UNIVERSITÀ DEGLI STUDI DI PADOVA

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

PALMO: un predittore di aggregazione proteica

Relatore:
prof. Carlo Ferrari

Correlatore:
prof. Silvio C. E. Tosatto

Candidato:
Mattia Meneguzzo

Anno accademico 2012-2013

A mio nonno Massimiliano,
nel centenario della nascita:
alpino, instancabile
lavoratore, uomo grande
nella sua umiltà.

Sommario

Questo elaborato presenta un nuovo metodo computazionale, finalizzato a prevedere l'amiloidogenicità delle proteine. Il termine *amiloidogenicità* fa riferimento alla propensione con cui una sequenza proteica dà origine ad aggregati fibrillari, detti appunto *amiloidi*, cui è riconosciuto un ruolo cruciale nello sviluppo di varie patologie neurodegenerative.
La comunità bioinformatica ha ideato un gran numero di approcci algoritmici che, a partire dalla sola sequenza di amminoacidi della proteina e in base a varie assunzioni di carattere biochimico sulla struttura molecolare delle fibrille, tentano di dare risposta a due distinti problemi predittivi: riconoscere, in un insieme di frammenti proteici dalle proprietà sconosciute, quelli responsabili di innescare fenomeni aggregativi e, in secondo luogo, rilevare eventuali regioni amiloidogeniche all'interno di proteine complete.
Il metodo qui esposto, denominato Palmo (*Protein Aggregation Likelihood and Mutation Optimization*), propone un approccio fondato sul riconoscimento delle interazioni esistenti tra gli elementi strutturali fondamentali (i *filamenti-β*) degli aggregati fibrillari, grazie ad un modello in grado di catturare l'influsso del contesto amminoacidico circostante su tali interazioni. Da un punto di vista computazionale, Palmo combina un modulo di apprendimento automatico, sotto forma di una rete neurale, e un algoritmo basato sul paradigma della programmazione dinamica.
Benché sussistano margini di miglioramento, la verifica del nuovo metodo su proteine di struttura nota evidenzia già ora un promettente livello di accuratezza, specie se confrontato con analoghi algoritmi, con risultati di particolare rilievo in termini di specificità di classificazione.

Ringraziamenti

Desidero innanzitutto esprimere la mia gratitudine al prof. Carlo Ferrari, per avermi dato l'opportunità di intraprendere questo lavoro di tesi.
Un sentito ringraziamento va al prof. Silvio Tosatto e al dott. Ian Walsh: senza il loro indispensabile supporto di carattere tecnico-scientifico e la comprensione dimostrata nei miei confronti durante il lungo e accidentato processo di sviluppo, il progetto di ricerca all'origine di questa tesi non sarebbe stato possibile. Non posso non citare anche gli altri componenti del Laboratorio di Biologia Computazionale *BioComputing UP*, quotidiani colleghi di ricerca nei mesi appena trascorsi.
Ultimi, ma non meno fondamentali, i miei familiari, i quali, credendo in me giorno dopo giorno e fornendomi un sostegno morale e affettivo (oltre che finanziario), mi hanno accompagnato in ogni passo del lungo cammino di crescita che culmina, oggi, con il traguardo della laurea. Grazie di cuore!

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

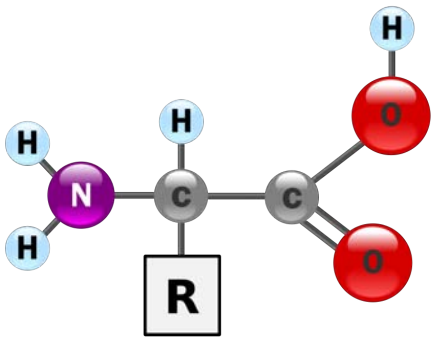
Obiettivo di questo capitolo introduttivo è, innanzitutto, fornire una sommaria panoramica su alcuni concetti chiave della biochimica, quali amminoacidi e proteine, con maggiore attenzione agli aspetti funzionali allo scopo del presente elaborato. Si approfondirà, poi, il tema dell'aggregazione proteica, dando conto delle cause e delle peculiarità strutturali del fenomeno. Si passerà, infine, ad una visione d'insieme sullo stato dell'arte nel campo della predizione di aggregazione proteica per via computazionale, descrivendo gli approcci algoritmici adottati dai numerosi metodi che la comunità bioinformatica ha elaborato nel tentativo di affrontare il problema e con i quali PALMO, prodotto finale del lavoro di progettazione e sviluppo qui presentato, si trova a doversi confrontare.

1.1 Aminoacidi e proteine

Le **proteine** sono macromolecole biologiche consistenti di una o più catene di **amminoacidi**. Di cruciale importanza in qualsiasi organismo vivente, esse rivestono un'ampia varietà di ruoli cruciali, quali il trasporto di molecole, la catalisi di reazioni metaboliche, la difesa immunitaria, la replicazione del codice genetico e varie funzioni strutturali e meccaniche.
Gli amminoacidi, 20 in tutto, sono molecole organiche più semplici, costituite (Figura 1.1a) da un atomo di carbonio tetraedrico C_α, cui si legano un atomo di idrogeno -H, un gruppo funzionale amminico -NH_2, un gruppo carbossilico -COOH ed un gruppo laterale -R, che, specifico di ciascun amminoacido, ne determina le proprietà chimiche (Tabella 1.1b).

1.1.1 Struttura delle proteine

Tra il gruppo carbossilico di un primo amminoacido ed il gruppo amminico di un secondo può instaurarsi un legame covalente, detto *legame peptidico*, che costituisce il fondamentale fattore di coesione delle **catene polipeptidiche** alla base delle proteine. In particolare, l'instaurarsi in successione dei legami peptidici dà luogo alla *catena principale* (*backbone*), da cui sporgono i gruppi laterali degli amminoacidi, detti anche, in quest'ambito, *residui*,



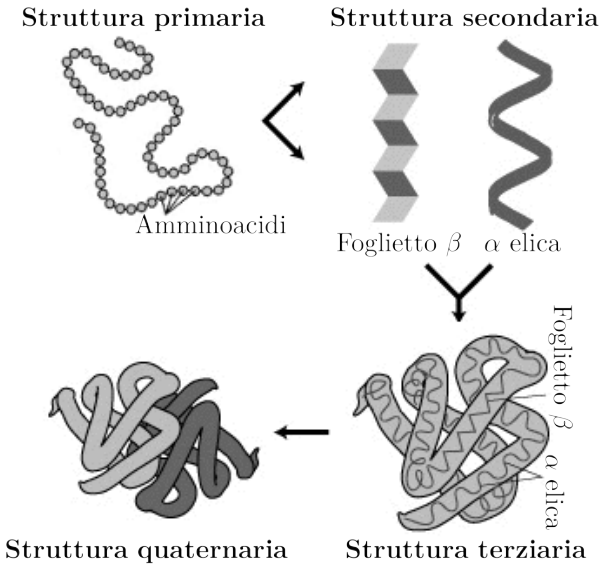
(a) Struttura chimica di un amminoacido. [Wik13b]

Nome	Simbolo		Tipo di R
Alanina	A	Ala	idrofobo
Cisteina	C	Cys	idrofilo
Acido aspartico	D	Asp	acido
Acido glutammico	E	Glu	acido
Fenilalanina	F	Phe	idrofobo aromatico
Glicina	G	Gly	idrofobo
Istidina	H	His	basico
Isoleucina	I	Ile	idrofobo
Lisina	K	Lys	basico
Leucina	L	Leu	idrofobo
Metionina	M	Met	idrofobo
Asparagina	N	Asn	idrofilo
Prolina	P	Pro	idrofobo
Glutammina	Q	Gln	idrofilo
Arginina	R	Arg	basico
Serina	S	Ser	idrofilo
Treonina	T	Thr	idrofilo
Valina	V	Val	idrofobo
Triptofano	W	Trp	idrofobo aromatico
Tirosina	Y	Tyr	idrofilo aromatico

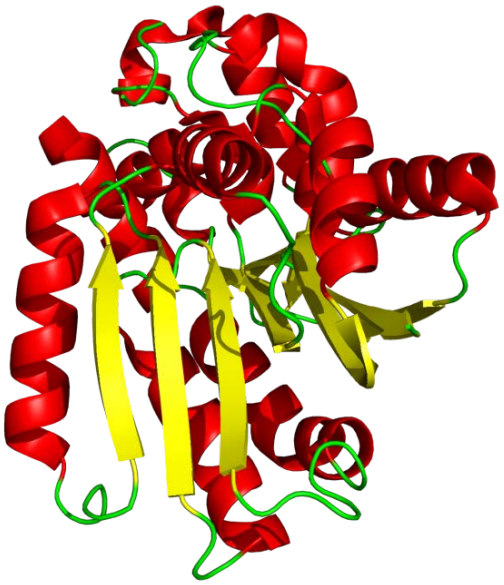
(b) I 20 amminoacidi standard: di ciascuno è riportato il nome, il simbolo (a una e a tre lettere) ed una classificazione in base alle proprietà chimiche del gruppo laterale. [Wik13b]

Figura 1.1: Gli amminoacidi: struttura chimica e classificazione.

a formare la *catena laterale* (*side chain*). La sequenza ordinata con cui gli amminoacidi, così legati, si succedono lungo la catena polipeptidica costituisce la **struttura primaria** della proteina.
La conformazione spaziale della catena a livello locale determina, invece, la **struttura secondaria**. Nella maggior parte dei casi, questa è data dalla combinazione di due tipi di strutture locali, caratterizzati da una topologia ben definita: α eliche e foglietti β (Figura 1.2b). L'***α* elica** (α-*helix*), la più semplice e diffusa conformazione secondaria, si contraddistingue per una disposizione a spirale della catena peptidica, consolidata da legami tra amminoacidi situati a distanze regolari lungo la catena stessa. Il **foglietto *β*** (β-*sheet*) consiste in una struttura planare molto compatta, composta da segmenti peptidici adiacenti e reciprocamente legati. Gli elementi di raccordo tra le strutture dei due tipi precedenti, dove la catena polipeptidica cambia la propria direzione complessiva, sono detti **curve** (*turns*). Così come il legame peptidico è il collante della struttura primaria, le strutture secondarie sono stabilizzate da *legami idrogeno*. Segmenti della catena che non assumono alcuna delle conformazioni appena descritte o delle loro varianti sono indicati con il termine **coil**. Essendo di primario interesse per gli scopi della presente tesi, si torneranno ad approfondire il legame idrogeno e le strutture β nel seguito.
La **struttura terziaria** consiste nella configurazione tridimensionale globale della catena



(a) I quattro livelli di organizzazione di una proteina. [Wik13e]



(b) Modello a nastro di una proteina, dove le α eliche sono evidenziate in rosso, i foglietti β in giallo e le curve in verde. In primo piano, un foglietto β parallelo composto da tre filamenti β.

Figura 1.2: Struttura delle proteine.

polipeptidica, definita dalle coordinate atomiche degli amminoacidi. Unità fondamentali della struttura terziaria sono i *domìni*, ovvero porzioni della catena polipeptidica, costituiti da combinazioni di elementi di struttura secondaria, le quali si ripiegano stabilmente in modo autonomo dal resto della proteina. Il consolidamento della struttura terziaria dipende principalmente da ponti disolfuro e forze di Van der Waals.
Infine, il modo in cui più catene polipeptidiche, dette *subunità*, si associano reciprocamente nello spazio costituisce la **struttura quaternaria**.
Un aspetto di rilevanza fondamentale è lo stretto legame esistente tra la sequenza amminoacidica e la struttura tridimensionale di una proteina [Anf73] e tra quest'ultima e la sua funzione: proteine con struttura primaria analoghe sono solitamente accomunate da una disposizione spaziale ed una funzione simili e, d'altra parte, piccole modifiche nella sequenza di amminoacidi possono provocare drastici cambiamenti nella funzionalità.

1.1.2 Sintesi e classificazione delle proteine

Le informazioni sulla composizione delle proteine sono contenute nel codice genetico degli organismi, sotto forma di DNA. La sequenza di DNA relativa ad una proteina viene trascritta in una corrispondente sequenza di mRNA, che è poi tradotta in una particolare catena lineare di amminoacidi durante la sintesi proteica. Le interazioni reciproche fra gli

amminoacidi della catena, che inizialmente è priva di una conformazione spaziale definita, avviano il cosiddetto *ripiegamento* (*folding*), nel quale viene a formarsi la struttura tridimensionale stabile che costituisce lo *stato nativo* della proteina.
In prima approssimazione, le proteine si distinguono in tre classi principali: *fibrose* (con funzione tipicamente strutturale e solitamente insolubili in acqua), *globulari* (solubili in acqua, delle quali molte sono enzimi) e *di membrana*.

1.1.3 Legami idrogeno e strutture β

Come già accennato, la peculiare conformazione spaziale delle strutture secondarie α elica e foglietto β è stabilizzata da **legami idrogeno** fra atomi della catena principale.

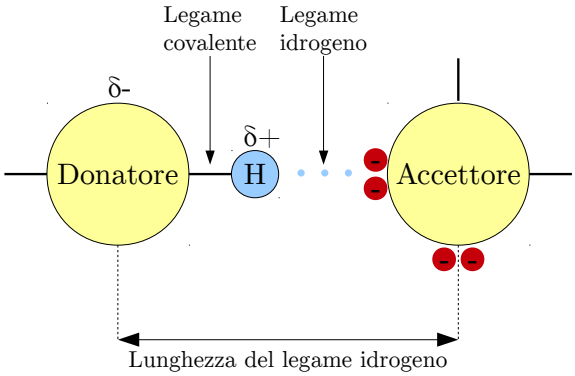
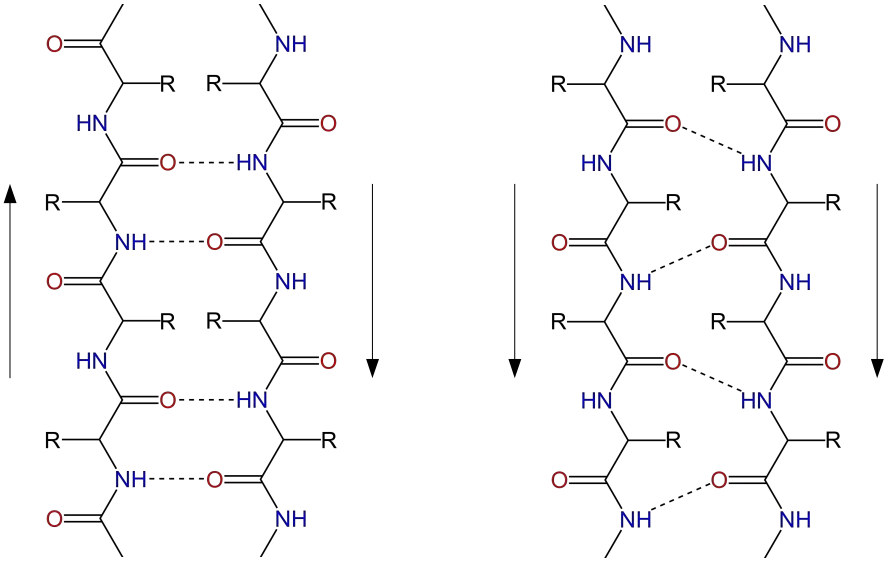


Figura 1.3: Rappresentazione schematica di un legame idrogeno.

In generale, un legame idrogeno (Figura 1.3) consiste in una interazione tra un atomo di idrogeno H coinvolto in un legame covalente con un primo atomo fortemente elettronegativo (ad esempio, ossigeno O o azoto N), che prende il nome di *donatore*, ed una coppia di elettroni liberi di un secondo atomo elettronegativo, detto *accettore*. Il donatore cattura l'elettrone dell'atomo di idrogeno, acquisendo una carica parziale negativa ($\delta-$) e lasciando all'idrogeno una carica parziale positiva ($\delta+$), che viene a sua volta attratta dalla coppia di elettroni del donatore.
Lungo la catena principale di una proteina, un legame idrogeno può coinvolgere un atomo di idrogeno del gruppo N−H di un primo amminoacido, che funge da donatore, ed un atomo di ossigeno di un gruppo C=O di un secondo amminoacido, che assume il ruolo di accettore.
Quando più segmenti peptidici, detti **filamenti β**, si dispongono adiacenti l'un l'altro ed instaurano reciprocamente legami idrogeno, si origina una conformazione planare molto compatta, definita **foglietto β**; denominazione che si deve al tipico aspetto di questa struttura, somigliante, per l'appunto, ad un foglio pieghettato (*pleated β-sheet*), caratterizzato solitamente da una leggera torsione (*twist*) destrorsa.
Nel legarsi reciprocamente, due filamenti β possono assumere orientazioni contrarie oppure

estendersi secondo la medesima orientazione:[1] nel primo caso, si parla di **allineamento antiparallelo**, nel secondo di **allineamento parallelo** (Figura 1.4). In orientamento an-



(a) Foglietto β antiparallelo. (b) Foglietto β parallelo.

Figura 1.4: Struttura atomica di due foglietti β, antiparallelo (a sinistra) e parallelo (a destra), costituiti ciascuno da una coppia di filamenti β, dove i legami idrogeno sono simboleggiati dalle linee tratteggiate. Le frecce puntano dall'N-terminale al C-terminale. [Wik13c]

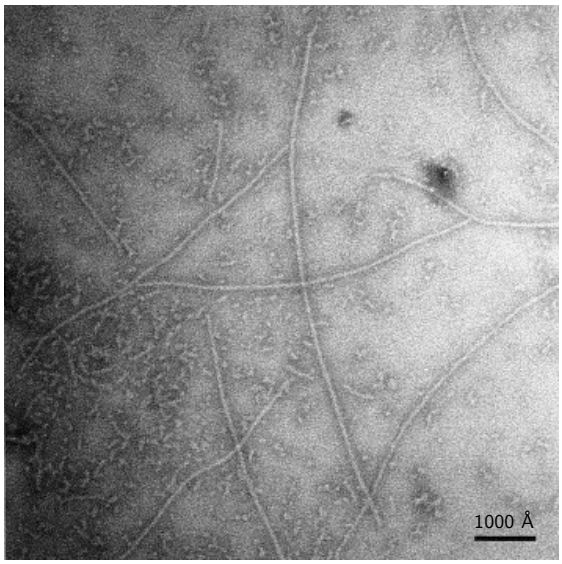
tiparallelo, il gruppo NH ed il gruppo CO di un amminoacido instaurano legami idrogeno con, nell'ordine, il gruppo CO ed il gruppo NH del corrispondente amminoacido nel filamento β adiacente. Nel caso parallelo, invece, un amminoacido vede il proprio gruppo NH legarsi al gruppo CO di un primo amminoacido del filamento β adiacente, mentre il proprio gruppo CO instaura un legame con il gruppo NH di un amminoacido a due posizioni di distanza dal precedente [BSTC02]. A sua volta, un foglietto β può essere *antiparallelo* se i filamenti β che lo compongono assumono a due a due orientamento antiparallelo, *parallelo* se tutti i filamenti condividono la medesima orientazione, *misto* altrimenti.
Nelle proteine globulari, i foglietti β comprendono normalmente da 2 a 22 filamenti β, 6 in media; questi ultimi si estendono solitamente per non più di 15 residui in lunghezza, con un valore medio di 6. I foglietti β paralleli raramente contengono meno di 5 foglietti, il che suggerisce che questa conformazione sia meno stabile della controparte antiparallela, a causa, forse, della differente disposizione dei legami idrogeno che ne consolidano la struttura [VV11].

[1] Alla catena polipeptidica, infatti, è attribuito un orientamento convenzionale, con estremo iniziale il gruppo amminico -NH_2 libero (detto *N-terminale*) ed estremo finale il gruppo carbossilico -COOH libero (chiamato *C-terminale*).

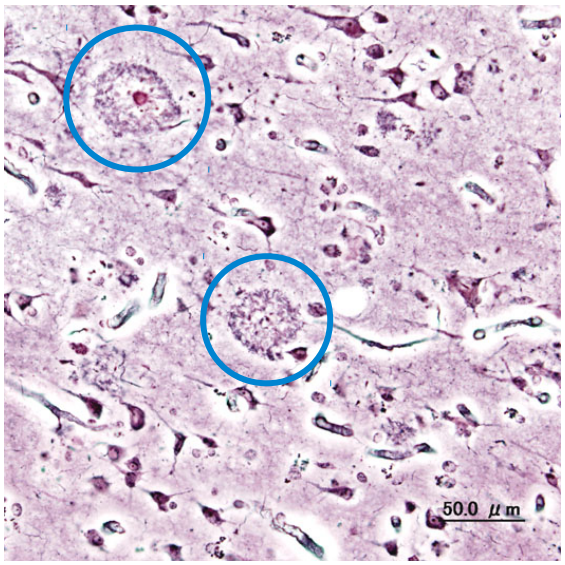
Le conformazioni β, oltre ad essere annoverate tra gli elementi costitutivi basilari della struttura secondaria delle proteine in stato nativo, rappresentano anche, in condizioni patologiche, l'unità strutturale fondamentale delle cosiddette *fibrille amiloidi* all'origine del fenomeno dell'*aggregazione proteica*, come si vedrà più in dettaglio nella prossima sezione.

