziale che possa tenere conto del contributo entalpico della struttura originaria della catena nello stato nativo solubile, anche disordinato. Ci si aspetta che questo contributo entalpico sia dovuto a un certo grado di struttura residua, che é sperimentalmente presente per proteine nativamente non strutturate.

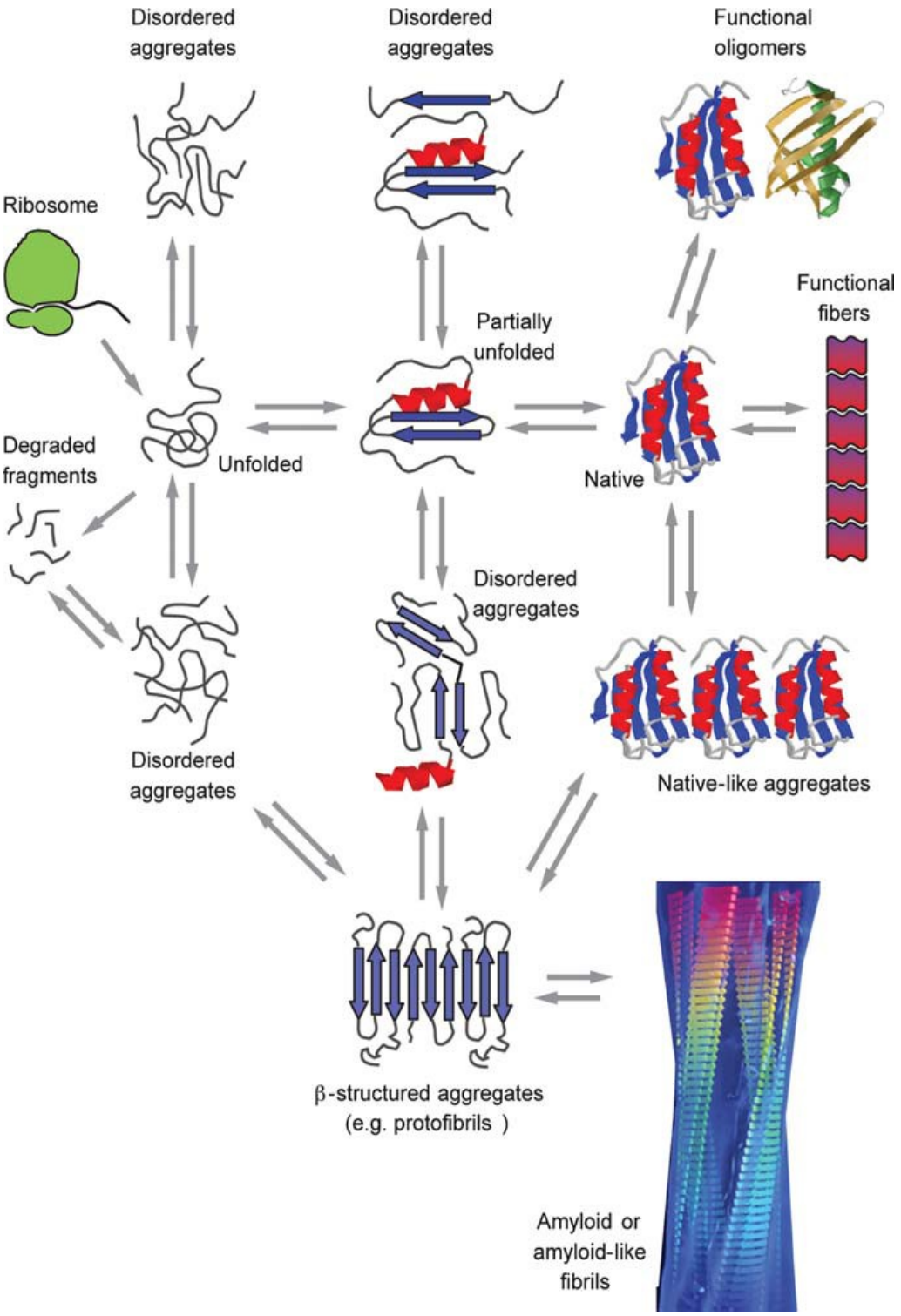

Figura 1.1: Rappresentazione schematica di alcuni degli stati conformazionali che le catene polipeptidiche possono assumere e delle relazioni che possono intercorrere tra loro

Capitolo 2

Predizione di propensitá ad aggregare: formalismo teorico

2.1 Struttura dell'algoritmo PASTA

L'algoritmo di base, PASTA, '*Prediction of Amyloid STructure Agregation*', si basa sull'ipotesi che le proteine che aggregano siano caratterizzate da un elevato stato di disordine nello stato solubile.

Si vuole cercare una funzione del tipo $F = H - TS$, dove H é l'entalpia, T la temperatura, S l'entropia e F l'energia libera. Ció che a noi interessa, e che PASTA calcola, é la differenza di energia libera, definita come

$$\Delta F = \Delta H - T\Delta S \tag{2.1.1}$$

Si puó immaginare la differenza di energia libera come differenza tra l'energia libera nello stato fibrillare e quella nello stato solubile, diciamo

$$\Delta F = F_f - F_s \tag{2.1.2}$$

Un profilo energetico in grado di rappresentare la situazione in maniera pittorica é quello in figura, dove S é il minimo di energia libera riferito allo stato solubile e F quello riferito allo stato fibrillare.

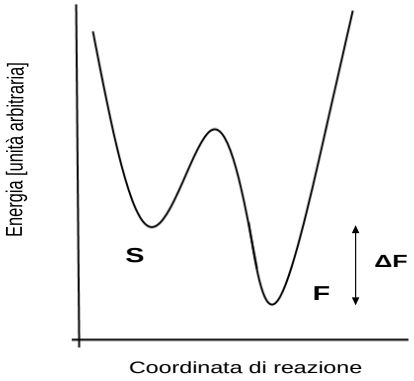

Figura 2.1: Esempio di un plausibile profilo energetico per una proteina che aggrega nello stato fibrillare

Come vedremo piú in dettaglio in seguito l'obiettivo dell'algoritmo é valutare ΔF per una generica struttura aggregata definita da uno specifico accoppiamento β fra due segmenti proteici di lunghezza L. Si arriva quindi all'espressione della differenza di energia libera

$$\Delta F = H_f - H_s + TS_f - TS_s \tag{2.1.3}$$

La differenza di entropia vuole rappresentare la perdita di entropia dovuta al maggior ordine della nuova fase amiloide strutturata. In PASTA originale si é posto $S_f = 0$. L'entropia dello stato solubile é invece calcolata supponendo il cambiamento di entropia lineare con la lunghezza dei segmenti coinvolti nell'accoppiamento, cioé supponendo che ogni contatto tra residui sottragga la stessa quantitá di entropia al sistema. La temperatura é supposta costante, nel seguito si userá $k_s T = 1$, utilizzando quindi unitá di misura adimensionali. Si ha quindi

$$\Delta S = -L\Delta s \tag{2.1.4}$$

dove $\Delta s = 0.2$ é stato determinato empiricamente nella prima elaborazione dell'algoritmo[2].

L'entalpia H_s é posta uguale a zero, assumendo lo stato solubile comletamente disordinato e privo di interazioni tra amminoacidi della catena, mentre per calcolare l'entalpia H_f é stato elaborato un metodo piú complesso. Dato che la conoscenza della struttura secondaria richiede grande potenza computazionale e contiene molte piú informazioni rispetto alla sola struttura primaria, si é deciso di sviluppare un algoritmo che tenga conto solo della struttura primaria. Questo porta chiaramente a una notevole dose di approssimazione, ma consente di calcolare l'entalpia in modo molto semplice e permette di evidenziare come alcuni aspetti del processo di aggregazione delle proteine in fibrille amiloidi dipendano in effetti fortemente dalla struttura primaria.

A ogni coppia di residui viene associato un potenziale, ricavato dalle formule

$$E^a_{ab} = -\log\left(\frac{\frac{n^a_{ab}}{n_{ab}}}{\frac{\sum_{ab} n^a_{ab}}{\sum_{ab} n_{ab}}}\right), \qquad E^p_{ab} = -\log\left(\frac{\frac{n^p_{ab}}{n_{ab}}}{\frac{\sum_{ab} n^p_{ab}}{\sum_{ab} n_{ab}}}\right) \tag{2.1.5}$$

dove n_{ab} é il numero di coppie di residui a contatto dentro al foglietto β e n^a_{ab} é il numero di coppie di residui dentro un foglietto β antiparallelo (parallelo). In questa maniera le coppie di residui trovate piú frequentemente in contatto fra loro in foglietti β avranno una energia di interazione favorevole (negativa). Si é deciso di definire a contatto quelle coppie di residui che sono stati riconosciuti formare ponti β dall'algoritmo DSSP[12]. Con legame parallelo si intende un legame che, date due sequenze e dati due residui, il primo in posizione i-esima nella prima catena e il secondo in posizione j-esima nella seconda catena, leghi al residuo i+1 il residuo j+1. Il legame antiparallelo lega invece il residuo i+1 al residuo j-1. Un esempio grafico degli accoppiamenti parallelo e antiparallelo é mostrato in figura.