1.2 Aggregazione proteica e fibrille amiloidi

Varie patologie, talvolta fatali, sono correlate con l'accumulo extracellulare, in certi tessuti dell'organismo, di proteine normalmente solubili sotto forma di aggregati insolubili, di struttura estremamente ordinata, noti come **amiloidi**[2] (Figura 1.5). Tra di esse, si an-



(a) Fibrille (diametro: $70 \div 80$Å) formate da β amiloide A$\beta_{1\text{-}42}$. [Ser00]



(b) Cerchiate in blu, due placche senili, dovute al deposito di β amiloide, nella corteccia cerebrale di un paziente affetto da morbo di Alzheimer. [Wik13a]

Figura 1.5: Immagini al microscopio di fibrille amiloidi e di un esempio degli effetti provocati dal loro accumulo in tessuti umani.

noverano alcune malattie neurodegenerative, quali il **morbo di Alzheimer** e il **morbo di Parkinson**, note per il loro drammatico impatto su ampie fasce di popolazione in età avanzata, le **encefalopatie spongiformi trasmissibili** (TSE o *transmissible spongiform encephalopathies*), queste ultime infettive, e, in genere, tutte le patologie che rientrano nella più ampia definizione di **amiloidosi**, nelle quali il deposito di aggregati proteici in

[2]Il nome *amiloide*, letteralmente "simile all'amido", fu in origine attribuito a questo tipo di aggregati in base alla convinzione, rivelatasi poi errata, che essi avessero una costituzione chimica affine a quella dell'amido.

vari organi (milza, cuore, cervello, fegato, rene) interferisce con le normali funzioni delle cellule, causando la morte di queste ultime fino a provocare, talora, la completa disfunzione dell'organo interessato.
D'altra parte, è stata anche dimostrata l'esistenza di amiloidi funzionali che, in alcuni organismi, assumono importanti ruoli benefici [ON08].
Per quanto le proteine amiloidogeniche esibiscano, in stato nativo, le più varie conformazioni tridimensionali, gli amiloidi che ne hanno origine sono accomunati da un'unico elemento strutturale di base: la **fibrilla amiloide**.
Per la verità, la formazione di fibrille coinvolge non solo *in vivo* le proteine amiloidogeniche nei contesti, spesso patologici, sopra elencati, ma può anche essere indotta *in vitro* con tecniche di laboratorio: le fibrille sintetiche ottenute in questo modo hanno struttura del tutto analoga alle prime, ma sono più correttamente indicate con l'aggettivo *simil-amiloidi* (*amyloid-like*) [WBB$^+$05].

1.2.1 Fattori determinanti l'amiloidogenesi

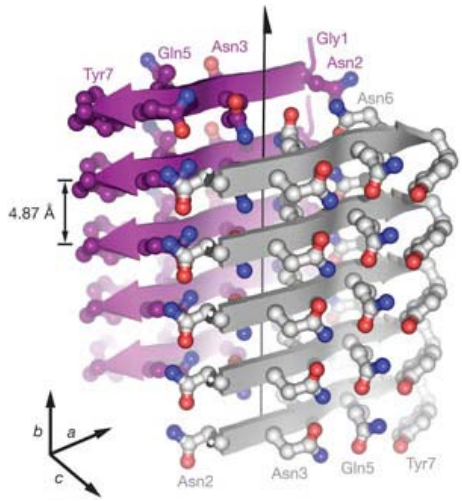
È ormai ampiamente condivisa l'ipotesi che la capacità di formare fibrille amiloidi sia una proprietà comune a molte catene polipeptidiche esistenti in natura (non limitata, dunque, ad una ristretta categoria di proteine degeneri), sotto opportune condizioni di soluzione [CWT$^+$99].
D'altra parte, è dimostrato che la propensione di un polipeptide alla formazione di aggregati amiloidi risulta strettamente legata alla sua **composizione amminoacidica** [WWP$^+$99, HJH$^+$02]. Per inciso, questo è un aspetto di grande importanza per molte tecniche algoritmiche finalizzate alla predizione di aggregazione, come si vedrà estesamente nel seguito.
Vari studi, inoltre, mostrano come la tendenza di una proteina all'amiloidogenesi sia concentrata in particolari regioni della sequenza e, più specificamente, in brevi frammenti peptidici [VZN$^+$04, dlPS04, PECS07],[3] solitamente compresi tra 4 e 7 residui di lunghezza [FP09] . Tali frammenti, isolati dal resto della catena amminoacidica di origine, sono in grado di produrre aggregati fibrillari in vitro: solo per citare alcuni esempi, è questo il caso dell'eptapeptide `GNNQQNY`, ricavato dal prione Sup35 del lievito [NSB$^+$05], e degli esapeptidi `VQIVYK` e `VQIINK`, tratti dalla proteina τ, responsabile del morbo di Alzheimer [VBFB$^+$00].

1.2.2 Struttura cross-β delle fibrille amiloidi

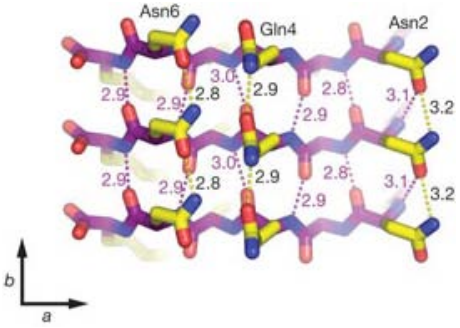
Le fibrille amiloidi (e, ugualmente, le simil-amiloidi) presentano una morfologia tipicamente allungata e priva di ramificazioni (ben distinguibile in Figura 1.5a), con un diametro compreso tra 40 e 130 Å [AK12]. Una fibrilla è costituita da più *protofilamenti*, in numero variabile da 2 a 6, che si intrecciano intorno all'asse longitudinale o si associano lateralmente a formare lunghe strutture simili a nastri [TCMS06]. La conformazione allungata, quasi filiforme, è determinata, a livello atomico, dalla caratteristica disposizione assunta

[3] Nel seguito, ci si riferirà con l'aggettivo **amiloidogenico** a regioni o peptidi caratterizzati dalla tendenza a dare origine a fibrille amiloidi.

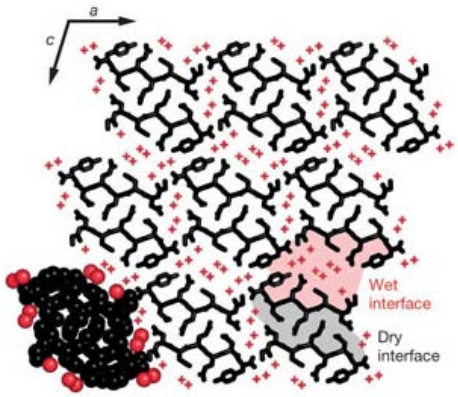
dai peptidi amiloidogenici nell'assemblarsi reciprocamente, nota con il nome di **struttura cross-β** [EG68]. A comporre tale struttura, molto ordinata, è un insieme di **foglietti β**, che corrono paralleli all'asse della fibrilla, con i filamenti β disposti perpendicolarmente all'asse stesso.



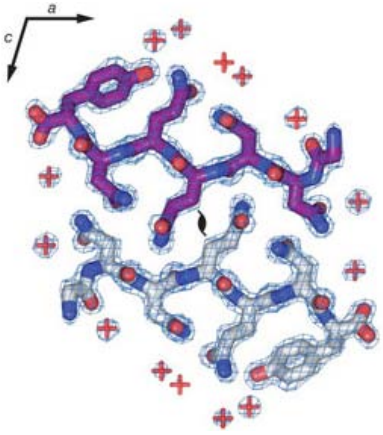
(a) Una coppia di foglietti β, disposti uno di fronte all'altro lungo l'asse della fibrilla. I nastri a freccia indicano i filamenti β, con i gruppi laterali sporgenti.



(b) Vista frontale di tre filamenti β sovrapposti, con le linee punteggiate a segnalare i legami idrogeno (in viola quelli fra catene principali, in giallo quelli fra catene laterali).



(c) Vista dall'alto del cristallo ricavato in vitro dal peptide, con sei file orizzontali di foglietti β (in nero). I segni + rossi rappresentano le molecole d'acqua.



(d) Dettaglio della cerniera sterica tra due foglietti β, vista dall'alto.

Figura 1.6: Struttura cross-β delle fibrille amiloidi originate dall'eptapeptide `GNNQQNY`, tratto dal prione Sup35 del lievito. In viola o bianco-grigio gli atomi di carbonio, in rosso quelli di ossigeno, in blu quelli di azoto. [NSB+05]

Con l'obiettivo, ora, di chiarire l'architettura della conformazione cross-β, si fa riferimento al caso delle fibrille ottenute in vitro a partire dal già citato peptide `GNNQQNY` [NSB$^+$05]. Numerosissime repliche del peptide, sotto forma di filamenti β orientati perpendicolarmente all'asse della fibrilla, si impilano l'uno sull'altro con allineamento reciproco parallelo in registro. Ciascun filamento è connesso ai propri due vicini tramite molteplici legami idrogeno, che coinvolgono sia le catene principali sia le catene laterali (Figura 1.6b). Questa fitta rete di legami idrogeno consolida ciascuno dei foglietti β che, estendendosi lungo l'asse longitudinale della fibrilla, conferiscono a quest'ultima il caratteristico aspetto allungato. A loro volta, i foglietti β si organizzano in coppie, collocandosi uno di fronte all'altro in modo che ogni filamento del primo sia orientato antiparallelamente rispetto al corrispondente filamento del secondo (Figura 1.6a).
Tra foglietti β vicini esistono due tipi di interfaccia (1.6c): la prima è caratterizzata dalla presenza di molecole d'acqua (da cui il nome *interfaccia umida* o *wet interface*), che determinano una separazione piuttosto ampia; la seconda, più sottile, quasi del tutto priva d'acqua (e detta, perciò, *interfaccia secca* o *dry interface*), vede le catene laterali di un filamento compenetrarsi con quelle del filamento di fronte, dando luogo ad interazioni di Van der Waals, a formare una struttura molto solida detta *cerniera sterica* (*steric zipper*) per via della somiglianza visiva dei gruppi laterali con i denti di una cerniera (Figura 1.6d).
L'estrema stabilità della cerniera sterica permette di qualificare l'unità formata da una

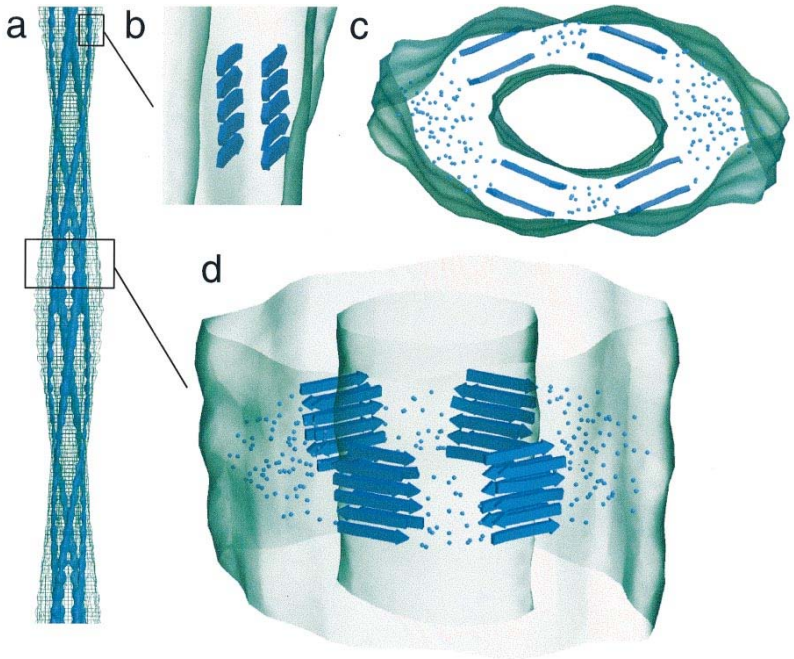


Figura 1.7: Struttura della fibrilla amiloide formata dal dominio SH3 dell'enzima PI3K. Ciascuno dei quattro protofilamenti (in blu) è costituito da una coppia di foglietti β in struttura cross-β. [JñGO$^+$99]

coppia di foglietti β collegati da un'interfaccia secca come elemento strutturale stabile della conformazione cross-β. Al di là dello specifico caso in esame, varie osservazioni inducono a supporre che l'**architettura "a coppie di foglietti β"** rappresenti una caratteristica fondamentale di qualsiasi fibrilla simil-amiloide (Figura 1.7). Inoltre, la solidità della cerniera sterica, nonostante il ridotto numero di residui che la costituiscono, giustifica il fatto che frammenti peptidici, anche molto brevi, siano sufficienti ad innescare l'amiloidogenesi nelle proteine.
In definitiva, l'architettura delle fibrille amiloidi si articola su una **gerarchia a tre livelli**:

1. l'allineamento di molteplici filamenti β, repliche del peptide amiloidogenico, in un foglietto β stabilizzato da legami idrogeno;

2. l'associazione di due foglietti β a formare una coppia, consolidata dalle interazioni di Van der Waals che si generano nell'interfaccia secca tra i due;

3. l'interazione fra coppie di foglietti β da cui trae origine la fibrilla.

1.2.3 Genesi e propagazione delle fibrille amiloidi

Come già accennato, l'amiloidogenesi si concretizza nella conversione di una proteina dalla forma solubile tipica dello stato nativo ad una forma fibrosa insolubile. In anni recenti, sono stati proposti vari modelli per spiegare questo processo di conversione [NE06].
Il modello a **rinaturazione** (*refolding model*) riconosce l'esistenza di due stati nettamente distinti per la proteina amiloidogenica: uno stato nativo ed uno fibrillare. Nella transizione dal primo al secondo, la proteina perde la propria conformazione originaria (*unfolding*) e, in una seconda fase, acquisisce una struttura differente (*refolding*), ricca di strutture β. Secondo alcuni studiosi [FFD01], la conformazione fibrillare è determinata da legami idrogeno coinvolgenti la catena principale, mentre le interazioni tra gruppi laterali hanno un influsso nettamente minore.
In base al modello a **disordine nativo** (*natively disordered model*), una proteina originariamente disordinata assume (tutta o una parte di essa) una struttura ben definita, basata sulla conformazione cross-β. Rientra in questa classe, ad esempio, il β amiloide Aβ all'origine delle placche cerebrali con cui si manifesta il morbo di Alzheimer.
Infine, il modello a **guadagno di interazione** (*gain-of-interaction model*) individua la causa dell'amiloidogenesi in un cambiamento di conformazione limitato ad una ristretta regione della proteina; ciò provoca l'esposizione di una superficie fino ad allora irraggiungibile, che si lega ad una superficie di un'altra proteina, dando origine alla fibrilla. A differenza di quanto accade nei precedenti modelli, lo stato fibrillare conserva inalterata buona parte della struttura della proteina nativa. In questo modello ricade un'ampia varietà di fenomeni, tanto che sono state definite ulteriori sottoclassi.[4] Ad esempio, può accadere che la fibrilla abbia origine dall'accumulo di proteine identiche l'una sull'altra; oppure,

[4]Per una trattazione più completa delle sottoclassi del modello a guadagno di interazione si rimanda a [NE06].

brevi segmenti amiloidogenici, una volta resi accessibili, possono impilarsi in conformazione cross-β, mentre la porzione restante della proteina mantiene la propria struttura nativa; o, ancora, può succedere che le proteine si aggreghino l'una all'altra per mezzo di uno scambio reciproco di domìni, con o senza la formazione di una struttura cross-β.

1.3 Predizione di aggregazione proteica: lo stato dell'arte

Alla base di qualsiasi approccio computazionale che miri a prevedere la tendenza delle strutture proteiche ad effettuare la transizione dal rispettivo stato nativo ad uno stato fibrillare è il seguente assunto, già accennato in precedenza.

Assunzione. *L'**amiloidogenicità** di una proteina (ovvero, la propensione della stessa a formare amiloidi) è codificata nella sua **sequenza amminoacidica**. [AK12]*

Nell'ultimo decennio, la comunità bioinformatica ha proposto un gran numero di metodi algoritmici, che declinano in vari modi la precedente assunzione. Passiamo ora in rassegna i principali, mettendo in evidenza per ciascuno gli aspetti più rilevanti.

1.3.1 Una panoramica dei metodi computazionali oggi disponibili

AggreScan[5] [CSdGA+07] identifica in brevi segmenti, detti *hot spot*, di lunghezza compresa tra 5 e 11 residui, i responsabili dell'aggregazione. A ciascun amminoacido è associato un valore numerico assoluto di attitudine all'aggregazione, precalcolato sulla base di dati sperimentali. Il grado di amiloidogenicità complessivo di un segmento è calcolato come la media delle propensioni individuali dei suoi amminoacidi e l'attribuzione della qualifica di hot spot aggregante al segmento stesso viene stabilita sulla base del confronto della propensione media con un valore soglia prefissato.
FoldAmyloid[6] [GLG10], con un approccio analogo, ricerca le eventuali regioni amiloidogeniche tra i segmenti di almeno 5 residui di lunghezza, impiegando una finestra scorrevole e comparando la media delle propensioni dei singoli amminoacidi all'interno di questa con una soglia predeterminata. Ancora, le attitudini individuali all'aggregazione degli amminoacidi sono precalcolate a partire da informazioni sperimentali su proteine globulari, ma qui sono definite tre distinte scale di valori, che prendono in considerazione vari parametri correlati con l'amiloidogenicità (il numero medio di contatti interatomici per residuo e il numero medio di legami idrogeno intracatena per residuo, distinguendo tra accettori e donatori). In fase di ricerca dei segmenti aggreganti, l'utente può selezionare quale scala impiegare oppure può optare per una scala ibrida, che tiene conto di tutte e tre.

Nel complesso, i metodi appena illustrati tentano di stimare il grado di amiloidogenicità di una regione a partire dalla tendenza intrinseca all'aggregazione dei singoli amminoacidi,

[5] `http://bioinf.uab.es/aggrescan/`

[6] `http://bioinfo.protres.ru/fold-amyloid/oga.cgi`

senza considerare in alcun modo quei fattori strutturali che, nella realtà, hanno un ruolo rilevante nella formazione di aggregati. Le seguenti tecniche, al contrario, tengono conto del fatto che mattoni basilari delle fibrille amiloidi sono i filamenti β.
Zyggregator[7] [TV08] calcola il tasso di aggregazione proprio di ciascun amminoacido sulla base non solo di caratteristiche fisico-chimiche, quali l'idrofobia e la carica elettrica, ma anche della tendenza ad assumere conformazioni di tipo α elica o filamento β. Nel ricercare le regioni aggreganti, vengono prese in considerazione finestre di 7 o più residui consecutivi, tenendo anche conto dell'eventuale influsso elettrostatico esercitato dai residui circostanti sulla propensione all'aggregazione del segmento in esame. Opzionalmente, Zyggregrator può modificare i risultati di predizione in funzione del grado di stabilità della proteina, sulla base dell'assunto che la transizione di un polipeptide allo stato fibrillare, richiedendo la perdita della conformazione nativa, è tanto più probabile quanto più la proteina è instabile.
Tango[8] [FERSS04] si fonda sull'assunzione che sussista una stretta correlazione tra la tendenza di una segmento a formare strutture β e la sua propensione all'aggregazione, che tutti i residui di un segmento β siano sepolti nella parte interna idrofobica dell'aggregato e su considerazioni sull'influenza di fattori elettrostatici sulla tendenza ad aggregare. Esso effettua, in primo luogo, una predizione della struttura secondaria della proteina basata su quattro classi (random coil, curva β, α elica, foglietto β); dopodiché, individua un peptide come amiloidogenico se questo contiene una sequenza lunga almeno 5 residui cui la precedente predizione abbia assegnato una conformazione β.

Se è vero che i filamenti β sono le unità basilari degli aggregati, è altrettanto noto che una fibrilla non potrebbe esistere se, ad un livello più alto, i filamenti non interagissero reciprocamente a formare foglietti β. Un'ulteriore classe di metodi computazionali sfrutta le informazioni su queste interazioni.
BetaScan[9] [BJMC+09] tenta di stimare l'attitudine di segmenti β ad accoppiarsi reciprocamente, concentrandosi principalmente sull'orientamento parallelo, in base alla considerazione che questo tipo di allineamento si presenta con maggior frequenza nei foglietti β costituenti le fibrille amiloidi. In particolare, questo metodo considera sia potenziali filamenti singoli sia coppie di filamenti, calcolando per ciascun candidato un punteggio di probabilità derivante da considerazioni di carattere statistico sulle strutture β di proteine dalla conformazione nota per via sperimentale. Il punteggio complessivo di una sequenza tiene conto tanto della sua probabilità di costituire un filamento β quanto della sua propensione a formare una coppia. BetaScan, inoltre, verifica se l'introduzione di modifiche nelle regioni predette come amiloidogeniche (ad esempio, l'aggiunta o rimozione di residui, oppure lo scorrimento reciproco dei segmenti accoppiati) possa dar luogo a strutture contraddistinte da una maggiore tendenza alla formazione di fibrille. La finestra scorrevole impiegata per il rilevamento delle strutture β varia da 3 a 13 residui di ampiezza.