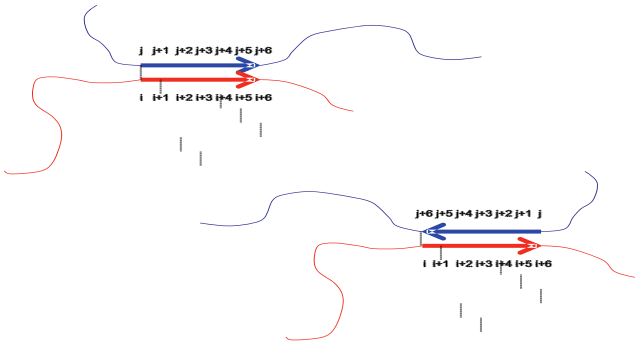

Figura 2.2: Schema della struttura dei legami parallelo e antiparallelo

Come giá anticipato, il verso di numerazione della catena proteica é fissato per convenzione dall'N al C terminale. Ipotizzeremo la prima e la seconda catena essere uguali, condizione quasi sempre verificata per le catene che compongono le fibrille amiloidi. In questo lavoro ci occuperemo solo di questo caso.

I contatti sono conteggiati a partire da quelli presenti nel database top500H[14]. Tale database é un set di 500 strutture globulari ricavate in alta risoluzione da esperimenti di diffrazione ai raggi-X su cristalli. Caratteristica di questo database é quella di essere stato raffinato in modo da non essere ridondante, per evitare di introdurre nelle analisi dei bias dovuti alla presenza di sequenze troppo simili fra loro. Si noti che si ipotizza che i legami idrogeno, protagonisti dei legami considerati da PASTA, siano simili tra residui analoghi per lo stato fibrillare e per quello globulare.

Gli accoppiamenti sono definiti immaginando di mettere a contatto due segmenti di catena, a partire dal residuo i nella prima catena e dal residuo j nella seconda, per poi proseguire considerando come adiacenti i residui i+1 e j+1 in regime parallelo e i+1 e j-1 in regime antiparallelo. I residui vengono aggiunti fino a che i segmenti a contatto non contano L residui.

Le energie di accoppiamento $\varepsilon^p_{i,j}(L)$ e $\varepsilon^a_{i,j}(L)$, dove 'p' e 'a' stanno ancora per parallelo e antiparallelo, sono definite come somma del potenziale calcolato e del termine di entropia. Questo porta, nel caso di accoppiamenti paralleli (la formula per il caso antiparallelo non verrá mostrata), a

$$\varepsilon^p_{i,j}(L) = \sum_{a<L} E^p_{i+a,j+a} - L\Delta s \tag{2.1.6}$$

dove L é la lunghezza del frammento considerato. Ci sará comodo indicare $E_1 = \sum_{a<L} E^p_{i+a,j+a}$. Con $E^p_{i+a,j+a}$ si vuole indicare il potenziale tra il residuo nella posizione i+a nella prima catena e quello in posizione j+a nella seconda catena.

Si procede quindi a definire una funzione di partizione Z come somma su tutti i possibili $L > 3$ e $L \leq Lmax$, ossia

$$Z = \sum_{i,j,L>3} L\{exp(\zeta\varepsilon^p_{i,j}) + exp(\zeta\varepsilon^a_{i,j})\} \tag{2.1.7}$$

La condizione $L > 3$ rappresenta il fatto che non si é osservata nessuna struttura a croce β con filamenti piú corti di 4 residui. $\zeta = 2$ é un fattore adimensionale che fissa la scala di energia in modo che una unitá di misura 'PASTA' sia equivalente a $2K_bT$.

La funzione di partizione ridotta, che considera solo i termini di energia che contengono un particolare residuo, diciamo k, é definita come

$$z(k) = \sum_{i,j,L>3} \delta_{i\leq k<i+L} L[exp(\zeta\varepsilon^p_{i,j}) + exp(\zeta\varepsilon^a_{i,j})] \tag{2.1.8}$$

Da queste quantitá sono infine calcolati i valori che effettivamente vengono restituiti dal programma, cioé

- Il profilo di energia libera di aggregazione $g(k) = ln(Z(k))/\zeta$
- Il profilo di probabilitá di aggregazione $h(k) = \frac{Z(k)}{Z}$

Sono calcolate contemporaneamente anche le quantitá bidimensionali associate alle coppie in cui compaiono solo i residui k nella prima catena e m nella seconda catena. La funzione di partizione ridotta é in questo caso definita come

$$z(k,m) = \sum_{i,j,L>3} \delta_{i\leq k<i+L}\delta_{j\leq m<j+L} L[exp(\delta_{k-m+j-i}\zeta\varepsilon^p_{i,j}) + exp(\delta_{k+m+1-L-j-i}\zeta\varepsilon^a_{i,j})] \tag{2.1.9}$$

Gli accoppiamenti con energia piú bassa determinano i picchi dei profili di probabilitá e i minimi dei profili di energia libera. PASTA si é dimostrato in grado di predire con grande precisione le corrette zone di aggregazione per diverse proteine che formano fibrille amiloidi.

2.2 Contributo entalpico dello stato solubile

L'idea principale di questa tesi é che la forza dei legami nello stato solubile influenzi la propensitá ad aggregare delle catene polipeptidiche. Una catena con uno stato globulare molto stabile, infatti, dovrebbe avere minore probabilitá di passare allo stato fibrillare. Si é giá visto come uno stato pre-fibrillare rivesta una grande importanza per la formazione delle fibrille, che tendono a legarsi partendo spesso dalle zone non strutturate della catena. Si é quindi cercato di dare una stima dell'entalpia delle proteine, per allargare la capacitá predittiva di PASTA anche a quelle proteine meno disordinate nel loro stato nativo.

Utilizzeremo qui quanto riportato nell'articolo pubblicato su JMB da Istvan Simon et al., originariamente introdotto per affrontare il problema della predizione del grado di disordine delle proteine [3].

In questo approccio si é cercato prima di tutto di stimare l'entalpia della proteina nello stato nativo a partire dalla composizione amminoacidica per un database di strutture native note di proteine globulari. L'entalpia totale é stata calcolata tenendo in conto tutti i contatti e pesandoli secondo la loro entalpia di interazione. L'entalpia é stata fatta dipendere solo dalle diverse coppie di amminoacidi in contatto tra loro, metodo che risulta in una matrice 20 x 20 simile a quella utilizzata per PASTA, che chiameremo **M**. Si ha quindi

$$H = \sum_{ij=1}^{20} M_{ij} C_{ij} \tag{2.2.1}$$

dove M_{ij} é l'entalpia di interazione tra un amminoacido di tipo i e uno di tipo j, mentre C_{ij} é il numero di coppie di residui i,j in contatto nella conformazione data. Si noti che in questo caso si considera una definizione semplice di contatto basata su una soglia di distanza di 6.5 Å fra i carboni beta dei gruppi R dei residui corrispondenti

L'etalpia per residuo é stata quindi approssimata con $\frac{H}{N}$, dove N é il numero di residui della proteina. Si é voluto rappresentare il fatto che l'entalpia di un residuo non dipende solo dal tipo di amminoacido, ma anche dai potenziali partner nella sequenza. Il metodo usato é il metodo piú semplice, é stata cioé introdotta una forma quadratica nella composizione amminoacidica:

$$\frac{H_{stimata}}{N} = \sum_{ij}^{20} n_i P_{ij} n_j \tag{2.2.2}$$

dove n_i é la frequenza di amminoacidi di tipo i nella sequenza e **P** é la matrice di predizione che verrá poi utilizzata per stimare l'entalpia basandosi solo sulle sequenze. Per calcolarla é stato eseguito un fit ai minimi quadrati. Prima di tutto l'entalpia totale di ogni proteina 'k' del database é stata scomposta nel contributo specifico degli amminoacidi, cioé $H^k = \sum_i e_i^k$. Gli e_i^k sono quindi stati associati alla formula quadratica. Le espressioni per $e_i^k(calcolato)$ e per $e_i^k(stimato)$ diventano quindi

$$e_i^k(calcolato) = \sum_{j=1}^{20} M_{ij} C_{ij}^k, \qquad e_i^k(stimato) = N_i^k \sum_{j=1}^{20} P_{ij} ni^k \tag{2.2.3}$$

Le righe della matrice **P** sono quindi ottenute minimizzando la funzione

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_i^k)^2 \tag{2.2.4}$$

cioé ponendo $\partial Z_i / \partial P_{ij} = 0$ per tutti i P_{ij}. I risultati sono poi stati testati su 674 proteine dalla struttura nota prese dal database Glob-list. In queste verifiche il coefficiente di correlazione lineare tra energie calcolate e energie stimate é risultato di 0.76.

Nel nostro lavoro abbiamo quindi utilizzato la matrice **P** per tentare di stimare l'entalpia H_s delle proteine nello stato solubile.