[7] `http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php`

[8] `http://tango.crg.es/`

[9] `http://groups.csail.mit.edu/cb/betascan/`

Un approccio analogo viene utilizzato da PASTA, che, a sua volta, costituisce il punto di partenza del metodo proposto in questo elaborato. In virtù di ciò, a PASTA sarà dedicato uno spazio maggiore nel seguito.

Tutti i metodi descritti finora si fondano sull'analisi di proteine globulari di struttura nota. Una nuova classe di metodi, invece, trae origine dalle informazioni sperimentali di recente acquisizione sulla struttura tridimensionale caratteristica della struttura cross β delle fibrille simil-amiloidi ottenute in vitro. Tra questi, si citano 3D Profile Method[10] [TSK+06] e Waltz[11] [MSDK+10]. Un'analisi di queste tecniche va oltre gli scopi del presente elaborato: ci si limita qui a citarli, rimandando ai rispettivi articoli per una trattazione completa.

1.3.2 All'origine di PALMO: il predittore PASTA

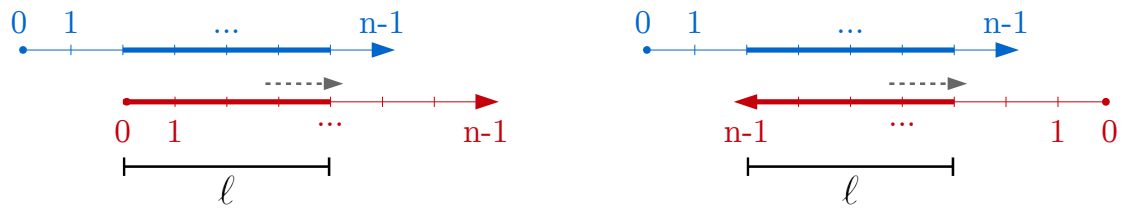
PASTA[12] [TCMS06], acronimo di *Prediction of Amyloid Structure Aggregation*, sviluppato presso l'Università degli Studi di Padova e pubblicato nel 2006, rientra nel filone dei predittori di aggregazione che, sfruttando le informazioni sperimentali sulla conformazione nativa di proteine globulari note per via sperimentale, mirano a prevedere quelle interazioni tra filamenti β di un polipeptide strettamente correlate al consolidamento di strutture cross-β fibrillari.
In via preliminare, PASTA analizza i segmenti β di un *data set* di strutture proteiche globulari risolte sperimentalmente,[13] per associare a ciascuna possibile coppia di amminoacidi una *energia elettrostatica di aggregazione* che ne riflette la frequenza con cui si trova coinvolta in strutture β. Il valore di energia dipende dal numero di occorrenze con cui la rispettiva coppia di amminoacidi compare, legata da legami idrogeno, in filamenti β appaiati. Poiché il conteggio è effettuato separatamente a seconda dell'orientamento reciproco dei filamenti, ogni coppia si vede assegnati, in effetti, due distinti valori di energia di aggregazione, relativi uno all'allineamento parallelo, l'altro all'antiparallelo.
Data in ingresso la sequenza amminoacidica di un polipeptide, PASTA passa in rassegna tutte le possibili coppie di segmenti di lunghezza compresa tra 4 e 23 residui, in orientamento sia parallelo sia antiparallelo. Per cogliere intuitivamente la dinamica dell'algoritmo, si immagini (Figura 1.8a) di sovrapporre in orizzontale due copie identiche della sequenza peptidica d'ingresso, tenendo fissa la prima e facendo scorrere progressivamente la seconda verso destra; tra un passo di scorrimento ed il successivo, si consideri una finestra di ampiezza $\ell \in \{4, \ldots, 23\}$ ad individuare coppie di segmenti di lunghezza crescente; una volta terminata la prima fase, si supponga (Figura 1.8b) di invertire la seconda copia del la sequenza ed effettuare un nuovo scorrimento, così da considerare anche gli allineamenti

[10] `http://services.mbi.ucla.edu/zipperdb/intro`

[11] `http://waltz.switchlab.org/`

[12] `http://biocomp.bio.unipd.it/pasta/`

[13] Il data set impiegato è il *Top500 Database*, per ulteriori dettagli sul quale si rimanda alla Sezione 2.1, pag. 16.



(a) Caso parallelo.

(b) Caso antiparallelo.

Figura 1.8: Rappresentazione grafica dell'algoritmo di PASTA.

antiparalleli. Ad ogni accoppiamento fra segmenti, PASTA attribuisce una propensione di aggregazione complessiva, data dalla somma dell'energia di ciascuna coppia di residui appaiati. In uscita si ottiene l'elenco delle 20 coppie di segmenti di più elevata attitudine all'aggregazione, con l'esplicita indicazione del rispettivo orientamento reciproco, ordinate per valori di propensione descrescenti.
Il collaudo di PASTA su polipeptidi dalle proprietà amiloidogenetiche note mette in evidenza prestazioni di predizione di buon livello, soprattutto in termini di specificità,[14] tanto nella classificazione di brevi frammenti quanto nel rilevamento di regioni amiloidogeniche in sequenze proteiche complete,[15] a conferma della correttezza delle assunzioni di base e della validità complessiva dell'approccio adottato. Tuttavia, l'algoritmo di ricerca degli accoppiamenti utilizza una procedura esaustiva di scarsa efficienza. E, soprattutto, il fatto di considerare singole coppie di residui, indipendentemente dal contesto circostante, limita fortemente la ricchezza informativa del modello che ne deriva.
Proprio dalla constatazione di questi limiti e dei margini di miglioramento che il loro superamento prometteva di conseguire, è nata l'idea di elaborare una tecnica computazionale che, pur basata su assunzioni biochimiche analoghe, adottasse un approccio algoritmico più efficiente ed un modello più sofisticato e aderente alla realtà. Il conseguente lavoro di progettazione e sviluppo si è, infine, concretizzato nel metodo presentato in questo elaborato: PALMO, *Protein Aggregation Likelihood and Mutation Optimization*.

[14] Per una trattazione sulle misure impiegate nella valutazione delle prestazione di un predittore, si rimanda alla Sezione 2.7, pag. 39.

[15] Un'analisi completa dei risultati di predizione conseguiti da PASTA è presentata in [TCMS06].

Capitolo 2

Dati e metodi

La Figura 2.1 fornisce una panoramica sulla struttura modulare di Palmo.

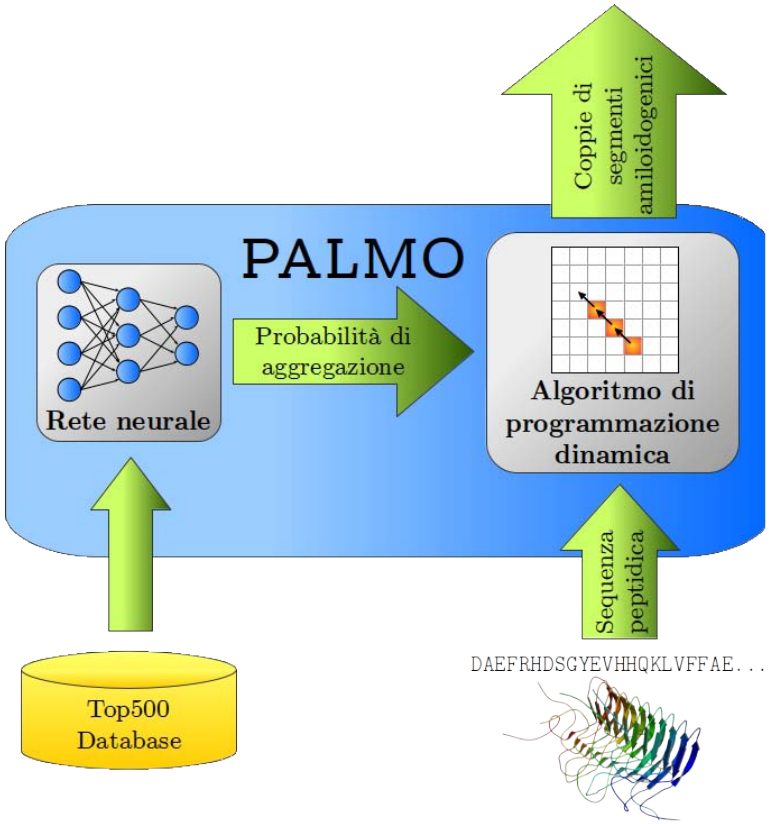


Figura 2.1: Schema a blocchi di Palmo.

Esso si fonda sull'interazione fra due componenti algoritmiche fondamentali:

1. una **rete neurale** (Sezione 2.4, pag. 23);
2. un **algoritmo di programmazione dinamica** (Sezione 2.5, pag. 28).

La rete neurale, addestrata sulla base di un *database* di strutture proteiche annotate (Sezione 2.1, pag. 16), è in grado di prevedere un valore di **probabilità di aggregazione** per una qualsiasi coppia di amminoacidi. Queste probabilità di aggregazione fungono da valori di ingresso all'algoritmo di programmazione dinamica, che, data unicamente la **sequenza primaria** di una proteina, identifica, all'interno di quest'ultima, gli **accoppiamenti fra segmenti** responsabili di innescare l'amiloidogenesi, distinguendone l'orientamento (parallelo o antiparallelo) ed assegnando a ciascuno un **punteggio** numerico che ne riflette la tendenza all'aggregazione. In uscita, PALMO fornisce l'elenco degli accoppiamenti amiloidogenici, in ordine di punteggio decrescente.

Nel seguito del capitolo, si procede ad un'analisi dettagliata delle componenti appena delineate; non prima, però, di avere descritto gli insiemi di dati impiegati sia nello sviluppo sia nella verifica di PALMO.

2.1 Training set

Tanto l'addestramento del sistema di apprendimento automatico alla base del metodo qui proposto, quanto la valutazione quantitativa dell'accuratezza dello stesso si sono scontrati con la limitata disponibilità di dati sperimentali sugli amiloidi. In concreto, i *data set* esistenti, tanto di peptidi isolati classificati sulla base di proprietà di aggregazione, quanto di intere strutture amiloidogeniche annotate, scarseggiano e sono perlopiù di dimensioni piuttosto ridotte.
La costruzione del *training set* per l'addestramento della rete neurale ha aggirato questo ostacolo in base ad una fondamentale considerazione.

Fatto. *I principali elementi strutturali degli aggregati amiloidi sono i* ***foglietti β****, disposti a formare una struttura detta* ***cross-β****.*[1]

È noto che la conformazione a foglietto β, a sua volta, è stabilizzata dai legami idrogeno tra le catene principali di filamenti β adiacenti. I legami idrogeno, dunque, rivestono un ruolo determinante nella formazione dei foglietti β tanto nelle proteine in stato nativo quanto nella struttura cross-β degli aggregati fibrillari. Questo fatto ha consentito di attingere alla vasta mole di dati sperimentali sulle proteine globulari, di molte delle quali sono note la struttura secondaria e, soprattutto, informazioni sui legami idrogeno che ne stabilizzano la conformazione.
In particolare, il training set è stato costruito a partire dal *Top500 Database*,[2] messo a punto dal *Richardson Lab* presso la *Duke University*.
Il Top500 Database consiste in una selezione di 500 strutture proteiche tridimensionali provenienti dal *Protein Data Bank*,[3] scelte sulla base di criteri comprendenti una buona

[1] Si veda la Sezione 1.2.2, pag. 7.

[2] `http://kinemage.biochem.duke.edu/databases/top500.php`

[3] `http://www.rcsb.org/pdb`

risoluzione (non peggiore di 1.8 Å), una bassa omologia e un'alta qualità in termini di accuratezza strutturale. Peculiarità del data set, cruciale per i nostri scopi, è il fatto che le strutture molecolari siano corredate di informazioni sugli atomi di idrogeno (coordinate ortogonali, fattore di occupazione, fattore di temperatura ed eventuale carica elettrica).

2.1.1 Costruzione del training set

Ciascuna struttura proteica, sotto forma di un file in formato `pdb`, è elaborata attraverso il programma DSSP (*Define Secondary Structure of Proteins*,[4] [KS83, JtBK$^+$11]). L'algoritmo, data una struttura proteica tridimensionale, è in grado di calcolarne la più probabile assegnazione di struttura secondaria, associando a ciascun residuo un codice alfabetico che identifica il tipo di conformazione di cui esso fa parte.[5] Nel calcolo, DSSP sfrutta sia le coordinate atomiche sia informazioni sui legami idrogeno intracatena;[6] di questi ultimi calcola l'energia elettrostatica in kcal/mol, distinguendo tra amminoacidi accettori e donatori.
A partire dall'output di DSSP, si prendono in considerazione i soli filamenti β coinvolti nella formazione di foglietti β. Ciascuno di tali filamenti può essere legato a uno o due suoi simili (a seconda che si trovi alle estremità o all'interno del foglietto di appartenenza), qui chiamati *partner*, con orientazione relativa parallela o antiparallela. Com'è noto, la struttura a foglietto β è stabilizzata dai legami idrogeno che sussistono tra coppie di amminoacidi (un donatore ed un accettore) appartenenti a filamenti adiacenti (Figura 2.2).
Per ogni residuo componente ad un filamento di un foglietto β, si considerano qui al più quattro legami idrogeno, uno per ciascuna delle seguenti classi, a seconda del ruolo assunto dall'amminoacido e dall'orientazione relativa del filamento cui appartiene il residuo con il quale è instaurato il legame stesso:

- *antiparallelo accettore*;
- *antiparallelo donatore*;
- *parallelo accettore*;
- *parallelo donatore*.[7]

In sintesi, di ciascuna proteina del data set di partenza si ricavano la sequenza primaria e la codifica di struttura secondaria, secondo una classificazione semplificata a tre sole "superclassi" (Tabella 2.1, pag. 18). I soli amminoacidi coinvolti in foglietti β sono associati

[4] `http://swift.cmbi.ru.nl/gv/dssp`

[5] `H`: α elica (α-*helix*); `B`: residuo in un ponte β (β-*bridge*) isolato; `E`: filamento β (β-*strand*) esteso; `G`: elica 3_{10} (3_{10}-*helix*); `I`: elica π (π-*helix*); `T`: curva (*turn*) stabilizzata da legame idrogeno; `S`: ansa (*bend*).

[6] Si veda la Sezione 1.1.3, pag. 4.

[7] Più in dettaglio: *a*) i legami idrogeno più deboli (ovvero, con energia elettrostatica di valore superiore a -0.5 kcal/mol) sono scartati; *b*) nel caso di più legami idrogeno di una stessa classe, se ne considera il più forte (cioè, quello con l'energia massima in valore assoluto); *c*) si tollerano legami idrogeno tra amminoacidi di cui uno sia esterno (di al più una posizione) agli estremi del filamento partner.

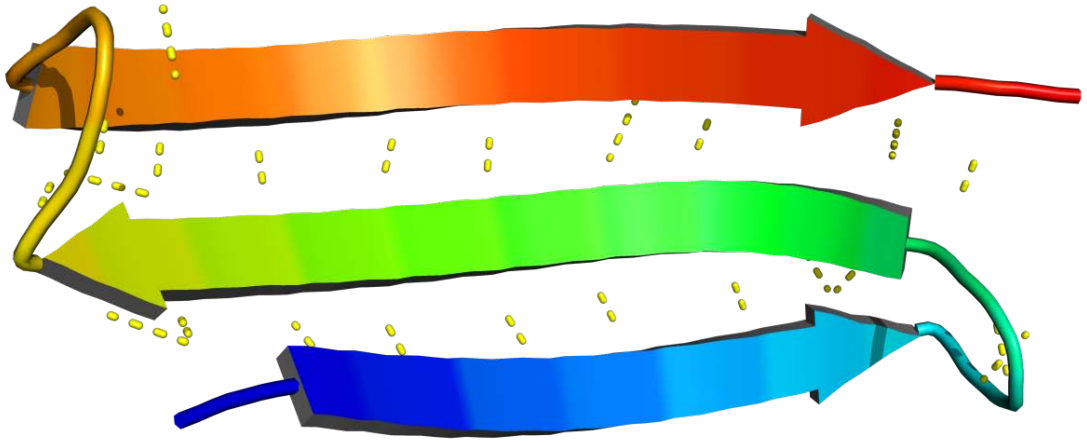


Figura 2.2: Un filamento β (al centro, in verde) con due filamenti partner in orientazione antiparallela. Evidenziati in giallo i legami idrogeno che stabilizzano il foglietto β.

Tabella 2.1: Classificazione semplificata di struttura secondaria.

Superclasse	**Identificatore**	**Classi DSSP unificate**
Elica (*helix*)	`H`	`H`, `G`, `I`
Filamento (*strand*)	`S`	`E`, `B`
Coil	`C`	`T`, `S`

all'orientazione relativa (parallela o antiparallela) degli (al più) due filamenti partner e, per ciascuno degli (al più) quattro legami idrogeno di cui sopra, alla rispettiva energia elettrostatica e all'indice sequenziale dell'amminoacido legato.
Dall'insieme di dati così organizzato sono ricavati gli esempi destinati a comporre il training set, sotto forma, in prima approssimazione,[8] di **coppie di residui** di una stessa sequenza proteica, uniti o meno da un legame idrogeno.
Le coppie di amminoacidi a_i, a_j che contribuiscono a stabilizzare una conformazione a foglietto β tramite un legame idrogeno rappresentano gli esempi **positivi**. I **negativi** consistono, invece, negli accoppiamenti di residui fra cui si assume non si instaurino legami idrogeno, perché appartenenti a combinazioni di strutture secondarie del tipo *elica-elica*,

[8] Si veda la Sezione 2.1.3, pag. 19.

elica-filamento, *elica-coil*, *filamento-coil*, *coil-coil* o viceversa.

2.1.2 Bilanciamento del training set

L'accentuata preponderanza numerica degli esempi negativi rispetto ai positivi comporterebbe una sproporzione altrettanto netta tra i valori di probabilità di legame prodotti dalla rete neurale, oltre ad una minore velocità di convergenza dell'algoritmo di addestramento. È emersa, dunque, l'esigenza di fornire alla procedura di apprendimento automatico un training set meglio bilanciato, includendovi solo una parte degli esempi negativi.
Più in dettaglio, data una sequenza proteica, sia S l'insieme dei residui appartenenti a filamenti β (in breve, *residui-β*), H l'insieme dei *residui-elica*, C l'insieme dei *residui-coil* e $s := |S|$, $h := |H|$, $c := |C|$ le rispettive cardinalità.[9] L'intero insieme di esempi positivi ricavabili da essa è incluso nel training set, ovvero tutti i p accoppiamenti di residui-β uniti da un legame idrogeno. Vengono, poi, selezionati casualmente $\hat{h}$ residui-elica da H e $\hat{c}$ residui-coil da C, a formare due sottoinsiemi $\hat{H}$ e $\hat{C}$, con

$$\hat{h} := \begin{cases} \lfloor \sqrt{p} \rfloor & \text{se } h > \sqrt{p} \text{ e } p > 0 \\ h & \text{se } h \leq \sqrt{p} \text{ e } p > 0 \\ \lfloor \frac{h}{10} \rfloor & \text{se } p = 0 \end{cases}, \quad \hat{c} := \begin{cases} \lfloor \sqrt{p} \rfloor & \text{se } c > \sqrt{p} \text{ e } p > 0 \\ c & \text{se } c \leq \sqrt{p} \text{ e } p > 0 \\ \lfloor \frac{c}{10} \rfloor & \text{se } p = 0 \end{cases}$$

Dai sottoinsiemi ridotti $\hat{H}$ e $\hat{C}$ e dall'intero insieme S si attinge nel comporre tutte le possibili coppie negative, secondo le combinazioni di classi di struttura secondaria descritte in precedenza. In questo modo, l'ammontare di esempi negativi è ridotto ad una quantità che è funzione del numero dei positivi.[10]

2.1.3 Rilevanza del contesto

Quanto affermato finora non ha tenuto conto, per semplicità, di un fatto rilevante.

Fatto. *La probabilità che si instauri un legame idrogeno, così come le caratteristiche di questo, non dipende solo dalla coppia di residui direttamente coinvolti, ma è influenzata anche dal* ***contesto locale****, ovvero dalle porzioni di sequenza primaria immediatamente circostanti.*

E, se si considera che proprio i legami idrogeno sono il principale fattore di stabilizzazione della conformazione a foglietto β (e non solo), è evidente come il contesto abbia un influsso non trascurabile anche sulla struttura secondaria delle proteine.

---

[9] Ad esempio, data la seguente sequenza peptidica e la relativa assegnazione di struttura secondaria "semplificata",

```
VPGFTPRLAILQVGNRDDSNLYINVKLKAAEEIGIKATHIKLPRTTTESEVMKYITSLNEDSTVHGFLVQLPLDSENSINTEEVINAIA
.CC...SSSSSSSC..HHHHHHHHHHHHHHHHHC.SSSSSSS.CC..HHHHHHHHHHHHH.CC..SSSS.CC...CC...HHHHHHC..
```

si ha $s = 18$, $h = 36$, $c = 13$.

[10] In base a semplici calcoli combinatori, nel caso peggiore (ovvero, se h, $c > \sqrt{p} > 0$, quindi $h = c := \lfloor \sqrt{p} \rfloor$), il numero di esempi negativi è al più pari a $4p + 4s\sqrt{p}$.

In base a queste considerazioni, vari algoritmi di predizione di struttura secondaria definiscono finestre di varie dimensioni per catturare un più ricco ventaglio di caratteristiche dell'input (ad esempio, PSIPRED, [RS93, Jon99]). Un approccio simile è adottato nel metodo qui proposto.
In generale, si definisce *contesto c* il numero di residui precedenti e seguenti l'amminoacido centrale, presi in considerazione a formare una *finestra* di lunghezza $w = 1 + 2c$. Data una sequenza proteica $\mathcal{S}$ di lunghezza n, dove a_k identifica l'amminoacido in posizione $k \in \{0,1,\ldots,n-1\}$ la finestra $\mathcal{W}_i^c$ centrata nel residuo a_i è evidenziata di seguito

$$\mathcal{S} := a_0 a_1 \ldots a_{(i-c-1)} \overbrace{a_{(i-c)} \ldots a_{(i-1)} a_i a_{(i+1)} \ldots a_{(i+c)}}^{\mathcal{W}_i^c} a_{(i+c+1)} \ldots a_{(n-2)} a_{(n-1)}$$

Nel nostro caso, l'informazione sulla sussistenza o meno di un legame idrogeno e gli eventuali dati supplementari sono associati non al singolo residuo, ma alla finestra centrata in esso. Così, il training set risulta costituito, in realtà, non da coppie di amminoacidi, bensì da accoppiamenti fra segmenti. Analogamente, in fase di predizione, ciascuna probabilità di aggregazione sarà calcolata in funzione di un accoppiamento di w-uple della sequenza primaria d'ingresso e poi attribuita alla coppia dei rispettivi residui mediani.
In linea di principio, quanto maggiore è la dimensione della finestra, tanto più è ampia la gamma di informazioni catturabili e ricco il modello che ne deriva. D'altra parte, vari studi sperimentali [PECS07, VZN$^+$04, TE09] suggeriscono come il processo di aggregazione fibrillare di proteine native sia innescato da brevi segmenti peptidici, di lunghezza compresa, solitamente, fra 6 e 8 residui. Inoltre, un valore eccessivo rischierebbe di dare luogo ad *overfitting*[11]. Nel caso del metodo qui proposto, si è ritenuto opportuno optare per un contesto di dimensione $c = 2$, corrispondente a una finestra di $w = 5$ residui. Addestramento e predizione, quindi, si basano su accoppiamenti fra quintuple peptidiche. In realtà, la dimensione del contesto può ridursi a 1 o 0 per tenere conto delle situazioni degeneri che si verificano agli estremi della sequenza proteica.[12] Ovviamente, l'adozione di differenti lunghezze di contesto ha richiesto l'addestramento di un apposito modello statistico per ciascuna, calcolato da un training set di coppie di w-uple, ricavate dalle stesse strutture del Top500 Database.

[11] Si dice *overfitting* (o *adattamento eccessivo*) il fenomeno per cui un algoritmo di apprendimento automatico costruisce un modello eccessivamente adattato a caratteristiche specifiche del training set, quindi incapace di generalizzare adeguatamente, risultando così poco accurato quando, in fase di predizione, gli vengono forniti dati sconosciuti.

[12] Data una sequenza peptidica completa $\mathcal{S} := a_0 a_1 a_2 a_3 \ldots a_{(n-4)} a_{(n-3)} a_{(n-2)} a_{(n-1)}$, si pone

- $c = 0$ e $w = 1$ in corrispondenza dei soli residui estremi a_0 e $a_{(n-1)}$, per i quali $\mathcal{W}_0^0 = [a_0]$, $\mathcal{W}_{(n-1)}^0 = [a_1]$;
- $c = 1$ e $w = 3$ per gli amminoacidi ad una posizione di distanza dagli estremi, ovvero a_1 e $a_{(n-2)}$, con $\mathcal{W}_1^1 = [a_0 a_1 a_2]$, $\mathcal{W}_{(n-2)}^1 = \left[a_{(n-3)} a_{(n-2)} a_{(n-1)}\right]$;
- $c = 2$ e $w = 5$ per tutti i restanti residui a_i, $i \in \{2, 3, \ldots, n-3\}$, nei quali casi $\mathcal{W}_i^2 = \left[a_{(i-2)} a_{(i-1)} a_i a_{(i+1)} a_{(i+2)}\right]$.

2.2 Test set di frammenti peptidici impiegato nella valutazione dell'accuratezza di classificazione

L'accuratezza di classificazione di Palmo, ovvero la misura della sua capacità di riconoscere sequenze proteiche particolarmente soggette alla formazione di fibrille amiloidi, è stata valutata utilizzando uno dei pochi insiemi di peptidi classificati ricavato dalla letteratura. In particolare, ci si è basati sul data set costruito dagli sviluppatori del predittore di aggregazione Tango [FERSS04].
L'insieme è composto da 177 peptidi corrispondenti a frammenti di 21 proteine (proteina τ, β-amiloide, α-sinucleina, acil-fosfatasi, β2-microglobulina, 434 cro-repressore, mioglobina di capodoglio, mioemeritrina, plastocianina di *Phaseolus vulgaris*, inibitore della tripsina di pancreas bovino, proteina ribosomiale L9, glutatione S-transferasi, spettrina, Ada, Ara, Com-A, Che-Y, flavodossina, P21-Ras, proteina PL B1, proteina G).[13] La propensione all'aggregazione di ciascun peptide è stata stimata attraverso metodi sperimentali, in base ai quali 65 peptidi sono risultati aggreganti, 112 non aggreganti.

2.3 Test set di proteine complete impiegato nella valutazione dell'accuratezza di individuazione di regioni amiloidogeniche

Per quanto riguarda, invece, l'individuazione di regioni propense all'aggregazione all'interno di sequenze proteiche intere, si è optato per il data set presentato in [OWL$^+$11] per la valutazione di AmyloidMutants,[14] un ulteriore predittore di aggregazione.
Si tratta di cinque fra le proteine amiloidogeniche presenti in natura (*wild type*) più studiate, sia patogeniche sia funzionali. In dettaglio:

β-amiloide (Aβ) Di questa proteina, riconosciuta come la causa di malattie neurodegenerative, sono note varie isoforme (Aβ_{1-40}, Aβ_{1-42}, A$\beta_{1-40/D23N}$, A$\beta_{1-40/E22Q}$) e sottosequenze (Aβ_{16-22}, Aβ_{11-25}) in grado di dare origine ad una vasta gamma di strutture fibrillari. Osservazioni sperimentali, basate su NMR, scambio idrogeno-deuterio (*H-D exchange*) e analisi mutazionale, hanno condotto all'elaborazione di due distinti modelli tridimensionali [PIB$^+$02, LRA$^+$05], il secondo dei quali rivela una struttura β solenoidale con due foglietti β, collegati da una piegatura, per catena (Figura 2.3, pag. 22). In particolare, è stata sottoposta a test l'isoforma A$\beta_{1\text{-}42}$, responsabile delle aggregazioni amiloidi alla base del morbo di Alzheimer.

HET-s Il prione HET-s$_{218-289}$ del fungo filamentoso *Podospora anserina* è l'amiloide più complesso di cui sia stata ricavata la struttura tridimensionale [WLVM$^+$08]. L'amiloide esibisce una conformazione a β elica, con due giri per ciascuna catena e

[13] Per l'elenco completo dei frammenti peptidici, si rimanda all'Appendice A, Tabella A.1, pag. 75.

[14] `http://amyloid.csail.mit.edu/`

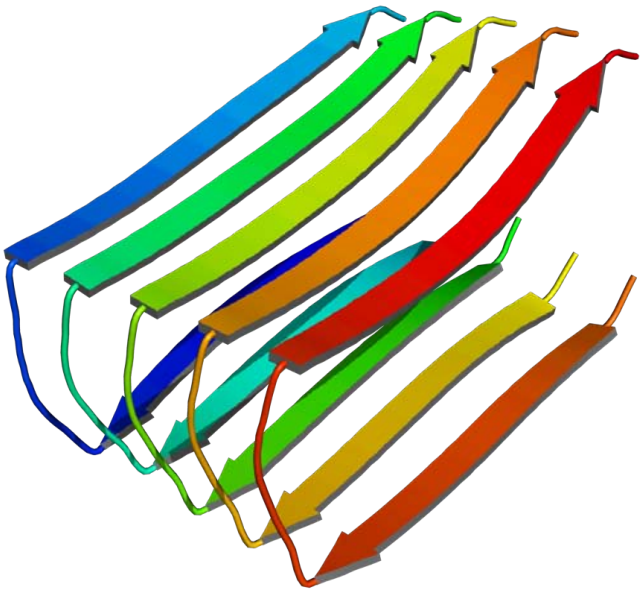


Figura 2.3: Struttura 3D di Aβ_{1-42}. [LRA$^+$05]

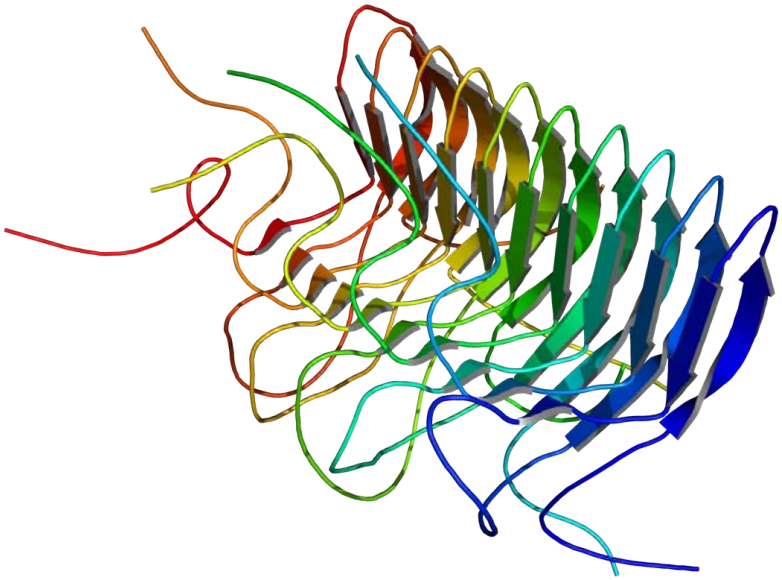


Figura 2.4: Struttura 3D di HET-s$_{218-289}$. [WLVM$^+$08]

quattro foglietti β, organizzati in due coppie separate da una curva (Figura 2.4). Si è presa in considerazione anche una lontana omologa, la proteina HET-s di *Fusarium*

graminearum (FgHET-s), che, benché abbia un grado di similarità di sequenza pari ad appena 38%, si è rivelata possedere una struttura β solenoidale analoga alla prima [WZS$^+$10].

Amilina (IAPP) Denominata anche *islet amyloid polypeptide*, questa proteina (37 residui) è all'origine degli amiloidi che si formano nel pancreas dei pazienti di diabete mellito tipo 2. Studi sperimentali hanno svelato una struttura 3D del tutto analoga a quella del β-amiloide [LYLT07]. Un diverso modello propone una struttura β sinusoidale con tre foglietti β per catena [KAS05].

α-sinucleina I modelli 3D elaborati concordano nell'assegnare all'α-sinucleina (140 residui), proteina responsabile degli amiloidi all'origine del morbo di Parkinson, una struttura basata su cinque foglietti β paralleli per catena [HHB$^+$05, VCL$^+$08].

Proteina τ Questo amiloide (441 residui) gioca un ruolo nello sviluppo di aggregati proteici alla base di molte malattie neurodegenerative, dette taupatie, tra cui il morbo di Alzheimer. La struttura della proteina τ si caratterizza per la presenza di vari foglietti β [MBK$^+$09], tra i quali due esapeptidi cruciali per l'avvio del processo di aggregazione: ^{274}`VQIINK`^{279} e ^{305}`VQIVYK`^{310} [VBFB$^+$00].

Le sequenze primarie delle proteine amiloidi appena descritte, con evidenziate le regioni sperimentalmente riconosciute come propense alla formazione di aggregati, sono riportate per intero in Appendice A, Tabella A.3, pag. 82.

2.4 Rete neurale

2.4.1 Reti neurali: un'introduzione generale

Nell'ambito dell'apprendimento automatico, una *rete neurale artificiale* (ANN, *artificial neural network*) è un modello matematico di calcolo che, ispirato alle interconnessioni neuronali del sistema nervoso centrale umano, può apprendere grandi quantità di informazioni complesse. In analogia con la controparte biologica, una rete neurale artificiale consiste in un insieme di unità computazionali o *neuroni*, connesse fra loro da *collegamenti sinaptici*; il processo di apprendimento avviene a partire da esempi e si concretizza attraverso una opportuna alterazione dell'intensità delle connessioni.
Ciascuna unità computazionale (Figura 2.5) ha per ingressi le uscite di altri neuroni o di una sorgente esterna. Formalmente, il generico nodo i riceve k valori d'ingresso y_j, $j \in \{1, \ldots, k\}$ attraverso altrettanti archi orientati, ciascuno contraddistinto da un peso $w_{i,j} \in \mathbb{R}$; un ulteriore parametro $b_i \in \mathbb{R}$, detto *bias*, rappresenta la soglia di attivazione del neurone.[15] L'unità calcola in uscita una funzione f della somma pesata degli ingressi, $y_i = f\left(\sum_{j=0}^{k} w_{i,j} y_j\right) = f\left(b_i + \sum_{j=1}^{k} w_{i,j} y_j\right)$, dove f è definita *funzione di attivazione*.
Un insieme di unità computazionali disposte a strati (un livello di ingresso, uno di uscita e uno o più strati intermedi, detti *nascosti*), tali che i nodi di uno strato instaurino

[15] Il bias, in alternativa, può essere visto come il peso $w_{i,0} = b_i$ associato ad un ingresso fittizio $y_0 = 1$.

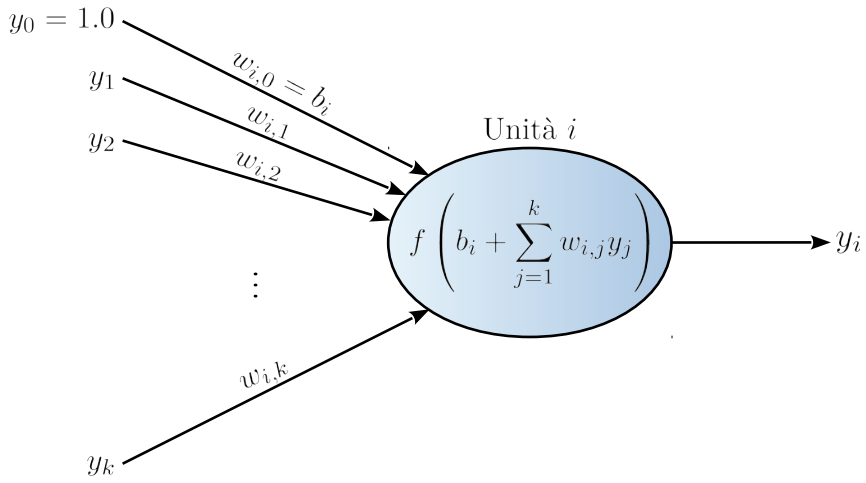


Figura 2.5: Una singola unità di una rete neurale.

connessioni dirette con le sole unità del livello immediatamente successivo, forma una *rete neurale multistrato alimentata in avanti* (*feed-forward multilayer neural network*) o, in breve, *percettrone multistrato* (MLP, *multilayer perceptron*, Figura 2.6). È dimostrato che

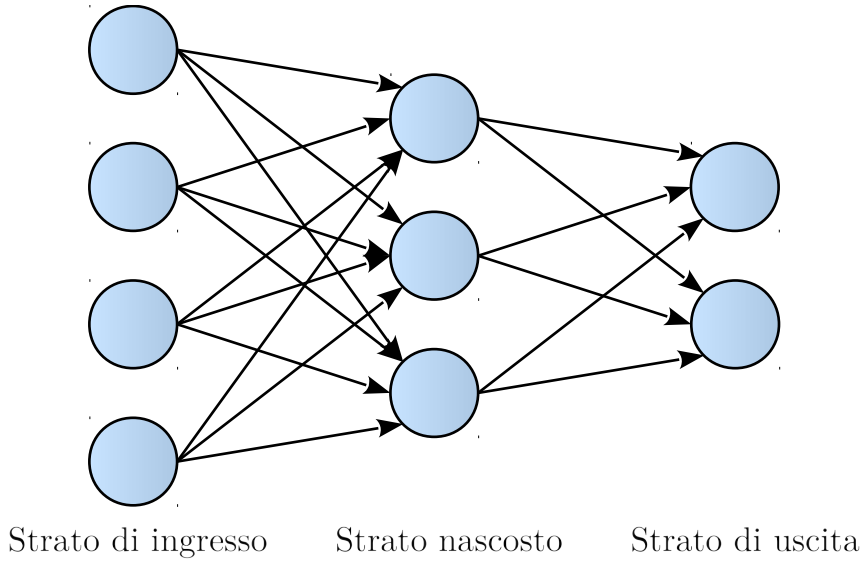


Figura 2.6: Esempio di percettrone multistrato con un livello nascosto.

un percettrone multistrato con un solo livello nascosto è in grado di approssimare una qualsiasi funzione non lineare (*teorema di approssimazione universale*, [HSW89]).