La nuova formula per le energie di accoppiamento di due segmenti di lunghezza L (anche ora riportata solo per il caso parallelo) diventa quindi

$$\varepsilon^p_{i,j}(L) = \sum_{a<L} E^p_{i+a,j+a} - L\Delta s + \lambda \sum_{a<L} \sum_{b<L} P_{i+a,j+b} \tag{2.2.5}$$

dove $P_{i+a,j+b}$ é il nuovo termine di potenziale preso dalla matrice $\mathbf{P}$. Chiamiamo ora H_2 questo secondo termine di potenziale, ossia $H_2 = \lambda \sum_{a<L} \sum_{b<L} P_{i+a,j+b}$. Un $\lambda < 0$ é quanto ci aspettiamo da quanto detto sopra, e significherebbe che il nuovo termine predice correttamente il termine di entalpia H_s cioé $H_2 = -H_s$. Se λ fosse invece positivo starebbe a significare che i legami tra residui anche lontani nella catena hanno effettivamente un ruolo nella aggregazione. In questo caso avremmo $H_s = 0$ e $H_f = H_1 + H_2$, cioé anche il termine quadratico contribuirebbe all'entalpia della proteina nel suo stato fibrillare. É anche possibile che una situazione di questo tipo rifletta un comportamento di tipo cooperativo, in cui la formazione di un legame influenza anche le energie dei legami giá formati tra i residui vicini.

Qui sotto é riportata una visione grafica di come agisce il nuovo termine di entalpia H_2. Nella parte 1 sono rappresentati i legami considerati in H_1. Nella parte 2 sono rappresentati i legami aggiunti da H_2 (in verde) solo per quanto riguarda il primo peptide della prima catena, mentre nella terza parte della figura sono rappresentati (sempre in verde) tutti i legami aggiunti dal nuovo termine di potenziale.

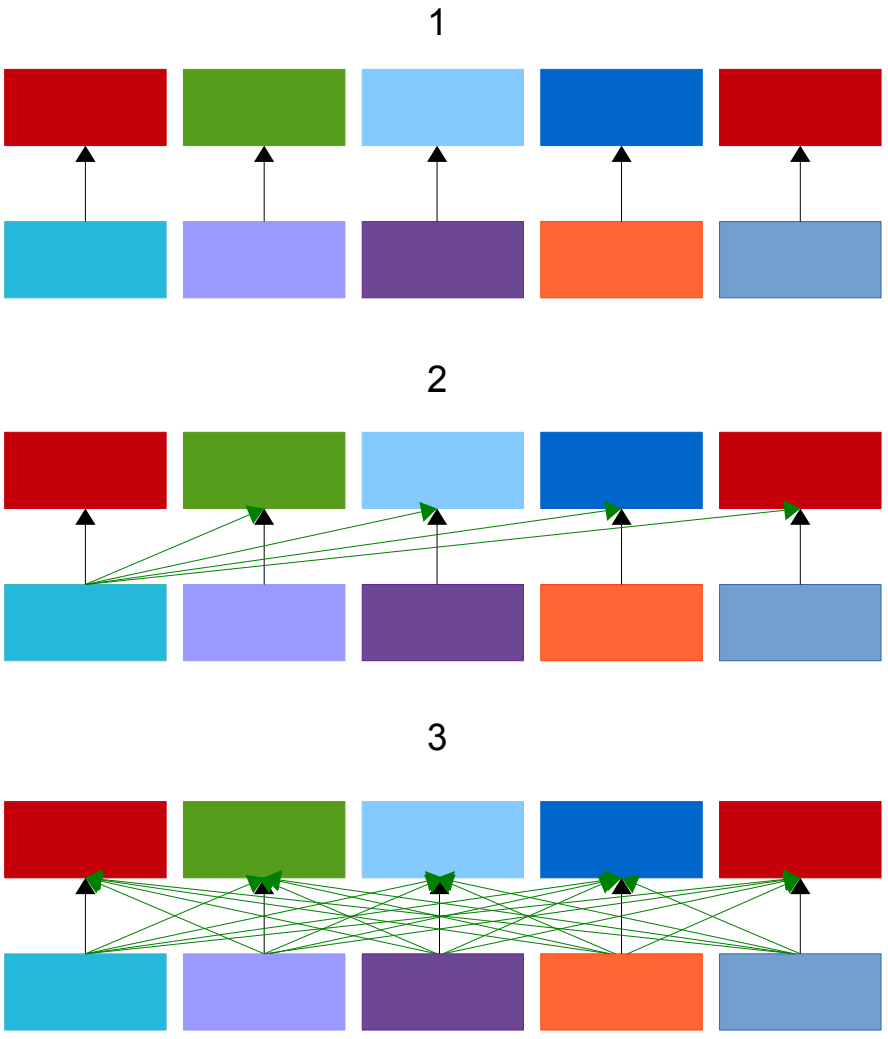

Figura 2.3: Visualizzazione grafica dei legami fra coppie di residui corrispondenti ai diversi termini di eq. 2.2.5

Capitolo 3

Predizione di propensitá ad aggregare: confronto con dati sperimentali

3.1 Predizione della propensitá ad aggregare di peptidi corti

Un primo passo molto importante é stato normalizzare il nuovo termine di potenziale a quello giá presente, cioé si é dovuto determinare il peso relativo di λ nella (2.2.5). Ci si é quindi riferiti a un problema specifico, cioé la capacitá dell'algoritmo di predire se una data sequenza forma o meno una struttura amiloide, e si é provveduto a trovare il λ ottimale per tale problema.

Il nostro problema é quindi l'ottimizzazione dell'efficienza di un predittore con classificazione binaria.

Dato un insieme di classi, per le quali si sappia giá quali sono positive e quali negative, e dato un predittore che permetta di separare le positive dalle negative, il confronto tra la predizione e il risultato sperimentale puó portare a quattro possibili risultati.

Si possono avere un vero positivo (TP), un falso positivo (FP), un vero negativo (TN) e un falso negativo (FN). Il predittore ideale é quello per cui ci sono solo veri positivi e veri negativi, cioé quello che riesce a predire correttamente tutti i risultati sperimentali. Ovviamente, i predittori sono in generale non ideali, cioé i risultati conterranno un certo numero di falsi positivi e falsi negativi.

Un tipico predittore assegna un punteggio (score) a un dato evento e poi lo confronta con un valore di soglia, restituendo una predizione di evento positivo nel caso di superamento della soglia. Quest'ultima si puó scegliere in maniera cosí restrittiva da non predire nessun evento positivo. Cominciando a variare la soglia cominceremo a trovare sempre piú eventi positivi, sia veri che falsi, fino a che non avremo solo eventi positivi.

Un modo per visualizzare questo problema é tramite una curva ROC (Reicever Operating Characteristics). In ascissa una curva ROC ha la frazione di falsi positivi sul numero totale di eventi negativi, ossia (1-TN/(TN+FP)). In ordinata, invece, c'é la frazione di veri positivi sul numero totale di eventi positivi, cioé TP/(TP+FN). Il valore TN/(TN+FP) é anche chiamato specificitá e misura la proporzione di eventi positivi che sono correttamente identificati come tali. La decisione su quale soglia scegliere per una predizione va fatta tenendo conto della situazione. Una soglia bassa corrisponde a un predittore molto specifico, in cui sono minimizzati i falsi allarmi, mentre una soglia alta corrisponde a un predittore sensibile, che cattura cioé tutti o quasi tutti gli eventi interessanti. Nella maggior parte delle applicazioni del nostro predittore, per esempio, é preferibile un'alta specificitá.

La soglia viene alzata poco a poco, in modo che ogni volta che il numero di falsi positivi aumenta di uno sia possibile calcolare il numero di veri positivi. L'area sotto tale curva equivale a 0.5 se il predittore é casuale, ed é tanto maggiore quanto maggiore é la bontá del predittore, dove 1 é l'area del predittore ideale. Abbiamo utilizzato un database fornito di 424 brevi catene

polipeptidiche [7-10], realizzato in condizioni di temperatura e composizione della soluzione non omogenee, di cui sapevamo sperimentalmente quali aggregavano e quali no.

Abbiamo quindi definito come classi positive i peptidi che aggregano, come classi negative i peptidi che non aggregano e come *score* l'energia piú bassa fra quelle degli accoppiamenti β testati da PASTA. Abbiamo quindi calcolato le curve ROC per valori di λ in [0.5:1.5] e intervalli $\Delta\lambda = 0.02$.

Questa operazione é stata fatta sia per l'intero database, sia per un sottoinsieme di *training* scelto in modo casuale. Il λ che ottimizza l'insieme di training é risultato pari a 0.018. I risultati sono poi stati controllati per il resto del database, la parte di *test*, che ha confermato i risultati dell'insieme di *training*.