2.4.2 La rete neurale di PALMO

PALMO si basa proprio su un percettrone multistrato (Figura 2.7, pag. 25). L'unico livello

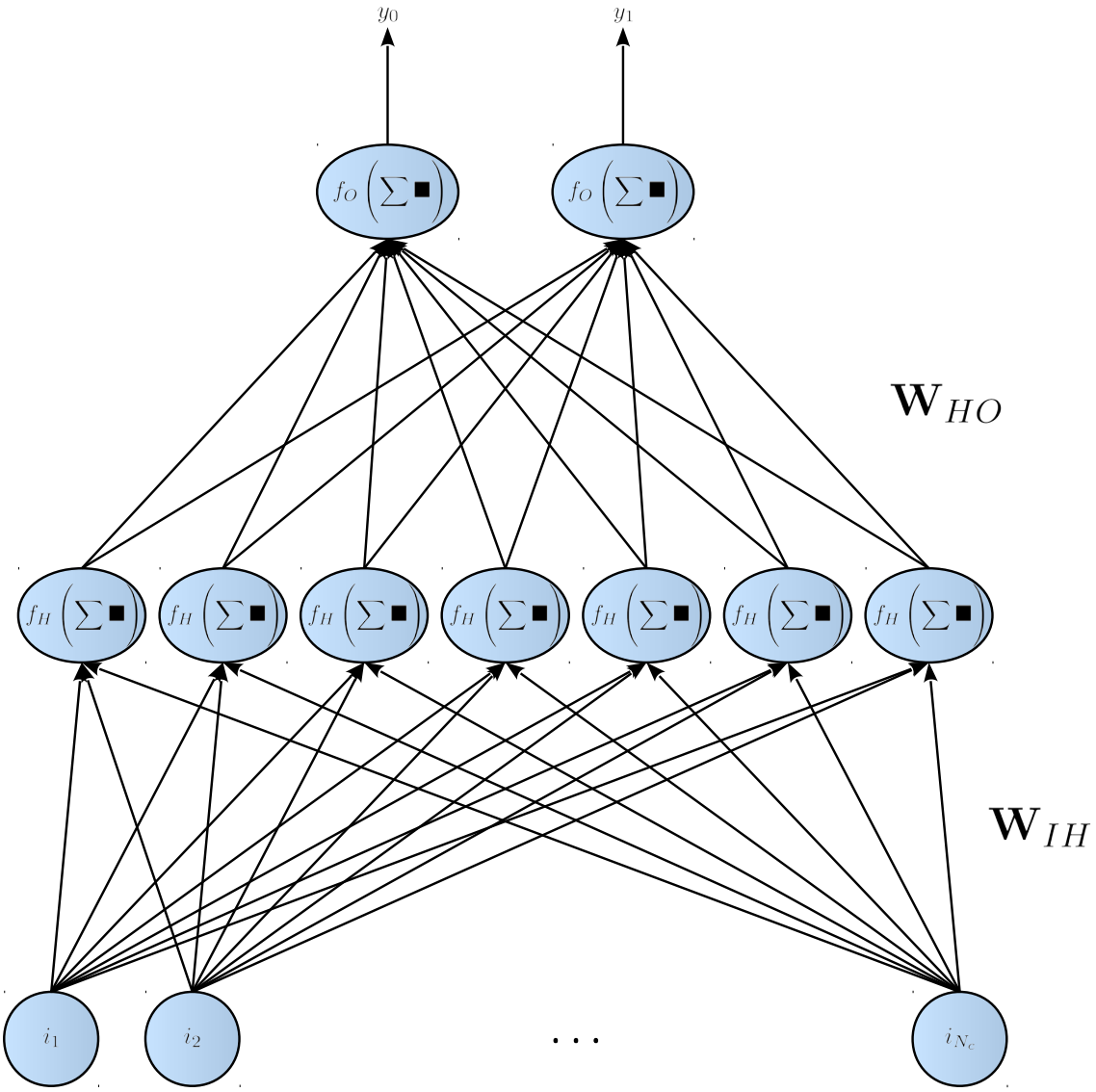


Figura 2.7: Schema del percettrone multistrato di PALMO. Le funzioni di attivazione sono, rispettivamente, $f_H(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$ per le unità dello strato nascosto e $f_O(x) = \frac{1}{1+e^{-x}}$ per i neuroni del livello di uscita. $\mathbf{W}_{IH}$ e $\mathbf{W}_{HO}$ sono le matrici dei pesi relativi agli archi, rispettivamente, dallo strato di ingresso a quello nascosto e da questo al livello di uscita. L'uscita y_0 indica la probabilità di non-formazione di legame idrogeno, y_1 la probabilità di formazione di legame idrogeno.

nascosto contiene 7 neuroni con funzione di attivazione $f_H(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$, mentre lo strato di uscita consiste in 2 neuroni con funzione di attivazione sigmoidale $f_O(x) = \frac{1}{1+e^{-x}}$ (Figura 2.8).
Obiettivo della rete neurale di PALMO è apprendere un *potenziale di legame idrogeno* fra

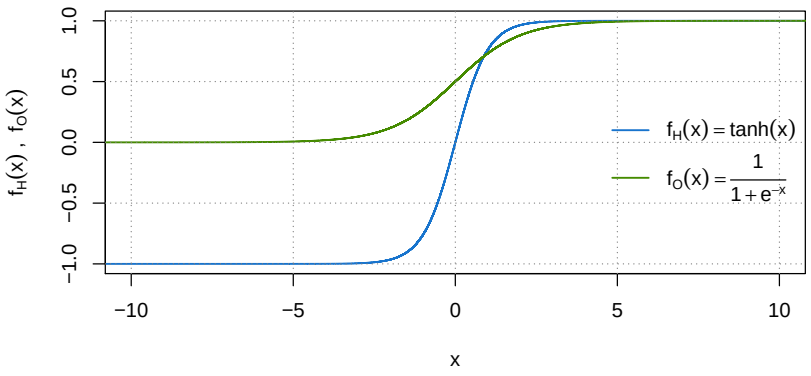


Figura 2.8: Grafico delle funzioni di attivazione $f_H(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$ e $f_O(x) = \frac{1}{1+e^{-x}}$, impiegate dalle unità dei livelli, rispettivamente, nascosto e di uscita del percettrone multistrato di PALMO.

coppie di amminoacidi, prendendo in considerazione l'intera finestra di ampiezza $w = 1 + 2c$ centrata nei residui in esame, dove $c \in \{0, 1, 2\}$ è la dimensione del contesto.[16] Il percettrone multistrato relativo al contesto c è attivato da N_c neuroni di ingresso a calcolare, in uscita, due valori di probabilità: una probabilità di formazione di legame idrogeno y_1 e una di non-formazione di legame idrogeno y_0.
Il training set, costruito secondo la procedura descritta in Sezione 2.1, pag. 16, contiene le informazioni sui legami idrogeno ricavate dalle strutture proteiche del Top500 Database. Ogni coppia di residui tratta da queste ultime è associata ad un valore obiettivo (*target*) binario, dove 1 segnala l'esistenza, 0 l'assenza, di un legame idrogeno.
Un singolo residuo è rappresentato sotto forma di un vettore di 20 elementi, ciascuno associato univocamente ad un amminoacido standard e inizializzato a 0, in cui si pone a 1 la sola cella corrispondente al residuo in esame (Figura 2.9, pag. 27). Nel caso $c \geq 1$, la concatenazione di $w = 1 + 2c$ vettori di questo tipo consente di codificare anche le informazioni sul contesto; un vettore composto di soli 0 segnala le situazioni degeneri in cui la finestra di lunghezza w si estende oltre gli estremi della sequenza proteica di partenza. Ciascun elemento del vettore concatenazione corrisponde ad una distinta unità di ingresso della rete neurale. Pertanto, per un contesto di dimensione c, lo strato di ingresso è composto da un numero di neuroni pari a

$$N_c = 20 \cdot 2w = 20 \cdot 2\,(1 + 2c) = \begin{cases} 40 & \text{se } c = 0 \\ 120 & \text{se } c = 1 \\ 200 & \text{se } c = 2 \end{cases}$$

[16] Si veda la Sezione 2.1.3, pag. 19.

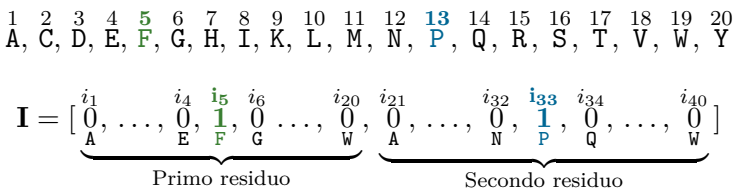


Figura 2.9: Esempio di vettore di ingresso al percettrone multistrato di PALMO, corrispondente alla coppia di residui (F, P) nel semplice caso $c = 0$. La seconda riga riporta la codifica ad una lettera dei 20 amminoacidi standard. Ad esempio, il primo amminoacido della coppia, fenilalanina, ha per simbolo F, quinto codice nell'ordine alfabetico: esso, pertanto, è rappresentato sotto forma di un vettore di lunghezza 20, con un unico 1 in quinta posizione e i restanti elementi posti a 0. Il vettore $\mathbf{I} \in \{0, 1\}^{40}$ di ingresso consiste nella concatenazione dei due vettori corrispondenti ai residui accoppiati.

Il percettrone multistrato riceve una coppia di vettori $(\mathbf{I}, \mathbf{t})$, dove

- $\mathbf{I} \in \{0, 1\}^{N_c}$ è il vettore di ingresso nella forma descritta sopra;
- $\mathbf{t} = (t_0, t_1) \in \{0, 1\}^2$ è il vettore obiettivo, con $\mathbf{t} = (0, 1)$ ad indicare la sussistenza di un legame idrogeno, $\mathbf{t} = (1, 0)$ l'assenza.

I pesi della rete neurale sono ottimizzati in modo da minimizzare la funzione di errore

$$E = t_0 \log(y_0) + t_1 \log(y_1),$$

con $\mathbf{y} = (y_0, y_1) \in \mathbb{R}^2$ vettore di uscita corrispondente all'ingresso $\mathbf{I}$. Al passo $h \geq 0$ della procedura di ottimizzazione, la matrice dei pesi $\mathbf{W} = (w_{i,j})$ (comprendente anche i bias $b_i = w_{i,0}$) viene aggiornata secondo la regola

$$\mathbf{W}_{h+1} = \mathbf{W}_h - \ell \left(\frac{\partial E}{\partial \mathbf{W}_h} \right),$$

dove $\ell \in \mathbb{R}_+$ indica il *tasso di apprendimento* (*learning rate*). Il gradiente $\frac{\partial E}{\partial \mathbf{W}}$ punta verso la direzione di discesa più ripida nell'andamento della funzione di errore, cosicché aggiornare i pesi in misura proporzionale a $-\frac{\partial E}{\partial \mathbf{W}_h}$ permette di minimizzare l'errore.
Questo algoritmo è noto come **propagazione a ritroso con discesa del gradiente** (*backpropagation with gradient descent*), per una trattazione più approfondita del quale si rimanda a [RHW02] e [Hay94].
La rete neurale di PALMO impiega la strategia dell'*apprendimento a lotti* (*batch learning*). L'insieme di strutture proteiche componenti il training set è partizionato casualmente in $n_b = 100$ sottoinsiemi, detti *lotti*.[17] Il calcolo del gradiente di errore e l'aggiornamento dei

[17] In particolare, poiché il training set è composto da 500 strutture proteiche, ne vengono ricavati $n_b = 100$ lotti di esattamente $\frac{500}{n_b} = 5$ proteine ciascuno.

pesi viene effettuato dopo avere sottoposto alla rete neurale un intero lotto.
La matrice dei pesi è inizializzata con valori casuali, mentre, detto n_{tot} il numero totale di residui componenti le strutture proteiche del training set, il tasso di apprendimento ℓ è inizialmente posto a $10\frac{n_b}{n_{\text{tot}}}$. Un'*epoca* di apprendimento si conclude una volta che tutte le strutture proteiche del training set sono state sottoposte una volta alla rete neurale, ovvero quando il processo di calcolo del gradiente e di aggiornamento dei pesi è stato eseguito una volta per ciascun lotto. I lotti, a loro volta, vengono re-inizializzati in modo casuale all'inizio di ogni epoca. Se per 50 epoche consecutive non si riscontra una diminuzione dell'errore, il tasso di apprendimento ℓ viene dimezzato.

2.5 Algoritmo di programmazione dinamica

2.5.1 Programmazione dinamica: un'introduzione generale

Nel campo dell'algoritmica, l'espressione *programmazione dinamica* identifica una classe di metodi finalizzati alla soluzione di **problemi di ottimizzazione**.
Analogamente alla tecnica detta *divide et impera*, questo paradigma algoritmico sfrutta la possibilità di suddividere il problema di partenza in sottoproblemi più piccoli e giunge ad una soluzione ottima complessiva combinando soluzioni ottime parziali dei sottoproblemi (proprietà di *sottostruttura ottima*). La programmazione dinamica si rivela particolarmente efficiente quando sottoproblemi uguali si presentano più volte nel corso della risoluzione (*sottoproblemi sovrapponibili*): memorizza, infatti, ciascun risultato parziale in una matrice, in modo da poterlo immediatamente recuperare in caso di necessità, senza doverlo ricalcolare da capo.
Un algoritmo di programmazione dinamica si articola in quattro fasi:

1. caratterizzazione della struttura di una soluzione ottima;
2. definizione ricorsiva di una soluzione ottima;
3. calcolo di una soluzione ottima secondo un approccio *dal basso verso l'alto* (o *bottom-up*, cioè a partire dai sottoproblemi più piccoli e computazionalmente più elementari a quelli progressivamente più complessi);
4. costruzione di una soluzione ottima a partire dalle informazioni memorizzate nella matrice dei risultati parziali.

Questo paradigma consente, così, di risolvere in tempo polinomiale problemi che risulterebbero altrimenti esponenziali [LRSC01].

2.5.2 Applicazioni della programmazione dinamica alla biologia computazionale: allineamento di sequenze

Vari problemi nel campo della biologia computazionale si prestano ad essere risolti secondo tecniche di programmazione dinamica: l'**allineamento di sequenze** è forse l'esempio più classico.

Data la struttura primaria di due proteine, l'obiettivo è allinearle in modo tale da massimizzare la loro similarità, espressa sotto forma di un punteggio numerico. In genere, è consentito inserire degli spazi vuoti (*gap*) per tenere conto di eventuali inserzioni o rimozioni di residui dovute a mutazioni, al prezzo di una penalità sul punteggio. L'utilità di una stima quantitativa della "somiglianza" tra una coppia di proteine consiste nella possibilità di evidenziare relazioni di carattere funzionale, strutturale o filogenetico tra di esse.
Il punteggio di similarità di un allineamento può essere calcolato ricorsivamente in funzione degli allineamenti ottimi delle sottosequenze precedenti. La matrice di programmazione dinamica, a partire dalla prima riga e della prima colonna (associate ai sottoproblemi più elementari), viene progressivamente riempita con i punteggi ottimi di tutti i sottoallineamenti, sfruttando la definizione ricorsiva (Figura 2.10). Terminata la fase di *riem-*

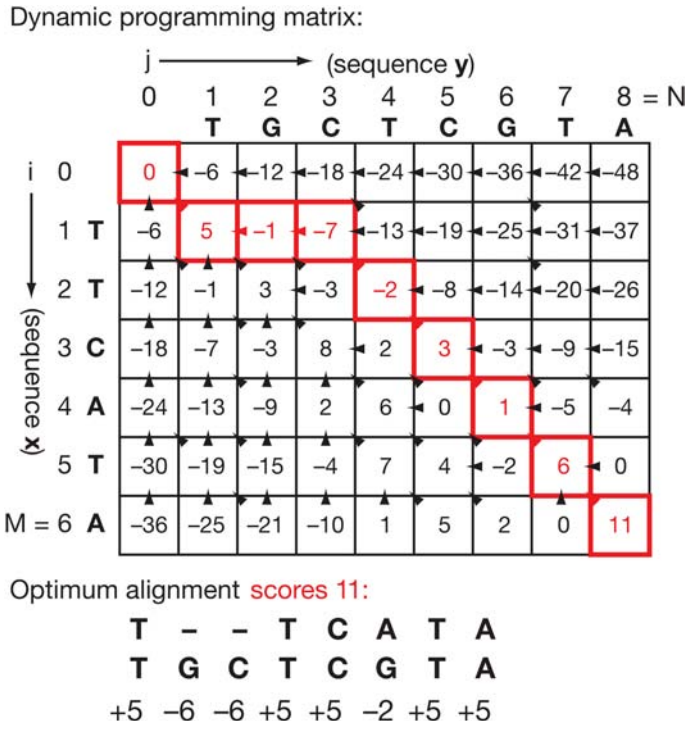


Figura 2.10: Esempio di matrice di programmazione dinamica per l'allineamento di due sequenze, $x =$ `TTCATA` e $y =$ `TGCTCGTA`. L'elemento $S(i, j)$ della matrice conserva il punteggio di similarità ottimo parziale per l'allineamento fra i prefissi $x_1 \dots x_i$ e $y_1 \dots y_j$. In rosso, il percorso seguito in fase di traceback nel costruire la soluzione ottima, a partire dal punteggio ottimo complessivo $S(M, N)$ a ritroso fino a $S(0, 0)$. [Edd04]

pimento (*filling-up*), la cella all'estremità in basso a destra contiene il punteggio ottimo dell'allineamento complessivo: si costruisce, quindi, la corrispondente soluzione ottima partendo da quest'ultima e compiendo una *risalita* (*traceback*) fino alla casella in alto a sinistra [Edd04].

L'algoritmo appena delineato, introdotto da Saul Needleman e Christian Wunsch nel 1970 [NW70], impone all'allineamento di includere le due sequenze proteiche nella loro interezza: in questo caso, si parla di **allineamento globale**. Una variante, presentata nel 1981 da Temple Smith e Michael Waterman [SW81] e detta **allineamento locale**, rimuove questo vincolo e consente, invece, di isolare sottosequenze di elevata similarità da sequenze proteiche complete. Tale tecnica è preferita quando le sequenze intere sono piuttosto divergenti, al punto da rendere poco significativo un tentativo di allinearle globalmente, e si vogliano mettere in evidenza relazioni di omologia fra regioni isolate (Tabella 2.2).

Globale	F T F T A L I L L A V A V F - - T A L - L L A - A V
Locale	F T F T A L I L L - A V A V - - F T A L - L L A A V - -

Tabella 2.2: Esempio di allineamento globale e locale fra le sequenze $x =$ `FTFTALILLAVAV` e $y =$ `FTALLLAAV`.

L'approccio adottato da Palmo trae ispirazione da quest'ultima tecnica, adattandola opportunamente all'obiettivo di individuare regioni peptidiche potenzialmente in grado di innescare la formazione di aggregati fibrillari.

2.5.3 L'algoritmo di PALMO

Come già accennato, il processo amiloidogenetico si concretizza nell'assemblaggio di una struttura cross-β coinvolgente un numero indeterminato di esemplari di una medesima catena polipeptidica. Pertanto, a differenza del consueto paradigma di allineamento locale, Palmo opera su una singola struttura proteica, nella quale la ricerca degli allineamenti si traduce nell'individuazione di coppie di regioni peptidiche dall'elevato potenziale aggregante.
D'altra parte, poiché i filamenti β all'origine della struttura amiloide corrispondono a segmenti di amminoacidi consecutivi, non è consentita l'introduzione di gap nella sequenza proteica.
Si noti che l'esistenza di due distinte orientazioni relative fra i segmenti peptidici coinvolti nella formazione di strutture β rende indispensabile affrontare separatamente la ricerca degli accoppiamenti paralleli e antiparalleli, rendendo necessarie strutture dati dedicate e variazioni, per quanto marginali, tra le procedure algoritmiche adottate nei due casi.

Definizioni preliminari

Sia $\mathcal{S}$ la sequenza primaria di una proteina, costituita da una successione di n residui amminoacidici,

$$\mathcal{S} := a_0 a_1 \dots a_{(n-2)} a_{(n-1)}.$$

Si definiscono due distinte matrici $n \times n$ di programmazione dinamica, P e A, a memorizzare le informazioni correlate agli accoppiamenti in orientazione, rispettivamente, parallela

e antiparallela.[18]
In particolare, il generico elemento $P_{i,\,j}$, con $i, j \in \{0,\, 1,\, \ldots,\, n-1\}$, tiene traccia della massima **propensione all'aggregazione** tra la seguente coppia di sottosequenze, entrambe di lunghezza $\ell = \min\{i,\, j\} + 1$:

$$\mathcal{S}\,[i-\ell+1,\, i] = a_{i-\ell+1}a_{i-\ell+2}\ldots a_{i-1}a_i$$
$$\mathcal{S}\,[j-\ell+1,\, j] = a_{j-\ell+1}a_{j-\ell+2}\ldots a_{j-1}a_j$$

Si ricorda, inoltre, che, data una qualsiasi coppia di residui a_h e a_k, la rete neurale è in grado di fornirne (tenuto conto del contesto di entrambi) la **probabilità di aggregazione** (**traslata e normalizzata**)[19] $\hat{p}_p\,(a_h,\, a_k)$.

Descrizione formale dell'algoritmo

Inizializzazione I casi più elementari consistono negli accoppiamenti fra singoli residui ($\ell = 1$). La matrice P, dunque, viene inizializzata a partire dalla prima riga e dalla prima colonna, ponendo

$$P_{0,\,j} := \max\{\hat{p}_p\,(a_0,\, a_j)\,,\ 0\} \quad \forall\, j \in \{0,\, \ldots,\, n-1\}$$
$$P_{i,\,0} := \max\{\hat{p}_p\,(a_i,\, a_0)\,,\ 0\} \quad \forall\, i \in \{0,\, \ldots,\, n-1\}$$

Gli accoppiamenti di lunghezza $\ell > 1$ sfruttano la semplice definizione ricorsiva della propensione all'aggregazione:

$$P_{i,\,j} := \max\{P_{i-1,\,j-1} + \hat{p}_p\,(a_i,\, a_j)\,,\ 0\} \quad \forall\, i, j \in \{1,\, \ldots,\, n-1\}$$

In questo modo, vengono progressivamente inizializzate le restanti righe e colonne. L'assenza di gap fa sì che il valore di ciascuna cella dipenda unicamente dall'elemento immediatamente precedente lungo la diagonale, $P_{i-1,\,j-1}$. Si noti, poi, come non sia ammesso l'inserimento di quantità negative: una propensione all'aggregazione nulla è indicata con uno zero.

Ricerca di regioni amiloidogeniche La ricerca dell'accoppiamento con la più accentuata propensione all'aggregazione inizia con l'individuazione dell'elemento massimo della matrice P. A partire da questo, ha inizio un percorso a ritroso (traceback) lungo la diagonale, che prosegue fintanto che si incontrano valori non nulli.
In altri termini, detto $P_{i_\omega,\,j_\omega}$ il valore massimo, la risalita tocca le celle

$$\underbrace{P_{i_\omega,\,j_\omega}}_{>0} \rightarrow \underbrace{P_{i_\omega-1,\,j_\omega-1}}_{>0} \rightarrow \underbrace{P_{i_\omega-2,\,j_\omega-2}}_{>0} \rightarrow \ldots \rightarrow \underbrace{P_{i_\alpha,\,j_\alpha}}_{>0},$$

[18] Per semplicità, si affronterà ora la sola orientazione parallela; per quella antiparallela, si veda la Sezione 2.5.3, pag. 33. I pedici p e a identificheranno le quantità riferite, rispettivamente, al primo e al secondo caso.

[19] Si veda la Sezione 2.6, pag. 37.

arrestandosi in corrispondenza dell'elemento $P_{i_\alpha, j_\alpha} > 0$ tale che $P_{i_\alpha - 1, j_\alpha - 1} = 0$ oppure si è raggiunto il bordo della matrice (cioè $i_\alpha = 0 \ \vee \ j_\alpha = 0$).
L'accoppiamento di massima propensione all'aggregazione risulta

$$\mathcal{S}\left([i_\alpha,\, i_\omega]\,,\ [j_\alpha,\, j_\omega]\right) = \begin{array}{ccccc} a_{i_\alpha} & a_{i_\alpha+1} & \dots & a_{i_\omega-1} & a_{i_\omega} \\ \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ a_{j_\alpha} & a_{j_\alpha+1} & \dots & a_{j_\omega-1} & a_{j_\omega} \end{array}$$

Considerazioni di carattere biochimico impongono un vincolo sulla lunghezza dei segmenti peptidici: poiché, in particolare, 4 è comunemente assunto come il numero minimo di residui necessari a formare un filamento β, solo gli accoppiamenti di lunghezza non inferiore a tale valore (tali che, cioè, $\ell = i_\omega - i_\alpha + 1 = |j_\omega - j_\alpha| + 1 \geq 4$) vengono presi in considerazione.
La coppia di segmenti con la seconda migliore propensione all'aggregazione può essere individuata con un'analoga procedura di traceback, a partire dal massimo tra gli elementi di P non considerati nella prima iterazione.
È possibile iterare l'algoritmo di ricerca un numero arbitrario di volte, fintanto che in P esistono celle non nulle ancora intatte, ad isolare accoppiamenti fra sottosequenze di propensione via via decrescente.

Un esempio passo-passo

Un esempio reale può aiutare a far apparire più chiara la procedura di ricerca appena formalizzata.
Data la sequenza peptidica $\mathcal{S} = \texttt{KKLVFFAED}$ di lunghezza $n = 9$, si supponga che la fase di inizializzazione abbia condotto alla matrice P degli allineamenti in orientazione parallela mostrata in Figura 2.11, pag. 33.
Il primo passo consiste nella ricerca dell'elemento massimo, che è individuato in $P_{6,6} = 153.1$. A partire da questo, si sale a ritroso in diagonale, lungo un cammino che si conclude al raggiungimento del primo valore nullo (Figura 2.12, pag. 34).
Associato al cammino di traceback è l'accoppiamento di massima propensione all'aggregazione:

$$\mathcal{S}\left([2,\, 6]\,,\ [2,\, 6]\right) = \begin{array}{ccccc} \texttt{L} & \texttt{V} & \texttt{F} & \texttt{F} & \texttt{A} \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ \texttt{L} & \texttt{V} & \texttt{F} & \texttt{F} & \texttt{A} \end{array}$$

Questo risulta ammissibile, in quanto di lunghezza $\ell = 5 \geq 4$.
L'algoritmo può ora essere ripetuto, avendo cura di escludere le celle di P corrispondenti alla coppia appena individuata. Così, la seconda iterazione si concretizzerà nel percorso tracciato in Figura 2.13, pag. 35. L'accoppiamento con la seconda migliore propensione all'aggregazione, anch'esso ammissibile, risulterà, pertanto,

$$\mathcal{S}\left([2,\, 6]\,,\ [1,\, 5]\right) = \begin{array}{ccccc} \texttt{L} & \texttt{V} & \texttt{F} & \texttt{F} & \texttt{A} \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ \texttt{K} & \texttt{L} & \texttt{V} & \texttt{F} & \texttt{F} \end{array}$$

E così via, finché tutti gli elementi diversi da zero non saranno stati presi in considerazione ed eliminati.

	0	1	2	3	4	5	6	7	8	
0	0.0	0.0	1.5	3.8	1.4	1.4	0.0	0.0	0.0	K
1	0.0	0.0	3.9	11.7	15.1	3.8	0.0	0.0	0.0	K
2	1.5	3.9	27.2	46.8	48.2	33.6	13.0	1.8	1.5	L
3	3.8	11.7	46.8	83.9	99.6	80.6	54.6	21.3	5.8	V
4	1.4	15.1	48.2	99.6	135.7	132.6	100.3	64.5	22.7	F
5	1.4	3.8	33.6	80.6	132.6	152.0	139.0	101.5	66.0	F
6	0.0	0.0	13.0	54.6	100.3	139.0	153.1	89.9	96.0	A
7	0.0	0.0	1.8	21.3	64.5	101.5	89.9	85.7	56.6	E
8	0.0	0.0	1.5	5.8	22.7	66.0	96.0	56.6	57.1	D
	K	K	L	V	F	F	A	E	D	

Figura 2.11: Esempio di matrice P di programmazione dinamica nel caso parallelo.

Adattamento al caso antiparallelo

Quanto descritto sopra può essere adattato alla ricerca di coppie amiloidogeniche in orientazione antiparallela con alcune immediate variazioni.
L'inizializzazione della matrice A inizia dalla prima riga e dall'ultima colonna, per poi estendersi alle restanti, in accordo con una definizione ricorsiva "speculare" rispetto alla precedente:

$$\begin{aligned}
A_{0,\,j} &:= \max\left\{\hat{p}_a\left(a_0,\, a_j\right),\ 0\right\} && \forall\ j \in \{0,\, \ldots,\, n-1\} \\
A_{i,\,n-1} &:= \max\left\{\hat{p}_a\left(a_i,\, a_{n-1}\right),\ 0\right\} && \forall\ i \in \{0,\, \ldots,\, n-1\} \\
A_{i,\,j} &:= \max\left\{A_{i-1,\,j+1} + \hat{p}_a\left(a_i,\, a_j\right),\ 0\right\} && \forall\ i \in \{1,\, \ldots,\, n-1\},\quad \forall\ j \in \{0,\, \ldots,\, n-2\}
\end{aligned}$$

Analogamente, il percorso di risalita a partire dall'elemento massimo $A_{i_\omega,\,j_\omega}$ segue la direzione antidiagonale, dalla sinistra in basso alla destra in alto,

$$\underbrace{A_{i_\omega,\,j_\omega}}_{>0} \rightarrow \underbrace{A_{i_\omega-1,\,j_\omega+1}}_{>0} \rightarrow \underbrace{A_{i_\omega-2,\,j_\omega+2}}_{>0} \rightarrow \ \ldots\ \rightarrow \underbrace{A_{i_\alpha,\,j_\alpha}}_{>0},$$

	0	1	2	3	4	5	6	7	8	
0	0.0	0.0	1.5	3.8	1.4	1.4	0.0	0.0	0.0	K
1	0.0	0.0	3.9	11.7	15.1	3.8	0.0	0.0	0.0	K
2	1.5	3.9	27.2	46.8	48.2	33.6	13.0	1.8	1.5	L
3	3.8	11.7	46.8	83.9	99.6	80.6	54.6	21.3	5.8	V
4	1.4	15.1	48.2	99.6	135.7	132.6	100.3	64.5	22.7	F
5	1.4	3.8	33.6	80.6	132.6	152.0	139.0	101.5	66.0	F
6	0.0	0.0	13.0	54.6	100.3	139.0	**153.1**	89.9	96.0	A
7	0.0	0.0	1.8	21.3	64.5	101.5	89.9	85.7	56.6	E
8	0.0	0.0	1.5	5.8	22.7	66.0	96.0	56.6	57.1	D
	K	K	L	V	F	F	A	E	D	

Figura 2.12: Prima iterazione della procedura di traceback su P.

dove $A_{i_\alpha - 1,\, j_\alpha + 1} = 0$ oppure $i_\alpha = 0 \ \vee \ j_\alpha = n - 1$.
L'accoppiamento che ne risulta è

$$\mathcal{S}\left([i_\alpha,\, i_\omega]\,,\ [j_\alpha,\, j_\omega]\right) = \begin{array}{ccccc} a_{i_\alpha} & a_{i_\alpha+1} & \dots & a_{i_\omega-1} & a_{i_\omega} \\ \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ a_{j_\alpha} & a_{j_\alpha-1} & \dots & a_{j_\omega+1} & a_{j_\omega} \end{array}$$

Calcolo del punteggio di un accoppiamento

Come si è visto, ciascuna coppia di sottosequenze peptidiche è contraddistinta da un valore di propensione all'aggregazione, cioè da quel $P_{i_\omega,\, j_\omega}$ che fa da punto di inizio del percorso di traceback corrispondente all'accoppiamento in questione.[20] In prima battuta, tale

[20] Come di consueto, si fa riferimento al solo caso parallelo. In quello antiparallelo, ovviamente, la propensione all'aggregazione sarà $A_{i_\omega,\, j_\omega}$.

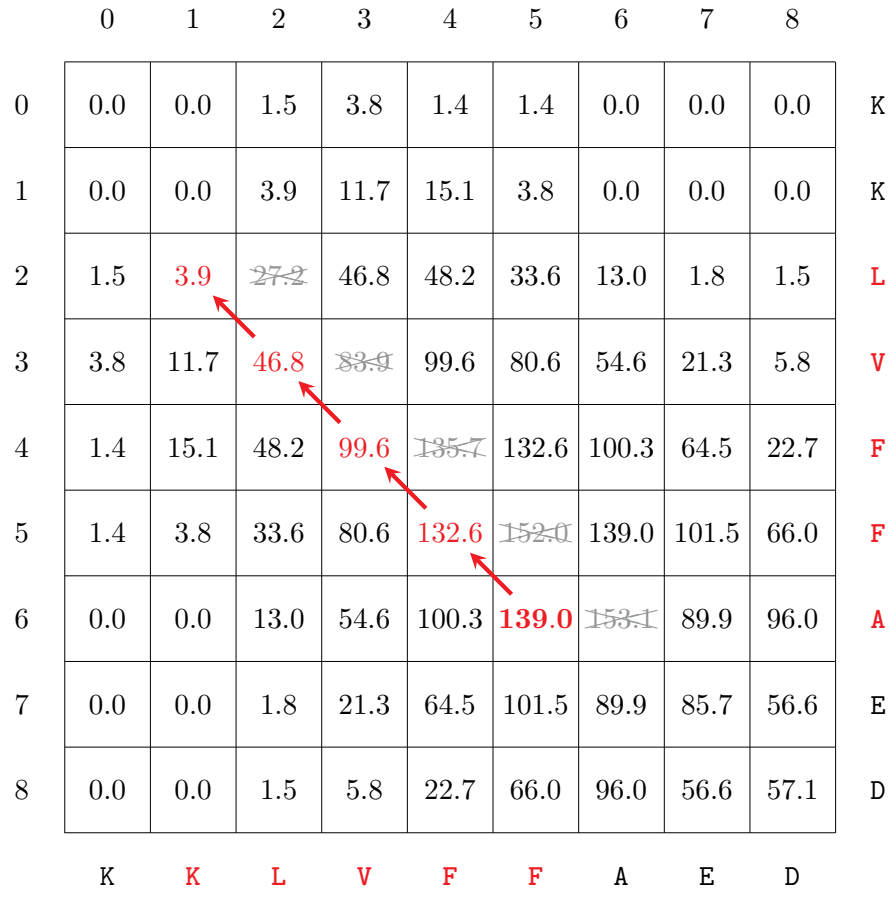


Figura 2.13: Seconda iterazione della procedura di traceback su P.

valore può essere ritenuto una buona stima quantitativa dell'effettiva tendenza a formare strutture fibrillari e, di conseguenza, un valido candidato nella scelta del punteggio da assegnare alla coppia.
Tuttavia, si è notato come la natura additiva della definizione di propensione tenda, in linea di massima, a premiare oltremodo le coppie di lunghezza maggiore, assegnando a queste ultime valori più alti.
Ciò contrasta con i dati sperimentali: infatti, l'analisi delle strutture proteiche del Top500 Database ha rivelato come i filamenti β siano solitamente piuttosto brevi. In dettaglio, la funzione di distribuzione delle lunghezze dei filamenti β (Figura 2.14) mostra come quasi il 90% sia entro i 6 residui di lunghezza.
In base a questa osservazione, si è ritenuto opportuno applicare alla propensione P_{i_ω, j_ω} delle coppie eccedenti tale soglia una penalità proporzionale alla lunghezza. In altri termini, detto $s\left(\mathcal{S}\left([i_\alpha, i_\omega], [j_\alpha, j_\omega]\right)\right)$ il **punteggio** (*score*) **di aggregazione** della coppia

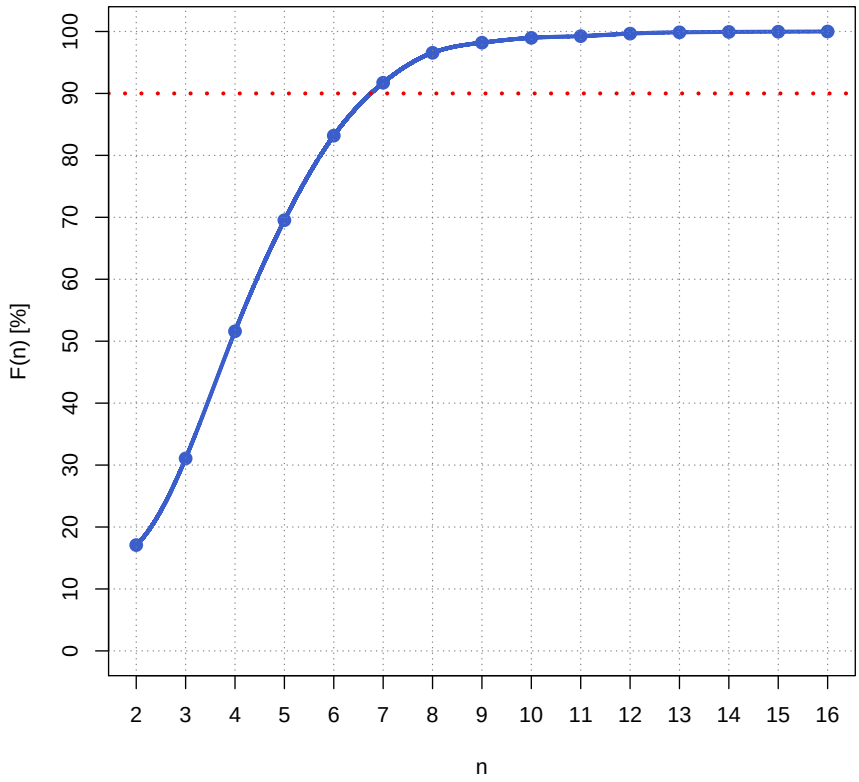


Figura 2.14: Il grafico mostra l'andamento della funzione $F(n)$ di distribuzione cumulativa delle lunghezze n dei filamenti β ricavati dalle strutture proteiche del Top500 Database, $n \in \{2, \ldots, 16\}$. $F(n)$ rappresenta la percentuale di filamenti β di lunghezza minore o uguale a n.

$\mathcal{S}([i_\alpha, i_\omega], [j_\alpha, j_\omega])$, di lunghezza $\ell = i_\omega - i_\alpha + 1 = |j_\omega - j_\alpha| + 1$, si è posto

$$s(\mathcal{S}([i_\alpha, i_\omega], [j_\alpha, j_\omega])) := \begin{cases} P_{i_\omega, j_\omega}, & \ell \in \{4, 5, 6\} \\ P_{i_\omega, j_\omega} - \ell K, & \ell \geq 7 \end{cases}$$

Se $s(\cdot) \leq 0$, l'accoppiamento viene scartato.[21]
L'ammontare unitario $K \in \mathbb{R}_+$ della penalità è stato determinato attraverso una procedura di ottimizzazione sulle strutture proteiche del Top500 Database. Essendo nota a priori la struttura secondaria di queste, si è impiegato PALMO come predittore di filamenti β, al fine di selezionare il valore di K che garantisce il massimo accordo con le informazioni

[21] Si ricorda, inoltre, che gli accoppiamenti di lunghezza $\ell < 4$ sono stati scartati in precedenza.

strutturali sul data set.
Più in dettaglio, si è assunto di poter approssimare i filamenti β con i segmenti peptidici componenti gli accoppiamenti $\mathcal{S}([i_\alpha, i_\omega], [j_\alpha, j_\omega])$ di punteggio $s(\cdot) > 0$ individuati da PALMO. Ovviamente, ciascuna distinta scelta di K genera un diverso insieme di filamenti β: infatti, la penalità $-\ell K$, combinata al vincolo di positività sul punteggio, fa sì che, all'aumentare di K, si riduca il numero di segmenti di lunghezza $\ell \geq 7$ predetti. L'accuratezza della predizione rispetto ai filamenti β noti sperimentalmente è stata quantificata in termini di *punteggio di sovrapposizione fra segmenti* (*segment overlap score* o SOV, [ZVFR99]). Una procedura di ottimizzazione di tipo *griglia di ricerca* (*grid search*) ha consentito di individuare il valore di K che produce il punteggio SOV massimo.[22]

2.6 Interfaccia fra rete neurale e algoritmo di programmazione dinamica

In fase di inizializzazione delle matrici P e A, l'algoritmo di programmazione dinamica invoca la rete neurale per ottenerne un valore numerico direttamente proporzionale alla probabilità di aggregazione di ciascuna coppia di residui in esame.
Il risultato dell'elaborazione della rete neurale consiste in una probabilità propriamente detta, ovvero in un numero reale $p \in [0, 1]$.[23] D'altra parte, il principio alla base dell'algoritmo di programmazione dinamica impone che le quantità numeriche in ingresso alla procedura di inizializzazione possano assumere valori negativi, ovvero $\hat{p} \in [-k, k]$, $k \in \mathbb{R}_+$. I valori negativi sono intesi a segnalare una coppia di residui particolarmente sfavorevole al consolidamento di strutture fibrillari, cosicché risulti "penalizzato" nel suo complesso l'accoppiamento fra segmenti cui questa coppia appartiene, grazie al termine additivo presente nella definizione ricorsiva della propensione all'aggregazione.
Pertanto, si è reso necessario introdurre una funzione

$$\begin{array}{rccc} f: & [0, 1] & \longrightarrow & [-k, k] \\ & p & \longmapsto & \hat{p} \end{array}$$

caratterizzata da

- uno zero (ovvero, un valore $p \in [0, 1]$ tale che $\hat{p} = f(p) = 0$) che rappresentasse un opportuno discrimine tra le coppie di residui propense all'aggregazione e quelle sfavorevoli;
- la capacità di distribuire i valori con omogeneità nell'intero codominio $[-k, k]$.

A tale proposito, sono state generate casualmente 25 milioni di coppie distinte di quintuple amminoacidiche e si è individuata la mediana p_{med} delle rispettive probabilità di

[22] Il valore ottimo è risultato $K_{\text{opt}} := 11.40$.

[23] Qui e nel seguito non si farà distinzione tra il caso parallelo e quello antiparallelo, che sono del tutto speculari.

aggregazione. Detti, poi, p_m e p_M i valori minimo e massimo assoluti di probabilità di aggregazione,[24] si è costruita la seguente funzione:

$$f(p) := \begin{cases} 100\,(p - p_\mathrm{med})\,, & p \geq p_\mathrm{med} \\ -\frac{p_\mathrm{M} - p_\mathrm{med}}{p_\mathrm{m} - p_\mathrm{med}} 100\,(p - p_\mathrm{med})\,, & p < p_\mathrm{med} \end{cases}$$

In altri termini, la funzione consiste in una traslazione verso sinistra pari a p_med e in una normalizzazione per una costante, diversa a seconda della posizione rispetto alla mediana (Figura 2.15).
Zero di tale funzione è proprio la mediana p_med. Inoltre, $\hat{p}_\mathrm{m} := f\,(p_\mathrm{m}) = -100\,(p_\mathrm{M} - p_\mathrm{med})$ e $\hat{p}_\mathrm{M} := f\,(p_\mathrm{M}) = 100\,(p_\mathrm{M} - p_\mathrm{med})$, ovvero le probabilità di aggregazione minima e massima vengono mappate in valori opposti, $\hat{p}_\mathrm{m} = -\hat{p}_\mathrm{M}$. Il fattore moltiplicativo 100 è una costante scelta arbitrariamente.