L'andamento dell'area sotto le curve ROC per l'insieme di training é riportato in figura 3.1. In figura 3.2 si possono invece vedere le curve ROC per l'insieme di test e per l'insieme di training per il valore ottimale ricavato sull'insieme di training $\lambda = 0.018$. In figura 3.3 é infine riportato l'andamento della curva ROC per l'intero database con $\lambda = 0$ e con $\lambda = 0.018$

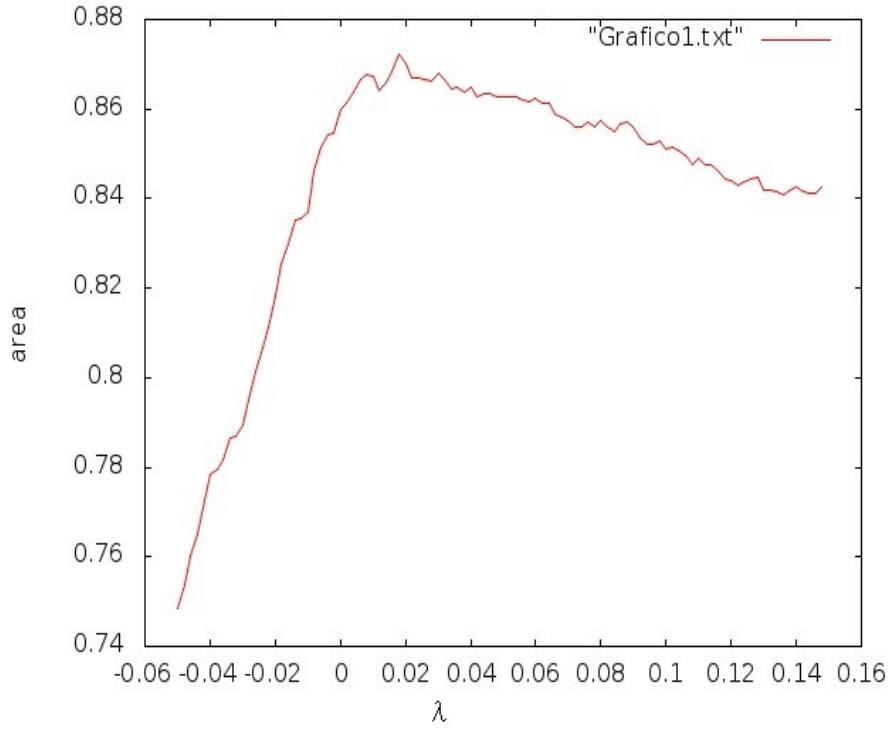

Figura 3.1: Area sotto le curve ROC per l'insieme di training al variare di λ. Il valore ottimizzato é $\lambda = 0.018$

Il risultato di questa operazione é stato piuttosto inatteso. Come si puó vedere dal grafico 3.1 la regione in cui il predittore migliora (un miglioramento di stretta misura ma comunque visibile) non é quella dove il segno é negativo, come atteso, quanto piuttosto quella dove λ é positivo. Il mancato miglioramento nella regione in cui λ é negativo potrebbe essere dovuto a un contributo minore di quanto sperato del termine entalpico H_s associabile allo stato solubile. Si noti che in questo caso abbiamo lavorato con catene piuttosto corte, al massimo venti o trenta residui, e quindi plausibilmente prive di un contributo entalpico significativo nello stato nativo. Il miglioramento nella regione positiva potrebbe invece significare, come spiegato precedentemente nel paragrafo 2.2, che il nuovo termine di potenziale contribuisce al termine di entalpia dello stato fibrillare H_f.

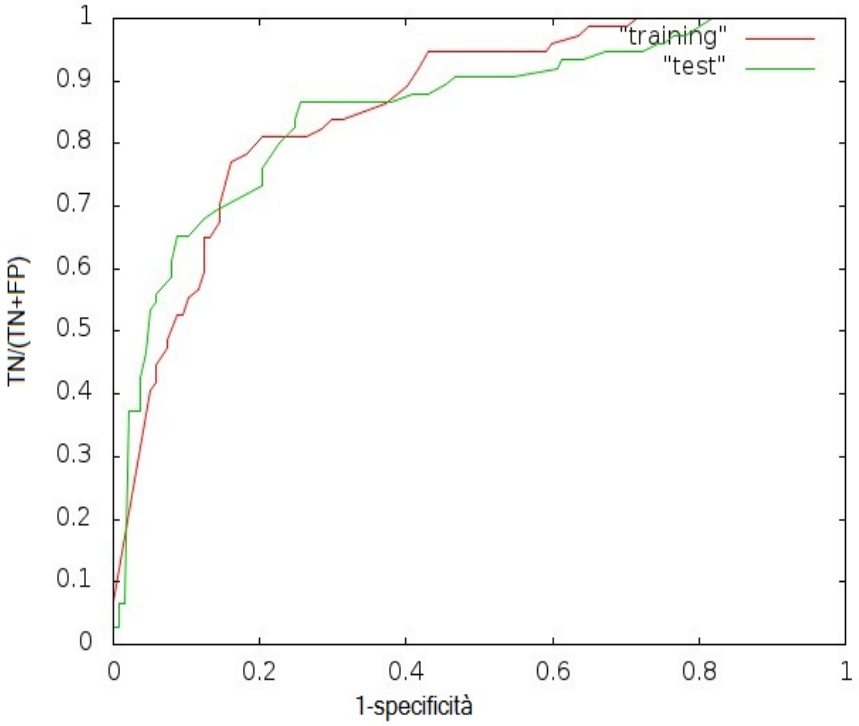

Figura 3.2: Curva ROC per gli insiemi di test e training $\lambda = 0.018$. Si noti che la performance nella regione ad alta specificitá ($x \backsimeq 0.1$) é migliore per l'insieme di test

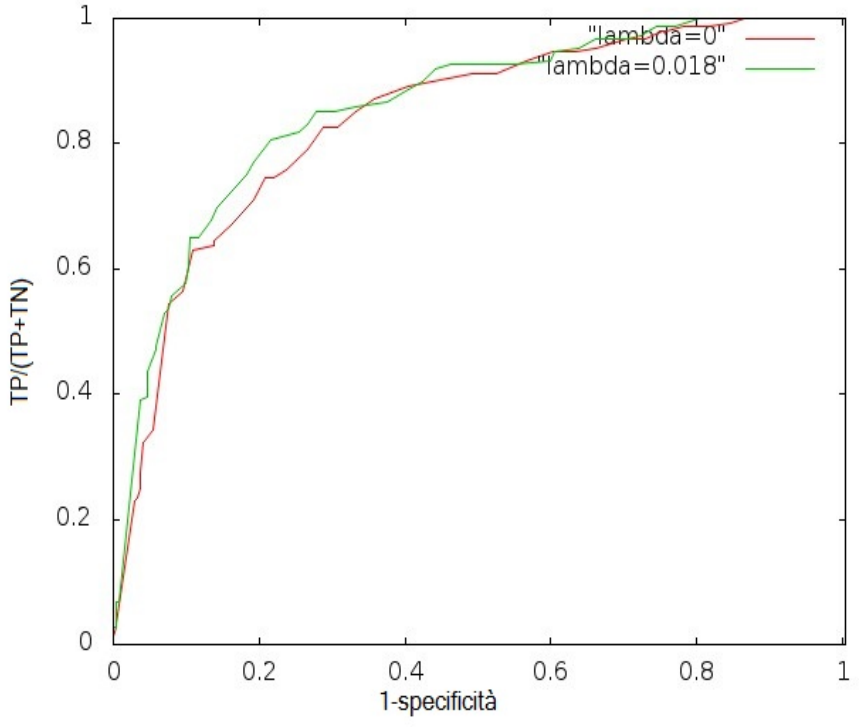

Figura 3.3: Curva ROC intero database $\lambda = 0, \lambda = 0.018$. Si noti che l'aggiunta del termine H_2 porta a un miglioramento delle prestazioni del metodo per quasi tutti i valori di specificitá

3.2 Predizione di profili di energia libera di aggregazione: peptide Aβ40 e prione HETs

Le due catene studiate nel maggiore dettaglio in questa tesi sono l'Aβ-42, insieme alla variante Aβ-40 e a una serie di 16 sue mutazioni, e il prione HETs della *Podospora anserina*.

- L'Aβ-40 é un peptide associato al morbo di Alzheimer, un disordine neurodegenerativo progressivo, caratterizzato da perdita di memoria e cambiamenti nella personalitá. Nel cervello dei pazienti sono stati infatti individuati dei depositi extracellulari, detti placche senili, che sono formati da fibrille del peptide β-amiloide, prevalentemente Aβ-40 e Aβ-42.

  Un modello di struttura della fibrilla amiloide formata dall'Aβ-40 a pH 7.4 e 24°C é stato realizzato recentemente usando il metodo SSNMR in congiunzione con metodi computazionali di minimizzazione dell'energia. In questa struttura ogni molecola contribuisce a due filamenti β, che coprono i residui tra il 12 e il 24 e tra il 30 e il 40. Questi due filamenti contribuiscono a formare due foglietti β distinti, in cui i filamenti sono legati parallelamente tra loro e in registro. Ulteriori analisi hanno portato a ritenere che i protofilamenti siano composti da quattro foglietti β separati da una distanza di circa 10 Å. Una visualizzazione tridimensionale é fornita in figura 3.4

  La predizione di PASTA identifica correttamente la zona legata in struttura β tra i residui 30 e 40, ma i residui tra il 12 e il 24, seppur a energia piú bassa delle zone adiacenti, risultano meno propensi a formare struttura beta aggregata.

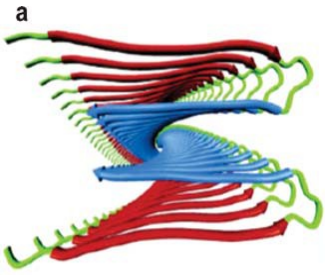

Figura 3.4: Modello tridimensionale del protofilamento Aβ visto dall'asse della fibrilla

- HETs é invece un prione, dall'inglese prion, acronimo di PRoteinaceus Infective ONly particle, cioé particella infettiva solamente proteica. I prioni sono delle proteine che causano malattie contagiose che attaccano di solito il sistema nervoso centrale, come l'encefalopatia spongiforme Bovina e il Morbo di Creutzfeldt-Jakob. I prioni fungono da seme di formazione per le fibrille amiloidi, e possono provocare reazioni a catena.

  Nel caso di HETs, e similmente per altri prioni di lievito, si ritiene che le caratteristiche prioniche abbiano un significato biologico funzionale [15]. Di HETs é stato studiato [5] un modello strutturale basato sulla risonanza magnetica nucleare sullo stato solido della proteina HETs del fungo filamentoso *Podospora anserina*.

  In particolare, quello che si é andato a studiare é l'interno rigido del prione. In tale modello si é rivelata una struttura a β elica, di complessitá strutturale maggiore rispetto ai foglietti β presenti nelle fibrille di Aβ40. La struttura pseudo-ripetuta permette la formazione di un solenoide in cui una catena contiene due giri, stabilizzati da tre ponti salini.

  Un modello di struttura tridimensionale del prione é riportato in figura 3.5

  I problemi principali del profilo energetico predetto da PASTA per il prione HETs sono che il segmento tra i residui 1 e 9 risulta legato in struttura β, mentre il segmento tra 9 e 16 risulta non legato, all'opposto di quanto rilevato sperimentalmente.

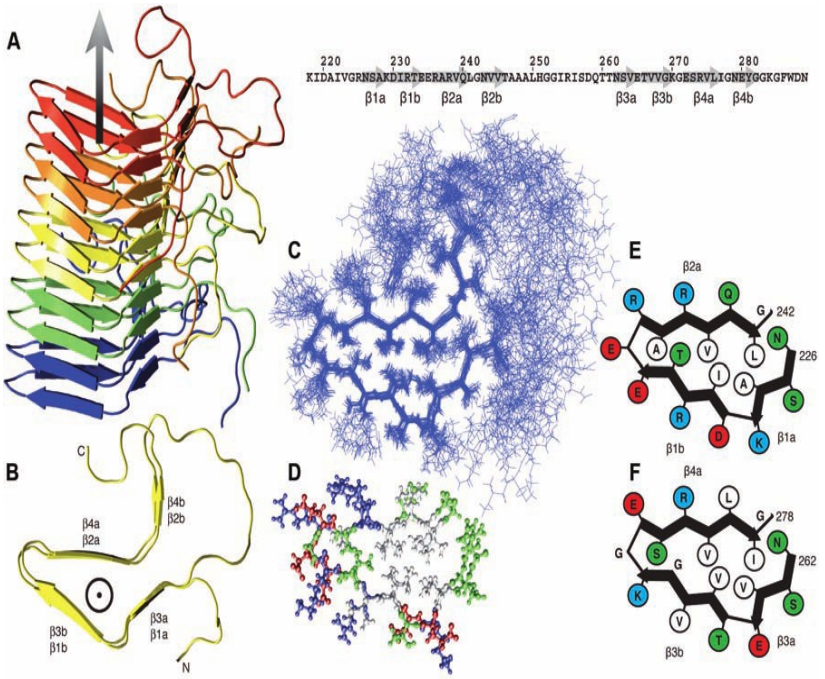

Figura 3.5: Immagine in 3D della struttura del prione HETs

Ponendo $\lambda = 0.018$, come ottimizzato nel paragrafo precedente, e valutando i corrispondenti profili di energia libera di aggregazione si sono subito resi evidenti alcuni problemi. Come si puó notare in figura 3.6, il profilo energetico risulta schiacciato e quasi costante a partire da circa il decimo residuo. La ragione é che gli accoppiamenti piú favorevoli sono in questo caso quelli con L_{max} grande, pari quasi all'intera lunghezza della catena, e sono di gran lunga piú favorevoli di tutti gli altri. Essi vengono aggiunti al conteggio energetico di ogni residuo, appiattendo appunto il profilo.

Il motivo é che il numero di contatti presi in considerazione dal nuovo termine entalpico H_2 aumenta in modo quadratico con l'aumentare di L. In PASTA originale il numero di contatti considerati in H_1 aumenta in modo lineare con L, ma il termine di entropia Δs, anch'esso lineare in L, bilancia questo effetto sfavorendo i contatti piú lunghi. L'introduzione del nuovo termine di potenziale rompe questo equilibrio generando il problema sopracitato.

Una possibile soluzione potrebbe essere di ottimizzare contemporaneamente sia Δs che λ, ma un tale lavoro é al di fuori degli obbiettivi di questa tesi. Abbiamo invece supposto che, per catene lunghe come quelle dell'Aβ-40 e dell'HETs, potesse valere la nostra ipotesi di partenza circa il segno e il significato del termine entalpico H_2, mantenendo per semplicitá il valore assoluto di λ come ottimizzato in precedenza. Abbiamo quindi cercato di valutare gli effetti dell'introduzione del nuovo termine con $\lambda = -0.018$ sul profilo di energia libera di aggregazione di queste due catene.

I grafici di riferimento per questa sezione sono riportati nelle figure 3.7, 3.8. Le barre in rosso rappresentano i segmenti di catena che sperimentalmente risultano partecipare alla struttura β.

Come si puó vedere i risultati ottenuti con $\lambda = -0.018$ alzano, in generale, il profilo di energia libera. Per quanto riguarda l'AB40 l'energia libera é minore rispetto alle zone adiacenti nella sezione tra 12 e 24, che sperimentalmente risulta coinvolta in struttura β. L'altra zona legata, da 30 a 40, é sempre predetta tale. Anche la predizione su HETs presenta qualche miglioramento. La zona tra 22 e 31 ha un'energia libera molto bassa, in corrispondenza segmento 20-28 indicato dal dato sperimentale, e anche l'ultimo minimo da 58 a 65 é in buon accordo con il dato sperimentale. Il primo minimo, che non corrisponde a nessuna zona sperimentale, é ancora presente anche con $\lambda = -0.018$, ma diventa quasi uguale a quello della zona tra 9 e 17 che invece corrisponde al