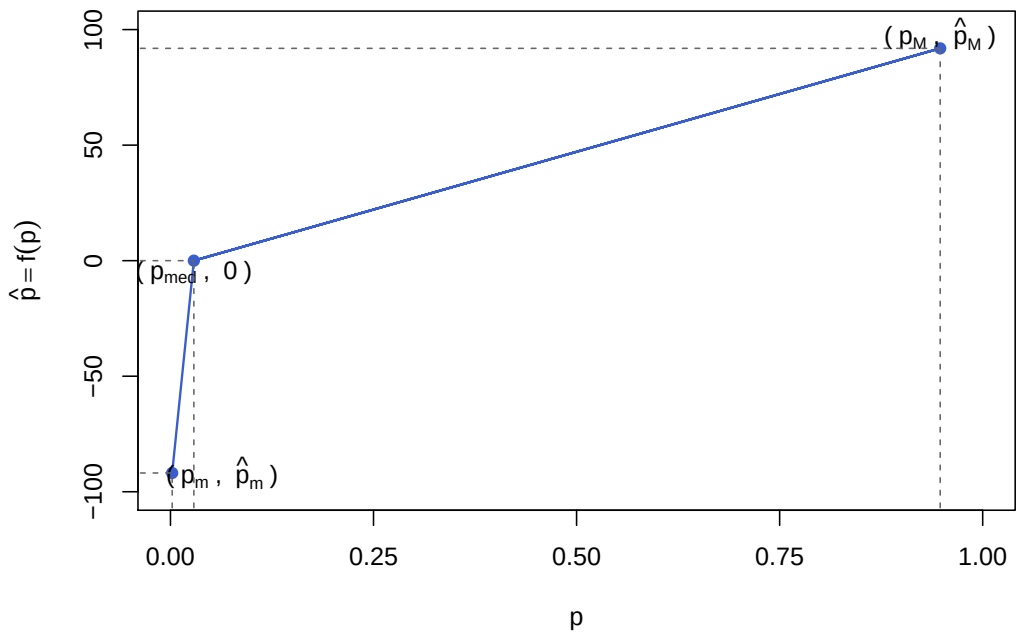


Figura 2.15: Grafico della funzione f di traslazione e normalizzazione della probabilità di aggregazione p utilizzata in Palmo.

[24] In dettaglio, p_m e p_M sono i valori restituiti dalla rete neurale in corrispondenza delle coppie di quintuple ([PPPPP], [PPPPP]) e ([VVVVV], [VVVVV]), rispettivamente. Ciò è in accordo con i dati sperimentali, che, da un lato, attribuiscono alla prolina (P) la capacità di rompere le strutture secondarie come i foglietti β, dall'altro riconoscono la valina (V) come uno degli amminoacidi più frequentemente coinvolti nella stabilizzazione di aggregati.

2.7 Metodi impiegati nella verifica dei risultati

Si illustrano di seguito le tecniche, sia algoritmiche sia statistiche, e le misure impiegate nel quantificare l'accuratezza di Palmo e nel compararlo ad analoghi metodi preesistenti. Viene trattati separatamente ciascuno dei due distinti problemi di interesse nello studio dell'amiloidogenesi: la **classificazione di frammenti peptidici** in aggreganti e non e il **rilevamento di regioni amiloidogeniche** in sequenze proteiche complete.

2.7.1 Classificazione di peptidi

Nel caso in esame, *classificare* un insieme di peptidi consiste nell'assegnarne ciascun elemento ad una e una sola tra le due classi *amiloidogenico* (o *positivo*) e *non amiloidogenico* (o *negativo*).
In concreto, Palmo (così come gli analoghi metodi scelti per la comparazione) può essere impiegato come classificatore binario, associando ogni frammento proteico ad un punteggio complessivo di aggregazione, coincidente con il massimo punteggio degli accoppiamenti fra sottosequenze individuati dall'algoritmo. La scelta di un valore soglia (*threshold*) consente di partizionare il test set, marcando come amiloidogenici tutti e soli i peptidi di punteggio non inferiore a tale soglia.
Se i frammenti proteici sono preclassificati sulla base di dati sperimentali (è questo il caso del data set *Tango*), è possibile valutare per confronto la qualità della predizione operata dal metodo in esame. In particolare, detti rispettivamente tp, fp, tn, fn il numero di *veri positivi*, *falsi positivi*, *veri negativi* e *falsi negativi*, alcune tra le misure di accuratezza più utilizzate sono

- **precisione** (*precision*):
$$pr := \frac{tp}{tp + fp} \in [0,\, 1]$$
- **sensibilità** (*sensitivity*) o **tasso di veri positivi** (*true positive rate*):
$$se = tpr := \frac{tp}{tp + fn} \in [0,\, 1]$$
- **specificità** (*specificity*):
$$sp := \frac{tn}{fp + tn} \in [0,\, 1]$$
- **tasso di falsi positivi** (*false positive rate*):
$$fpr := 1 - sp = \frac{fp}{fp + tn} \in [0,\, 1]$$
- **accuratezza bilanciata** (*balanced accuracy*):
$$ba := \frac{se + sp}{2} \in [0,\, 1]$$

- **coefficiente di correlazione di Matthews** (*Matthews correlation coefficient*):

$$mcc := \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)\,(tp + fn)\,(tn + fp)\,(tn + fn)}} \in [-1,\, 1]$$

In genere, sono desiderabili valori elevati di precisione, sensibilità, specificità, accuratezza bilanciata e coefficiente di correlazione di Matthews; al contrario, è preferibile un basso tasso di falsi positivi.

Curve ROC e AUC

Una rappresentazione grafica utile a visualizzare le prestazioni di un classificatore binario è la **curva ROC** (*receiver operating characteristic*) [Lus71]. Ad ogni valore di soglia t distinto corrisponde una diversa partizione dell'insieme di peptidi, caratterizzata da un tasso di veri positivi $tpr\,(t)$ e di falsi negativi $fpr\,(t)$. In un piano cartesiano avente il tasso di falsi positivi sulle ascisse e il tasso di veri positivi sulle ordinate, la curva ROC è data dai punti di coordinate $(fpr\,(t)\,,\, tpr\,(t))$ al variare del punteggio soglia t.
Una curva ROC coincidente con la bisettrice segnala un classificatore casuale, mentre un predittore perfetto genera una curva ROC che collega l'origine con il punto $(0,\, 1)$ e quest'ultimo con $(1,\, 1)$ (Figura 2.16). La capacità discriminante di un classificatore,

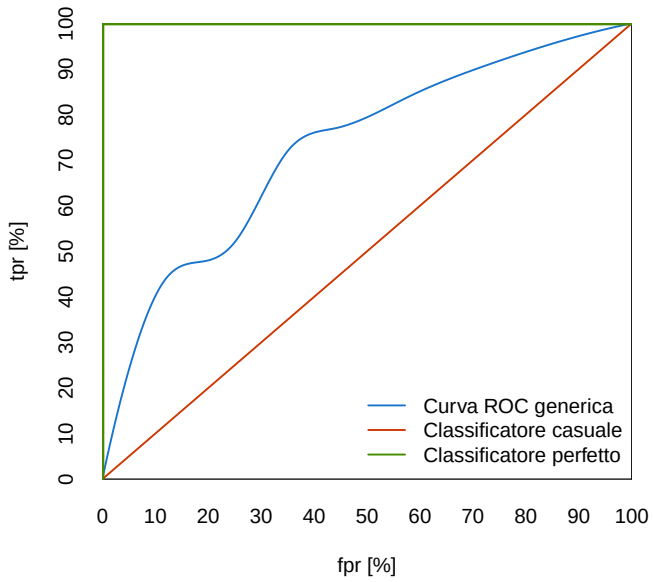


Figura 2.16: Esempio di curve ROC. Evidenziata in azzurro la AUC relativa alla curva ROC generica.

ovvero la sua attitudine a fornire un corretto partizionamento di una popolazione di peptidi

nelle due classi, è proporzionale al valore dell'**area sottesa dalla curva ROC** o **AUC** (*area under the curve*), coincidente con la probabilità che ad un elemento estratto a caso dall'insieme di peptidi riconosciuti sperimentalmente come amiloidogenici venga attribuito un punteggio di aggregazione maggiore rispetto ad un frammento scelto casualmente tra quelli non amiloidogenici.
In particolare, può essere utile focalizzare l'attenzione sulle situazioni di alta specificità, ovvero in corrispondenza ai soli valori soglia t che producono una predizione caratterizzata da un tasso di falsi positivi entro un valore $fpr_{\max}$ fissato: a questo scopo, è sufficiente considerare l'area sottesa alla curva ROC nell'intervallo di ascisse $[0,\ fpr_{\max}]$ (Figura 2.17).
Nella valutazione comparativa di PALMO rispetto agli analoghi predittori, si è ritenuto interessante confrontare le estensioni delle AUC parziali per tassi di falsi positivi entro il 5%, 10% e 20%.

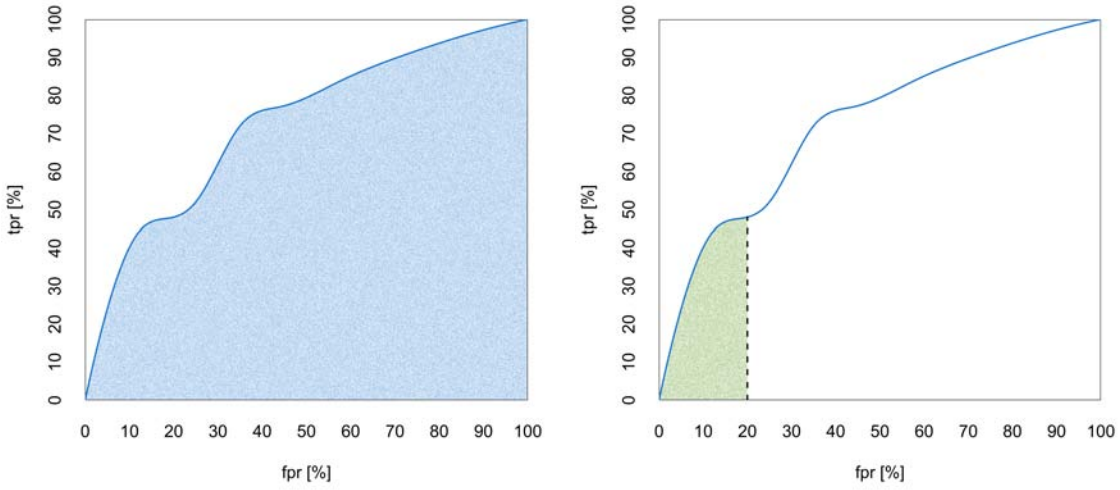


Figura 2.17: A sinistra, in azzurro, l'area sottesa alla curva ROC di Figura 2.16; a destra, in verde, la AUC parziale per $fpr \leq 20\%$, ovvero specificità $sp \geq 80\%$.

Jackknife e test t di Student

La ridotta dimensione del test set Tango ha reso indispensabile una verifica della significatività statistica dei valori di AUC, effettuata tramite l'applicazione combinata del metodo *jackknife* e del *test di Student*.
Si definisce **jackknife** una tecnica di ricampionamento che permette, quando le dimensioni del campione non siano adeguate a garantire una inferenza affidabile, di costruire uno stimatore di una grandezza statistica: nel caso in esame, l'area sottesa alla curva per un certo classificatore e per un tasso di falsi positivi entro un $fpr_{\max}$ prefissato. La procedura jackknife si è articolata in $n = 10^7$ iterazioni: in ciascuna, si è effettuato un

ricampionamento del test set di partenza, estraendone casualmente (senza ripetizioni) il 70% circa dei peptidi, e, sulla base di questo, si è calcolata la AUC relativa al predittore di interesse. Si sono, così, ottenute n stime dell'area sottesa alla curva.
A partire dalla media $\bar{x}$ delle n stime, un **test t di Student bilaterale a un campione** (*two-sided one-sample Student's t test*) ha consentito di ricavare un intervallo di confidenza $100\,(1-\alpha)\,\%$ per la media *vera* μ delle AUC: ovvero, un *range* $[\bar{x}-E,\ \bar{x}+E]$ tale che, nel $100\,(1-\alpha)\,\%$ dei casi, un intervallo costruito con la medesima tecnica di stima contenga μ [Ros09].[25]
Per verificare, invece, se i valori di AUC ottenuti per Palmo siano in media significativamente maggiori rispetto a ciascuno dei metodi concorrenti, si è fatto ricorso ad un **test t di Student unilaterale a campioni appaiati** (*one-sided paired Student's t test*).[26]

2.7.2 Individuazione di regioni amiloidogeniche in sequenze proteiche complete

Data la struttura primaria delle proteine descritte nella sezione 2.3, ci si è posti l'obiettivo di quantificare l'accuratezza di Palmo nell'isolare quelle regioni che, in base a studi sperimentali, sono riconosciute come particolarmente soggette alla formazione di strutture fibrillari.
La predizione si è basata sull'insieme degli accoppiamenti fra sottosequenze restituiti da Palmo, marcando come amiloidogenici tutti i residui costituenti le n coppie di punteggio più alto *non ridondanti*,[27] con n compreso tra 1 e 10.
Per quanto riguarda gli analoghi predittori con cui Palmo è stato confrontato, si è applicato, ove possibile, il medesimo criterio appena descritto.

---

[25] Si assume che i valori di AUC seguano una distribuzione normale $\mathcal{N}\left(\mu,\,\sigma^2\right)$ di media μ e varianza σ^2, entrambe ignote. Supponendo di non avere a disposizione altro che un campione di dimensione n, da cui si siano ricavate una stima $\bar{x}$ della media e s^2 della varianza, e di avere fissato un livello di confidenza $100\,(1-\alpha)\,\%$, con $\alpha \in [0,\,1]$, il *test t di Student a un campione* restituisce un intervallo $[\bar{x}-E,\ \bar{x}+E] = \left[\bar{x}-t_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}},\ \bar{x}+t_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}\right]$, dove $t_{\frac{\alpha}{2}}$ indica il $100\frac{\alpha}{2}$ percentile della *distribuzione t di Student con* $(n-1)$ *gradi di libertà*, data da $T := \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$. Nel caso specifico, si è posto $\alpha := 0.01$, pari ad un livello di confidenza del 99%.

[26] Supponendo di avere calcolato, per ciascun ricampionamento i del test set di partenza, la coppia $(x_{i,\,1},\ x_{i,\,2})$ di AUC relative, rispettivamente, a Palmo e a un secondo classificatore, siano $\bar{x}_d$ e s_d^2 la media e la varianza delle differenze $(x_{i,\,1}-x_{i,\,2})$. Se $p := P\left(T \leq t_{\frac{\alpha}{2}}\right) < \alpha$, dove $T := \frac{\bar{X}_d}{\sqrt{\frac{s_d^2}{n}}}$, è possibile concludere con confidenza $100\,(1-\alpha)\,\%$ che la media delle aree sottese alle curve ROC di Palmo è maggiore rispetto ai predittori concorrenti.

[27] Un accoppiamento è considerato *ridondante* se i residui che lo compongono risultano già marcati come amiloidogenici, perché facenti parte di una coppia di punteggio non inferiore considerata in precedenza. In altri termini, si considerano solo le prime n coppie che determinano un effettivo ampliamento della predizione, in ordine di punteggio non crescente.

Punteggio di sovrapposizione fra segmenti SOV

L'accuratezza di predizione è stata quantificata in termini di **punteggio di sovrapposizione fra segmenti** o **SOV** (*segment overlap score*) [ZVFR99]. Il punto di forza di questa misura, inizialmente concepita per valutare predizioni di struttura secondaria, risiede nell'assumere come unità strutturale il segmento anziché il singolo residuo. Tale assunzione consente di catturare le caratteristiche strutturali più importanti (i segmenti nel loro complesso) e di ridurre l'impatto di dettagli secondari (come piccole variazioni della lunghezza e nella posizione dei segmenti) rispetto alle stime basate sul singolo residuo (quali, ad esempio, sensibilità, specificità, ecc.), risultando, in definitiva, ben più significativo di queste ultime nel caso in esame.
Ovviamente, il diverso ambito di applicazione ha reso necessari alcuni marginali adeguamenti della definizione originaria di punteggio SOV, illustrati di seguito.
Data una sequenza proteica π, tanto la struttura derivata dagli studi sperimentali (*struttura osservata*) quanto il risultato della predizione (*struttura predetta*) risultano segmentati in regioni amiloidogeniche (stato `A`) e non amiloidogeniche (stato `N`).

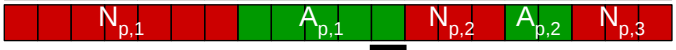
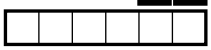
Struttura osservata	
Struttura predetta	
minov $(A_{o,2},\ A_{p,1})$	
maxov $(A_{o,2},\ A_{p,1})$	
minov $(A_{o,2},\ A_{p,2})$	
maxov $(A_{o,2},\ A_{p,2})$	

Tabella 2.3: Illustrazione del calcolo di un punteggio parziale $\mathrm{SOV}_\pi(\mathtt{A})$. I segmenti verdi rappresentano le regioni amiloidogeniche, quelli rossi le non amiloidogeniche. Le barre nere e bianche corrispondono, rispettivamente, a minov e maxov per i segmenti amiloidogenici sovrapposti.

Detto s_1 un segmento osservato e s_2 uno predetto, sia $S_\pi(i)$ l'insieme di tutte le coppie di regioni (anche solo parzialmente) sovrapposte $(s_1,\ s_2)$, entrambe in stato $i \in \{\mathtt{A}, \mathtt{N}\}$, e $S'_\pi(i)$ l'insieme di tutti i segmenti s_1 che non ammettono alcuna sovrapposizione con regioni nello stesso stato i, ovvero

$$S_\pi(i) := \{(s_1,\ s_2) : s_1 \cap s_2 \neq \varnothing,\ s_1 \text{ e } s_2 \text{ in stato } i\}$$
$$S'_\pi(i) := \{s_1 : \forall\ s_2,\ s_1 \cap s_2 = \varnothing,\ s_1 \text{ e } s_2 \text{ in stato } i\}$$

Per ciascuno stato i, il punteggio SOV parziale è dato da

$$\mathrm{SOV}_\pi(i) := 100 \frac{1}{N_\pi(i)} \sum_{(s_1,\, s_2) \,\in\, S_\pi(i)} \left[\frac{\mathrm{minov}(s_1,\ s_2) + \delta(s_1,\ s_2)}{\mathrm{maxov}(s_1,\ s_2)} \cdot \ell(s_1) \right]$$

dove (Tabella 2.3)

- $\ell(s_1)$ è il numero di residui di s_1;

- $N_\pi(i) := \sum_{(s_1,\, s_2) \,\in\, S_\pi(i)} \ell\,(s_1) + \sum_{s_1 \,\in\, S'_\pi(i)} \ell\,(s_1)$ è un fattore di normalizzazione;
- $\mathrm{minov}\,(s_1,\, s_2)$ è la lunghezza dell'effettiva sovrapposizione fra s_1 e s_2 (cioè, della sequenza i cui residui sono in stato i in *entrambe* le sequenze s_1 e s_2);
- $\mathrm{maxov}\,(s_1,\, s_2)$ è la lunghezza della sequenza i cui residui sono in stato i in *almeno una* fra s_1 e s_2;
- $\delta\,(s_1,\, s_2) := \min\left\{\left[\mathrm{maxov}\,(s_1,\, s_2) - \mathrm{minov}\,(s_1,\, s_2)\right],\ \mathrm{minov}\,(s_1,\, s_2)\,,\ \left\lfloor \frac{\ell(s_1)}{2} \right\rfloor,\ \left\lfloor \frac{\ell(s_2)}{2} \right\rfloor\right\}$.