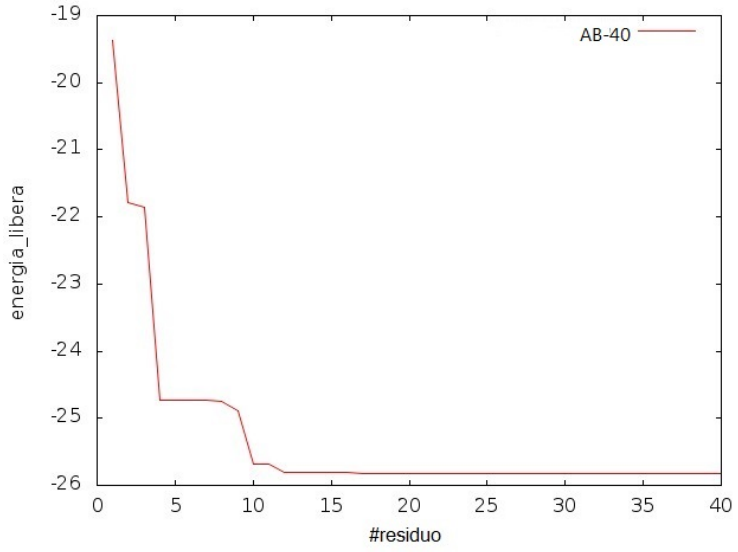

Figura 3.6: Profilo di energia libera del Aβ-40, $\lambda = -0.018$

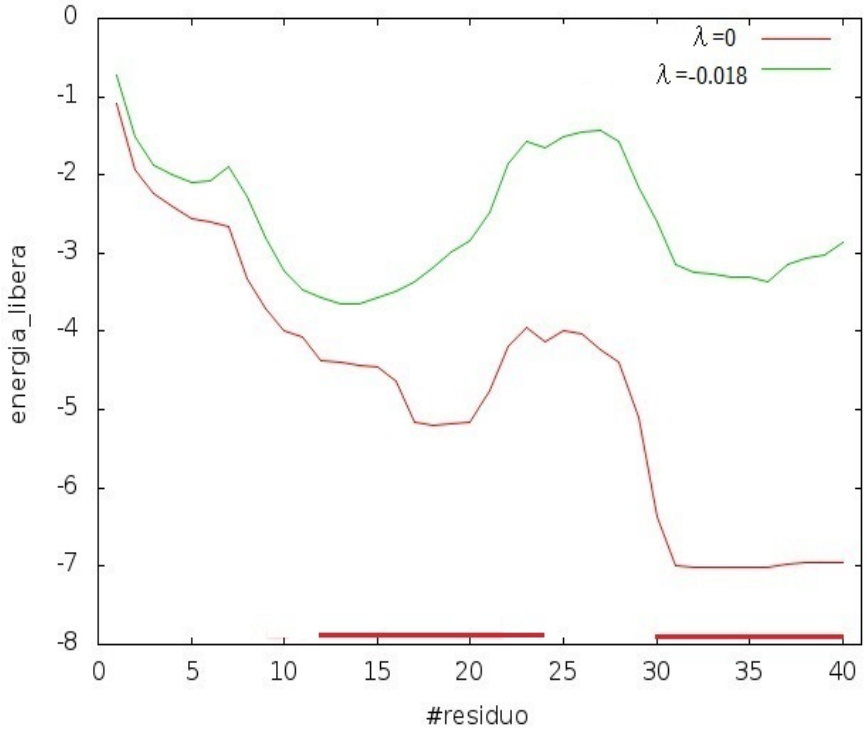

Figura 3.7: Confronto energia libera AB40, $\lambda = -0.018$, $\lambda = 0$

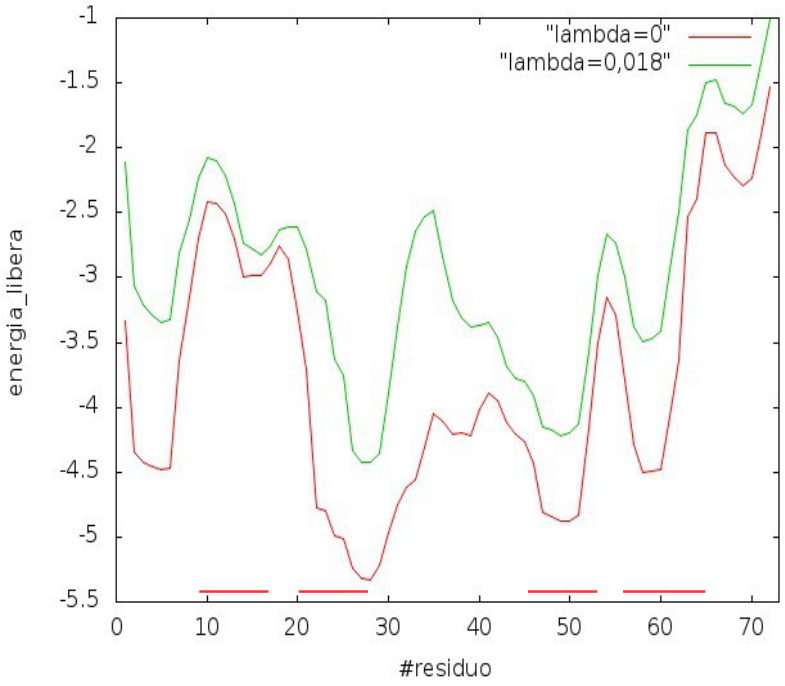

Figura 3.8: Confronto energia libera HETs, $\lambda = -0.018$, $\lambda = 0$

dato sperimentale.

3.3 Predizione di tossicitá in vivo

Le fibrille amiloidi, come giá spiegato precedentemente, sono generalmente collegate a malattie e a disfunzioni dell'organismo. É tuttavia interessante notare come recenti studi abbiano cominciato a suggerire che non sono le fibrille amiloidi ad essere nocive, quanto piuttosto gli aggregati oligomerici transienti, noti anche come protofibrille, anch'essi caratterizzati dalla presenza di strutture β. La velocitá di aggregazione delle fibrille, comunque, é ritenuta influente nel determinare il livello di tossicitá in vivo della proteina che le forma. Studi diretti in questa direzione sono giá stati compiuti in precedenza[4].

Un importante studio ha messo in relazione le propensitá ad aggregare predette da Zyggregator[11], un algoritmo ideato con scopi simili a PASTA, e la tossicitá degli aggregati in vivo. Zyggregator si differenzia da PASTA principalmente perché é stato ottimizzato per riprodurre le velocitá di aggregazione di fibrille in vitro misurate sperimentalmente, e perché non tiene conto degli accoppiamenti β tra i residui (come invece fa PASTA).

Nello studio riportato in [4] sono state create 17 mutazioni del peptide Aβ42, per poi ingegnerizzare mosche della frutta che esprimessero tali peptidi mutati. Si é poi valutata la tossicitá delle varie mutazioni in base al numero di giorni di sopravvivenza delle mosche e in base agli effetti che la mutazione aveva sul loro comportamento (per esempio riduzione dell'abilitá motoria).

La scala é stata rapportata al peptide Aβ40, che era giá stato mostrato non essere tossico. É stata trovata una forte correlazione lineare con il tempo di sopravvivenza, cioé 0.75, mentre una correlazione di 0.65 é risultata dal confronto tra previsioni e deficit motorio. Il valore Z_{aggr} é stato calcolato mediando dei profili di energia libera predetti da Zyggregator simili a quelli presentati precedentemente in questo lavoro. Il valore S_{tox} é ottenuto confrontando il tempo di sopravvivenza delle mosche, S_{mut}, al tempo di sopravvivenza massimo S_{max}, secondo la formula $S_{tox} = (S_{max} - S_{mut})/S_{max}$

Si é provato a valutare tale correlazione utilizzando i profili di energia libera predetti da PASTA con $\lambda = 0$ (figura 3.10). La lista di mutanti utilizzati e i correspondenti valori di tossicitá

Mutanti di Aβ42	Z^a_{aggr}	E_{PASTA}	S^b_{tox}
L17R	0.73	-5.787	0
F20E	0.63	-5.565	0.03
D7R	0.76	-6.116	0.19
K16W	0.76	-6.482	0.19
WT Aβ42	0.75	-5.972	0.20
R5Y	0.70	-6.040	0.23
A2F	0.72	-6.029	0.23
H14W	0.82	-6.109	0.27
E11G	0.79	-5.896	0.34
N27W	0.80	-6.198	0.45
M35F	0.79	-6.356	0.53
E22G	0.85	-6.346	0.73
H6W/E22G	0.83	-6.416	0.65
G9T/E22G	0.84	-6.509	0.77
F4D/E22G	0.84	-6.199	0.45
I31E/E22G	0.85	-6.482	0.13
Aβ40	0.80	-4.714	0

Tabella 3.1: Tabella delle tossicitá e dell'energia libera media. Nella prima riga é riportata la mutazione. La prima letterasi riferisce all'amminoacido che sta per essere sostituito, il numero si riferisce alla posizione in cui l'amminoacido verrá sostituito e la seconda lettera si rferisce all'amminoacido sostitutivo. Le penultime quattro mutazioni sono doppie mutazioni. Le quantitá riportate sono definite nel paragrafo 3.3

impiegati sono mostrati in tabella 3.1. Il profilo energetico di alcuni dei mutanti dell' Aβ-42 é visibile in figura 3.9. La migliore correlazione, attorno a 0.65, si ottiene lasciando la lunghezza massima L_{max} dei possibili accoppiamenti molto alta. Si é quindi testato se l'introduzione del nuovo termine entalpico H_2 portasse a dei miglioramenti.

I risultati si possono vedere in figura 3.11, dove mostriamo la correlazione lineare fra tossicitá e predizione al variare di λ e L_{max}.

Il coefficiente di correlazione lineare, che é negativo, tende al suo minimo in prossimitá dello 0 e ancora quando L_{max} é vicino o pari all'intera lunghezza della catena. Quando λ é positivo il coefficiente di correlazione sale rapidamente, mentre con $\lambda < 0$ tende a restare invariato, sopratutto con l'aumentare di L_{max}. Si noti che mano a mano che L_{max} tende all'intera lunghezza della catena subentrano due fattori. Il primo é che gli accoppiamenti ad energia minore sono generalmente formati da pochi residui, e quindi non ne compaiono di nuovi particolarmente significativi con L_{max} alto. Il secondo é che l'aggiunta di un legame é piú rilevante quando ci sono pochi legami, cioé quando L_{max} é basso. Entrambi questi fenomeni contribuiscono a stabilizzare l'andamento del coefficiente di correlazione lineare dopo un certo valore di L_{max}, di solito 17 o 18.

3.4 Predizione di accoppiamenti intra-catena in strutture prioniche

Incoraggiati dallo studio del prione HETs abbiamo provato a modificare ulteriormente l'algoritmo per tentare di capire se PASTA poteva predire il particolare registro dell'accoppiamento fra i segmenti coinvolti nella struttura fibrillare di HETs. Per far questo é necessario sfruttare l'informazione contenuta in mappe bidimensionali di energia libera di aggregazione, come quella mostrata in figura 3.12, ottenibili a partire da (2.1.9). Si noti in HETs la presenza di legami idrogeno intra-catena che stabilizzano la struttura fibrillare a croce β. Legami intra-catena non

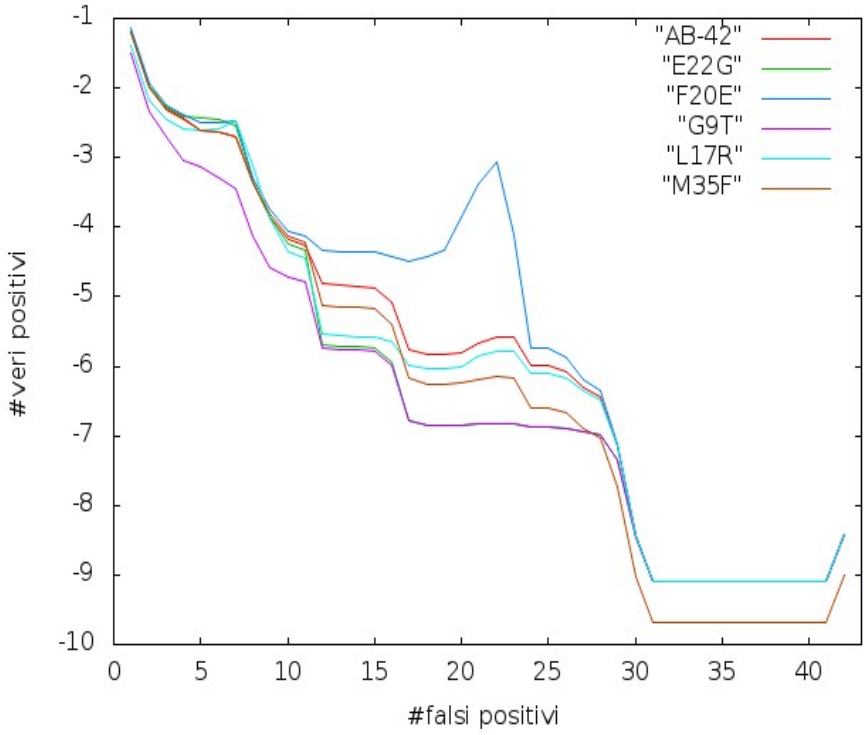

Figura 3.9: Profilo di energia libera per alcuni dei mutanti dell'Aβ-42 ($\lambda = 0$). Si noti che in genere i profili di mutanti piú tossici sono piú bassi che per il peptide wild type (vedi tabella 3.1)

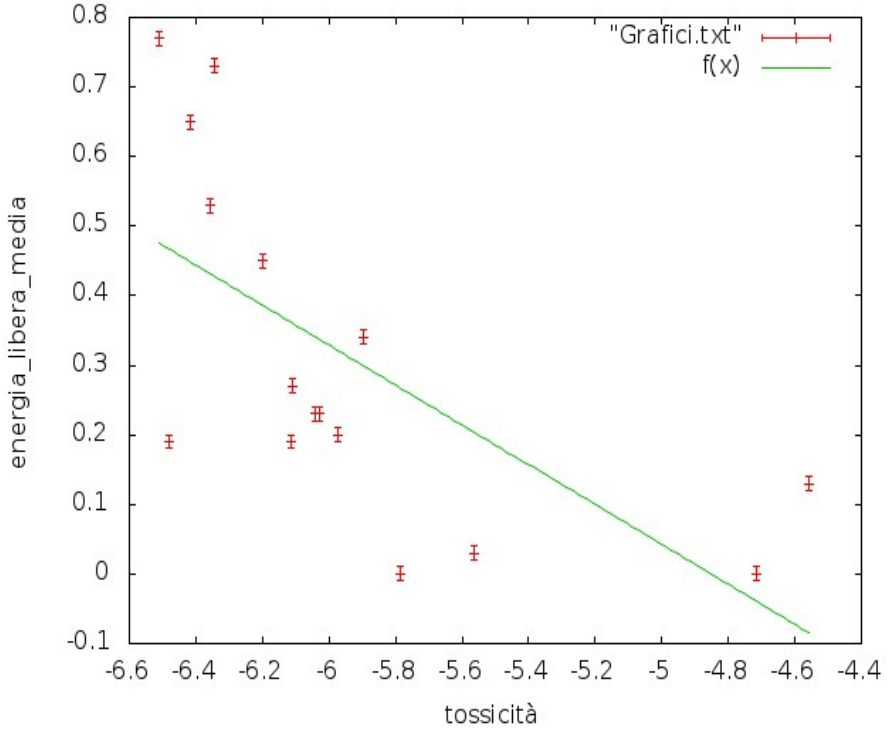

Figura 3.10: Correlazione fra tossicitá misurata in vivo ed energia libera di aggregazione media ($\lambda = 0$)

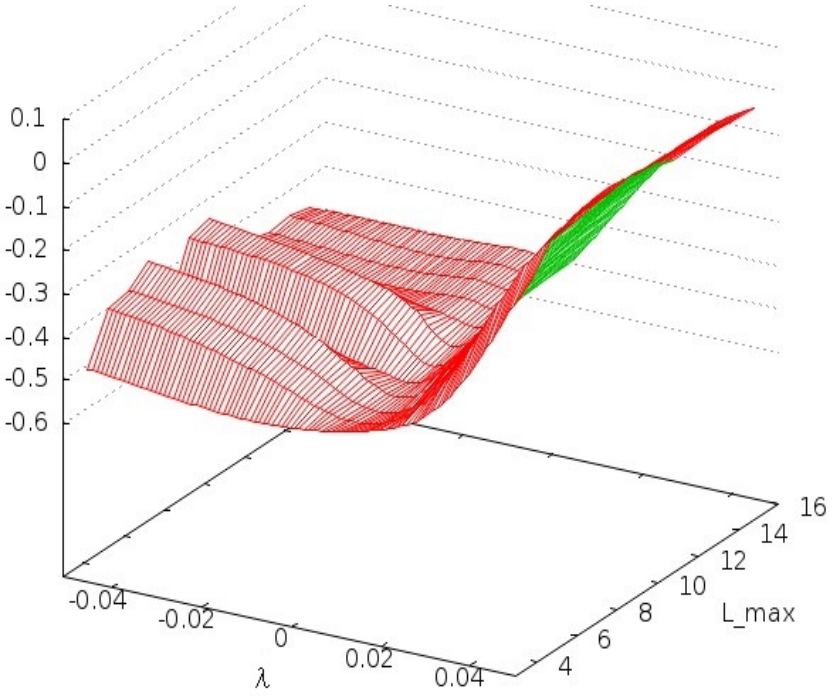

Figura 3.11: Correlazione lineare fra tossicitá misurata in vivo ed energia libera di aggregazione media al variare di λ e L_{max}

sono presenti in Aβ40. Il problema é quindi poter predire accoppiamenti intra-catena senza considerare gli effetti entropici dovuti alla presenza della catena.

Per questo motivo é stata inserita una condizione 'ad hoc' che togliesse tutti gli accoppiamenti in cui lo stesso residuo comparisse sia nel primo che nel secondo segmento. Ci aspettavamo delle zone a energia minima in corrispondenza degli accoppiamenti evidenziati nella tabella sottostante, paralleli e fuori registro.

9-12	45-48
13-17	49-53
19-24	55-60
26-29	62-65

Il programma cosí modificato, in effetti, elimina i contributi da accoppiamenti paralleli in registro, che sono altrimenti dominanti. Abbiamo inoltre aumentato la lunghezza minima degli accoppiamenti, per cercare di identificare lunghe porzioni di catena che potessero stabilizzare l'avvolgimento solenoidale di HETs. L'espediente ha funzionato (sia con $\lambda = 0.018$ che con $\lambda = 0$), supponendo di sapere giá che l'accoppiamento cercato é parallelo e utilizzando la sua lunghezza (21) per definire $L_{max} = L_{min} = 21$. Il migliore degli accoppiamenti paralleli, infatti, é risultato tra 8-28 e 44-64, esattamente quello visto sperimentalmente (figura 3.15).
Questi risultati preliminari sono incoraggianti, ma ulteriore lavoro sará necessario per elaborare un metodo predittivo che non sfrutti (come fatto in questo lavoro) conoscenze pregresse riguardo alle strutture che si vogliono predire. In particolare, lo scopo finale sará ottenere predizioni per gli accoppiamenti β del prione umano, per cui non é nota al momento una struttura ad alta risoluzione come quella di HETs.

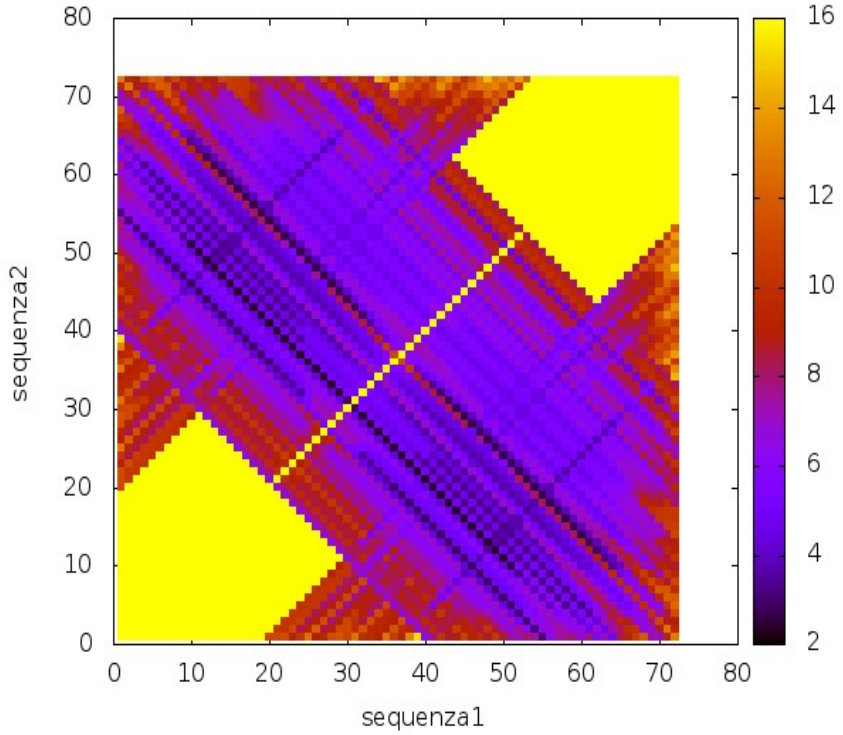

Figura 3.12: Energia libera per gli accoppiamenti $\ln(z(k,m))/\zeta$ del prione HETs con $L_{max} = L_{min} = 21$ e il vincolo che solo accoppiamenti possibili intra-catena sono considerati (con z(k,m) dato dalla 2.1.9). L'unico segnale di un accoppiamento parallelo fuori diagonale é quello fra 8-28 e 44-64, che corrisponde esattamente al dato sperimentale

Capitolo 4

Conclusioni

L'insorgere di molte malattie é legato all'aggregazione di proteine in fibrille amiloidi, caratterizzate da una struttura basata sui foglietti β. Poter predire la propensitá delle catene peptidiche ad aggregare per formare tali strutture é un problema di grande interesse biomedico, di cui il gruppo di ricerca nel quale ho svolto la tesi si é interessato. L'algoritmo sul quale ho lavorato, PASTA, é pensato per rispondere a questo bisogno.
PASTA si basa sull'ipotesi che le proteine considerate siano sostanzialmente non strutturate nel loro stato nativo, tiene conto della sola struttura primaria della catena e si basa sui contatti fra i residui nella particolare geometria dei foglietti β. Il numero di contatti considerati é inoltre lineare con il numero di residui nei segmenti considerati accoppiati nella struttura β aggregata. In questo lavoro é stato introdotto un nuovo termine che potesse valutare il contributo entalpico associato allo stato solubile delle proteine.
Il nuovo termine si é dimostrato in grado di migliorare la predittivitá rispetto all'aggregazione o meno di un database di 424 brevi peptidi, a patto peró di scegliere il coefficiente di normalizzazione λ maggiore di zero. Questo puó voler dire che, almeno per quanto riguarda catene brevi, cioé mediamente di una quindicina di residui, il nuovo termine va a migliorare la stima dell'entalpia H_f dello stato fibrillare piuttosto che stimare quella dello stato solubile. Una spiegazione di questo fenomeno potrebbe essere la presenza di un effetto cooperativo, cioé la presenza di un legame tra due residui potrebbe influenzare la forza dei legami tra i residui circostanti.
L'applicazione dell'algoritmo su sequenze piú lunghe ha peró mostrato come esse, pur essendo naturalmente non strutturate nel loro stato nativo, possano efettivamente mostrare un certo grado di struttura residua. I profili energetici dell'A-40 e del prione HETs hanno infatti mostrato dei miglioramenti rispetto ai dati sperimentali noti ponendo $\lambda < 0$.
Il numero di interazioni del termine introdotto in questo lavoro di tesi é quadratico con la lunghezza dei segmenti accoppiati, mentre lasciato come originale é lineare. In generale, i termini energetici tendono a favorire accoppiamenti molto lunghi; in PASTA originale questo effetto era bilanciato dal termine di entropia ΔS, anch'esso lineare nella lunghezza dell'accoppiamento.
Un'ipotesi di un lavoro per proseguire il progetto iniziato in questa tesi potrebbe essere di cercare di ottimizzare contemporaneamente sia ΔS che λ in modo da armonizzare meglio i due termini. Per farlo servirebbero delle catene lunghe, come per esempio i mutanti dell'Aβ-42 giá studiati in questa tesi. In questo lavoro non siamo riusciti a migliorare la correlazione fra tossicitá in vivo ed energia libera media per residuo, ottenuti con il predittore Zyggregator [11], ma é possibile che un'ottimizzazione di entrambi i parametri possa portare a dei miglioramenti.
Infine, anche l'aver individuato la zona di accoppiamento corretta del prione HETs é un successo di PASTA, nonostante esso sia stato possibile solo grazie a una conoscenza pregressa della sua struttura.

Bibliografia

[1] Fabrizio CHiti, Christopher M Dobson (2006) Protein Misfolding, Functional Amyloid, and Human Disease, Annu. Rev. Biochem. (333-365)

[2] Antonio Trovato1, Fabrizio Chiti, Amos Maritan1, Flavio Seno (2006) Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins, PLoS Computational Biology (1608-1618)

[3] Zsuzsanna Dosztanyi, Veronika Csizmok, Peter Tompa et al. (2005) The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Foded and Intrinsically Unstructured Proteins, J. Mol. Biol. (827-839)

[4] Leila M. Luheshi, Gian Gaetano Tartaglia, Ann Christin Brorsson et al. (2007) Systematic In Vivo Analysis of the Intrinsic Determiants of amyloid β Patogenicity, PLoS biology (2493-2500)

[5] Christian Wasmer, Adam Lange, Helene Van Melkebecke et al. (2008) Amyloid Fibrils of HET-s (218-289) prion form a β solenoid with a triangular Hydrophobic core, SCIENCE (1523-1526)

[6] Ian Walsh, Flavio Seno, Antonio Trovato (2014) PASTA 2.0 : an improved server for protein aggregation prediction, Nucleic Acids Research (301-307)

[7] Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. et al. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotech. (1302-1306).

[8] Roland,B.P., Kodali,R., Mishra,R. and Wetzel,R. (2013) A serendipitous survey of prediction algorithms for amyloidogenicity. Biopolymers, (780-789).

[9] Thompson,M.J., Sievers,S.A., Karanicolas,J. et al. (2006) The 3D prole method for identifying fibril-forming segments of proteins. Proc. Natl. Acad. Sci. U.S.A. (4074-4078).

[10] Garbuzynskiy,S.O., Lobanov,M.Y. and Galzitskaya,O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics (326-332).

[11] Pawar AP, Dubay KF, Zurdo J, et al. (2005) Prediction of aggregation-prone and aggregation-susceptible regions in proteins associated with neurodegenerative diseases. J Mol Biol (379-392).

[12] Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (2577-2637)

[13] Anfinsen CB (1973) Principles that govern the folding of protein chains, Science 181 (223-230)

[14] Lovell SC, Davis IW, Adrendall WB, et al. (2003) Structure validation by C-alpha geometry: phi,psi and C-beta deviation. Proteins 50 (437-450).

[15] Tuite MF, Serio TR. (2010) The prion hypothesis: from biological anomaly to basic regulatory mechanism. Nat Rev Mol Cell Biol 11 (823-833).

4.1 Ringraziamenti

Grazie prima di tutto a mia madre e mio padre, per il sostegno e per i sacrifici che mi danno la possibilitá di studiare. Un ringraziamento particolare va anche al Bigio, per avermi ospitato, e a Tino e Leo, per avermi preso come coinquilino. Avete fatto piú di quanto non crediate.
Ringrazio ovviamente Antonio Trovato, per la passione che ha messo nel cercare di trasmettermi il piú possibile, per la gentilezza e per la pazienza, sopratutto durante la fase di stesura.
Grazie ai PDF, per l'affetto e il divertimento, per essere sempre uniti. Grazie alla giaverna, perché ogni tanto nerdare fa bene. Non posso non citare direttamente il Budi, ma non staró qui a fare l'elenco dei motivi.
Grazie al Quiricio, a Isa e a tutti coloro che contribuiscono a rendere le lezioni un momento piacevole. Grazie a quelli che mi hanno concesso stima e amicizia, grazie per avermi aiutato a diventare quello che sono.