Il punteggio SOV finale per la proteina π è definito come

$$\mathrm{SOV}_\pi := 100 \frac{1}{N_\pi} \underbrace{\sum_{i \,\in\, \{\mathtt{A},\, \mathtt{N}\}} \ \sum_{(s_1,\, s_2) \,\in\, S_\pi(i)} \left[\frac{\mathrm{minov}\,(s_1,\, s_2) + \delta\,(s_1,\, s_2)}{\mathrm{maxov}\,(s_1,\, s_2)} \cdot \ell\,(s_1) \right]}_{(\star)}$$

con $N_\pi := \sum_{i \,\in\, \{\mathtt{A},\, \mathtt{N}\}} N_\pi\,(i)$.
Una stima complessiva per l'intero test set T è stata ricavata sommando i termini $(\star)$ relativi a ciascuna proteina e normalizzando per il fattore totale $N_{\mathrm{tot}} := \sum_{\pi \,\in\, T} N_\pi$.

$$\begin{aligned} \overline{\mathrm{SOV}} :&= 100 \frac{1}{N_{\mathrm{tot}}} \sum_{\pi \,\in\, T} (\star) \\ &= 100 \frac{1}{N_{\mathrm{tot}}} \sum_{\pi \,\in\, T} \ \sum_{i \,\in\, \{\mathtt{A},\, \mathtt{N}\}} \ \sum_{(s_1,\, s_2) \,\in\, S_\pi(i)} \left[\frac{\mathrm{minov}\,(s_1,\, s_2) + \delta\,(s_1,\, s_2)}{\mathrm{maxov}\,(s_1,\, s_2)} \cdot \ell\,(s_1) \right] \end{aligned}$$

Capitolo 3

Risultati e discussione

3.1 PALMO come evoluzione di PASTA

Lo sviluppo di Palmo si è posto come traguardo primario la realizzazione di un metodo che, pur costruito sulle stesse assunzioni biochimiche fondamentali alla base di Pasta, fosse in grado, grazie all'applicazione di tecniche algoritmiche più avanzate, di superare le già buone prestazioni di quest'ultimo in termini di accuratezza di predizione.
Entrambi i metodi, infatti, assegnano alla conformazione secondaria di tipo filamento β il ruolo di unità fondamentale nelle strutture cross-β tipiche degli aggregati fibrillari, supportati in questo da numerosi studi sperimentali, e assumono che la propensione ad instaurare i legami idrogeno in grado di stabilizzare tale conformazione sia una proprietà intrinseca della catena polipeptidica, codificata nella struttura primaria.
Sia Palmo sia Pasta, dunque, basano le proprie predizioni unicamente sulla sequenza amminoacidica. Ambedue, inoltre, in fase di addestramento, estraggono dal medesimo training set, il Top500 Database, informazioni sulla tendenza di ciascuna coppia di amminoacidi ad instaurare legami idrogeno e ricavano da tali dati, con tecniche differenti, una funzione in grado di esprimere la propensione all'aggregazione di ognuna, detta **energia** nel caso di Pasta e **probabilità di aggregazione** nel caso di Palmo.
Come già accennato, ciò che differenzia sostanzialmente i due metodi, consentendo di qualificare Palmo come un'evoluzione migliorativa dell'approccio di Pasta, risiede nelle tecniche algoritmiche adottate.
Innanzitutto, nell'individuare le coppie di segmenti di più elevata tendenza all'aggregazione, Palmo si avvale di una efficiente tecnica basata sull'algoritmo di programmazione dinamica per il calcolo degli allineamenti, laddove Pasta si limita ad un approccio *a forza bruta* che passa in rassegna tutti i possibili accoppiamenti fra i residui della sequenza in esame. In secondo luogo, Palmo deriva la propria funzione di probabilità tramite una rete neurale non lineare, mentre la funzione di energia di Pasta è ricavata con semplici calcoli statistici sulla frequenza con cui ciascuna coppia di amminoacidi stabilisce un legame idrogeno di un certo tipo. Ancora più rilevante è il fatto che la rete neurale di Palmo prenda in considerazione non solo la coppia di singoli residui, ma anche il contesto circostante, grazie all'impiego di una finestra scorrevole di ampiezza massima 5, riuscendo

così a catturare una gamma di informazioni nettamente più ricca.
Come, d'altra parte, è lecito attendersi, le due funzioni - l'energia di PASTA e la probabilità di PALMO - esibiscono una marcata correlazione, in quanto derivate da una base di dati comune e finalizzate a catturare una stessa proprietà. Un confronto diretto è possibile nella sola condizione in cui le due funzioni sono effettivamente paragonabili, ovvero quando PALMO impiega un contesto nullo, limitandosi a considerare, così come PASTA, coppie di singoli residui.

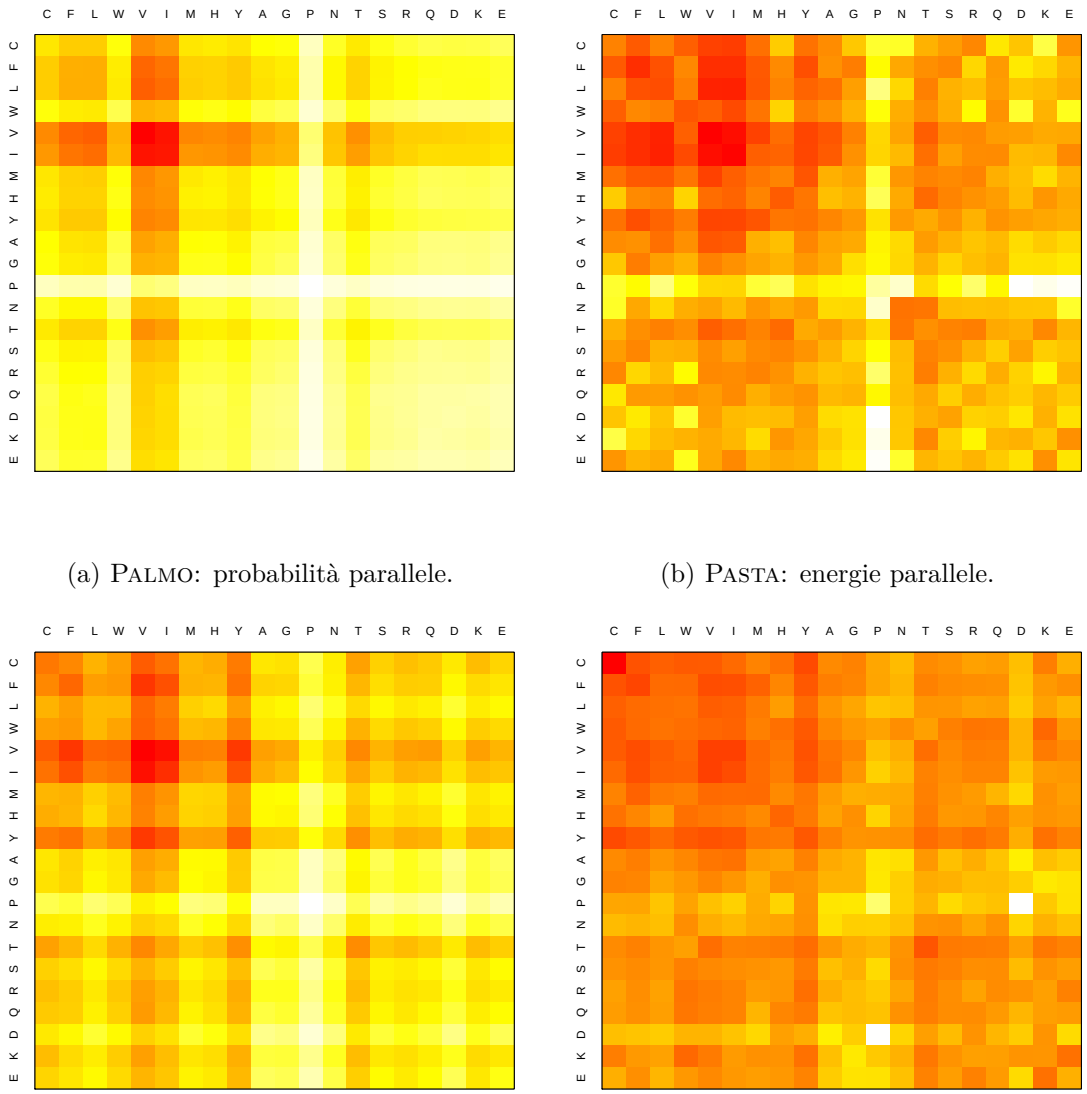
3.1.1 Un confronto tra le funzioni di aggregazione di PASTA e di PALMO

Una rappresentazione a *mappa di calore* (*heatmap*) fornisce una immediata impressione visuale del grado di correlazione fra le funzioni. Se, infatti, si associa a ogni possibile accoppiamento di amminoacidi una cella di una matrice 20×20, colorandola con una tonalità tanto più tendente al rosso quanto più la rispettiva funzione segnala una elevata propensione all'aggregazione, è facile constatare come i valori di energia di PASTA e di probabilità di PALMO esibiscano un andamento complessivamente simile (Figura 3.1).
In tutti e due i casi, gli amminoacidi valina (`V`) o isoleucina (`I`) si distinguono per una spiccata tendenza all'aggregazione, tanto che entrambi gli algoritmi attribuiscono valori tra i più elevati alle coppie (`V`, `V`), (`V`, `I`) e (`I`, `I`) (con l'unica eccezione di PASTA in orientazione antiparallela, dove queste sono superate, e di gran lunga, dalla sola coppia (`C`, `C`), Tabella 3.1). Anche la fenilalanina (`F`) ricorre con particolare frequenza in accoppiamenti aggreganti.
Per contro, l'amminoacido prolina (`P`) evidenzia una propensione all'aggregazione mediamente molto bassa, tanto da essere coinvolto negli accoppiamenti di valori minimo. Per ambedue i metodi, (`P`, `P`) e (`P`, `D`) risultano le coppie meno aggreganti in assoluto, con l'aggiunta, nel caso parallelo, di (`P`, `E`).
Queste osservazioni non solo mostrano una sostanziale coerenza di base tra PALMO e PASTA, ma risultano anche in accordo con gli studi sperimentali, che assegnano agli amminoacidi idrofobici, categoria cui appartengono proprio valina, isoleucina e fenilalanina, un ruolo preminente nel favorire la stabilizzazione di aggregati e, d'altra parte, riconoscono nella prolina un fattore di rottura dei legami idrogeno intracatena alla base della struttura cross-β amiloide.
La stretta affinità fra l'energia di PASTA e la probabilità di PALMO è confermata in termini quantitativi dal *coefficiente di Pearson* (PCC, *Pearson's correlation coefficient*)[1] fra

[1]Dette X e Y le variabili aleatorie date dai valori assunti, rispettivamente, dalla probabilità di PALMO e dalla funzione di energia di PASTA per ciascuna possibile coppia di residui, il *coefficiente di correlazione di Pearson* $\rho_{X,Y}$ è definito come

$$\rho_{X,Y} := \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \in [-1, 1]$$

dove μ indica la media e σ la deviazione standard di una variabile aleatoria. Un valore pari a 1 indica correlazione positiva totale, 0 nessuna correlazione, -1 correlazione negativa totale [Wik13d].



(a) PALMO: probabilità parallele.

(b) PASTA: energie parallele.

(c) PALMO: probabilità antiparallele.

(d) PASTA: energie antiparallele.

Figura 3.1: Confronto fra le rappresentazioni a mappa di calore delle probabilità di aggregazione di PALMO con contesto 0 (a sinistra) e dei potenziali di PASTA (a destra) fra tutte le coppie di amminoacidi in orientazione parallela (in alto) e antiparallela (in basso).

In realtà, poiché la funzione di energia di PASTA ha per codominio un intorno di 0 e (al contrario della probabilità di PALMO) assegna valori di energia decrescenti all'aumentare della propensione all'aggregazione, se ne sono considerati i valori cambiati di segno sia nel calcolo del PCC sia nella realizzazione della mappa

i valori assunti dalle due funzioni, che, sia nel caso parallelo sia in quello antiparallelo, segnala una correlazione molto marcata (Tabella 3.1).

		PCC	Massimo		Minimo	
			PALMO	PASTA	PALMO	PASTA
Orientazione	parallela	0.802304	(V, V)	(V, V)	(P, P)	(P, D)
	antiparallela	0.855720	(V, V)	(C, C)	(P, P)	(P, D)

Tabella 3.1: Coefficienti di correlazione di Pearson tra le matrici delle probabilità di aggregazione (normalizzate e non) di PALMO in orientazione parallela e antiparallela e le corrispondenti matrici dei potenziali di PASTA. Sono riportate, inoltre, le coppie di amminoacidi alle quali ciascun metodo assegna la massima/minima propensione all'aggregazione.

Si ribadisce, tuttavia, che quanto affermato in questa sezione non consente di cogliere appieno le potenzialità di PALMO: infatti, il confronto con PASTA ha imposto l'impiego del solo contesto nullo, trascurando così una delle migliorie più rilevanti apportate dal nuovo metodo.

3.2 Classificazione di peptidi

Obiettivo di questa sezione è fornire una stima quantitativa dell'accuratezza di PALMO come classificatore, valutata su un insieme di brevi peptidi con propensione all'aggregazione nota sperimentalmente, il già citato data set *Tango* (Sezione 2.2, pag. 21), secondo le tecniche descritte in Sezione 2.7.1, pag. 39. Come riferimento per una valutazione comparativa, si presentano i risultati ottenuti da tre analoghi classificatori: PASTA, BETASCAN e FoldAmyloid.

3.2.1 Curve ROC

Inizialmente, le prestazioni di PALMO sono state valutate senza applicare alcuna penalità (Sezione 2.5.3, pag. 34) al punteggio degli accoppiamenti fra segmenti (ovvero, $K = 0$). La curva ROC riferita al nostro metodo (Figura 3.2) esibisce un andamento sensibilmente migliore rispetto agli altri classificatori per tasso di falsi positivi fpr compreso nell'intervallo [1.8%, 39.3%].[2] Il vantaggio sui concorrenti raggiunge il massimo in corrispondenza di fpr 5%,[3] dove PALMO consegue un tasso di veri positivi tpr pari a 58.5%, superiore di

di calore.

[2] Si ricorda che un tasso di falsi positivi $fpr \in [1.8\%, 39.3\%]$ equivale ad una specificità $sp = (1 - fpr) \in [60.7\%, 98.2\%]$.

[3] L'esito della classificazione con tasso di falsi positivi $fpr = 5\%$ attraverso PALMO è riportato integralmente in Appendice A, Tabella A.2 pag. 79.

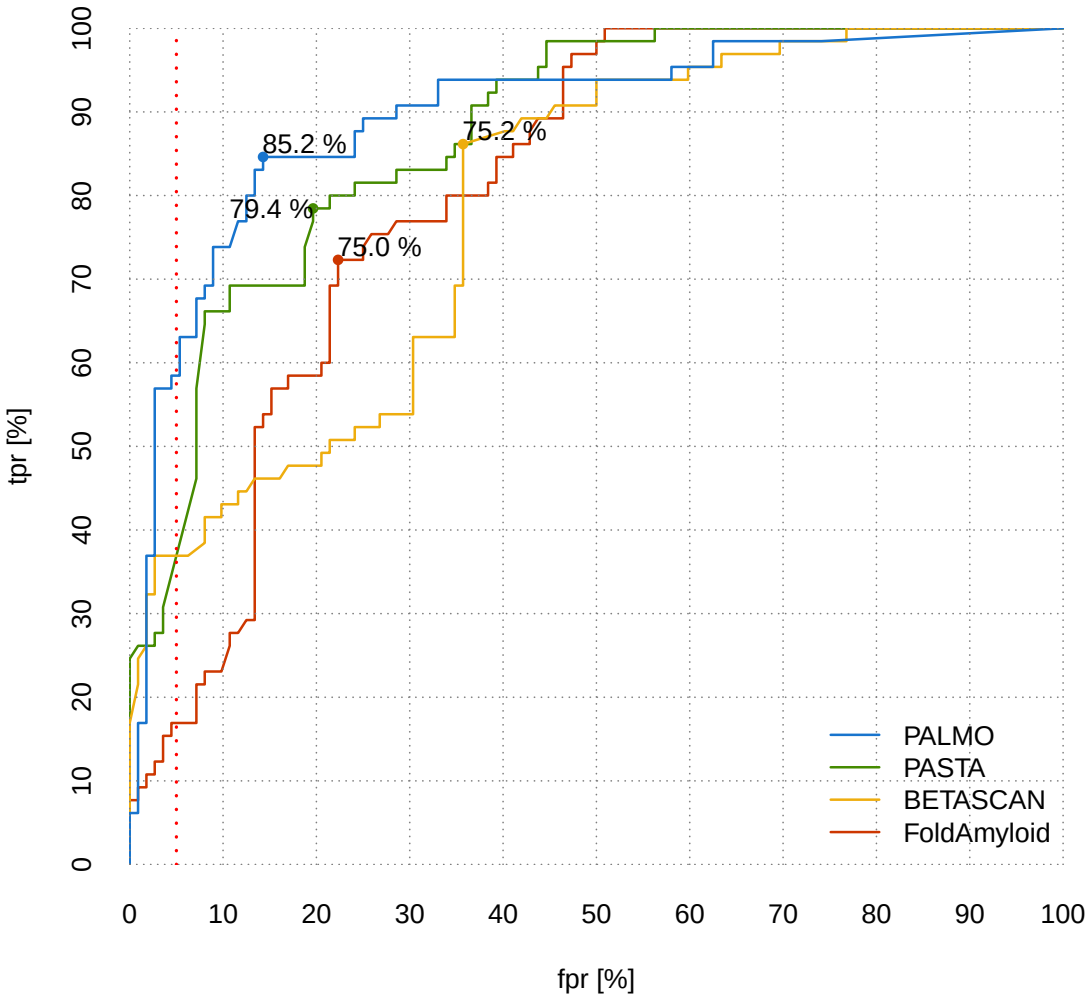


Figura 3.2: Curve ROC relative alla classificazione del data set *Tango* attraverso PALMO (senza penalità), PASTA, BETASCAN, FoldAmyloid. La linea rossa punteggiata evidenzia i punti di ascissa $fpr = 5\%$. I valori percentuali riportati nel grafico indicano la massima accuratezza bilanciata ottenuta da ciascun classificatore.

oltre il 20% rispetto a PASTA e BETASCAN ($tpr = 36.9\%$) e di più del 40% a FoldAmyloid ($tpr = 16.9\%$).
Al contrario, PALMO ottiene prestazioni leggermente inferiori per fpr compreso nell'intervallo [0%, 1.8%), dove il tasso di veri positivi risulta minore rispetto sia a PASTA sia a BETASCAN. Si tenga conto, tuttavia, che questo degrado è dovuto a due soli falsi positivi; sufficienti, però, ad influenzare la stima di accuratezza in misura significativa, a causa delle

ridotte dimensioni del test set. PALMO risulta in svantaggio anche per $fpr > 39.3\%$, dove l'alto numero di falsi positivi rende comunque poco attendibile l'esito della classificazione. Nel complesso, il nostro metodo è preferibile rispetto ai concorrenti nei casi in cui al classificatore sia richiesta un'**elevata specificità**, ossia un'alta probabilità che un peptide non amiloidogenico venga predetto come negativo. In queste situazioni, a parità di specificità, PALMO è in grado di riconoscere correttamente un numero maggiore di peptidi amiloidogenici rispetto agli altri classificatori: ha, in altri termini, una **migliore sensibilità**.
La capacità di conseguire ottimi risultati in termini di sensibilità senza penalizzare eccessivamente la specificità, requisito fondamentale di un buon classificatore, si riflette nell'**accuratezza bilanciata**, che con PALMO raggiunge un valore massimo di **85.2**%,[4] con quasi 6 punti percentuali di vantaggio su PASTA e almeno 10 sui restanti metodi.

Confronto tra gli esiti della classificazione attraverso PALMO e PASTA

Un raffronto tra i risultati ottenuti da PALMO e PASTA nella classificazione del data set *Tango* può aiutare a fare luce sulle relazioni esistenti tra i due metodi e sulle ragioni che determinano la migliore accuratezza del primo sul secondo. In particolare, si sono messe a confronto le classificazioni con tasso di falsi positivi pari a 5%, riportando in Tabella 3.2 i peptidi per cui PALMO e PASTA forniscono risultati discordanti.
Comparando le porzioni di sequenza evidenziate in rosso, salta all'occhio come, nella gran maggioranza dei casi, le regioni di massima amiloidogenicità individuate da ciascun algoritmo siano, quando non completamente sovrapponibili, differenti solo agli estremi, essendo, di fatto, centrate nel medesimo gruppo di amminoacidi. Quest'osservazione fornisce un'ulteriore conferma della forte correlazione esistente tra la funzione di probabilità di PALMO e la funzione energia di PASTA, già emersa in precedenza (Sezione 3.1, pag. 45).
Altro aspetto che accomuna i due metodi è la netta prevalenza dell'**allineamento parallelo in registro**, in accordo con varie evidenze sperimentali: molteplici studi, in effetti, hanno evidenziato come la struttura delle fibrille amiloidi sia frequentemente caratterizzata da foglietti β paralleli i cui filamenti si allineano in registro, cosicché residui dello stesso tipo (idrofobici o idrofilici) si impilano l'uno sull'altro in file disposte lungo l'asse della fibrilla [TCMS06, NSB$^+$05, NE06]. Anche l'orientamento antiparallelo in registro è presente, benchè meno comune (si vedano, ad esempio, i peptidi E1 ed E). In quest'ultimo caso, tuttavia, emergono alcune incongruenze: è il caso dei frammenti K19Gluc782, K19Gluc41 e K19, per i quali i due metodi, pur individuando, in sostanza, il medesimo segmento amiloidogenico (`KVQIVYK`/`VQIVY`), sono in disaccordo nello stabilire l'allineamento più favorevole alla formazione di aggregati (antiparallelo in registro per PALMO, parallelo in registro per PASTA). In base agli studi citati in precedenza, gli allineamenti fuori registro, al contrario, tendono ad ostacolare la formazione delle strutture altamente ordinate alla base delle fibrille amiloidi. Non a caso, essi compaiono di rado tra i risultati dei due metodi, spesso in corrispondenza di errori di classificazione: tra gli esempi riportati in

[4] L'accuratezza bilanciata massima $ba_{\max} = 85.2\%$ corrisponde ad una sensibilità $se = 84.6\%$ e ad una specificità $sp = 85.7\%$.

tabella, allineamenti fuori registro si presentano solo due volte, associati nel primo caso ad un falso positivo (peptide E-Helix), nel secondo ad un frammento (H2 Mt) cui PALMO assegna (erroneamente) una scarsa propensione all'aggregazione.
Il vantaggio in termini di accuratezza conseguito da PALMO rispetto a PASTA è determinato dal minor numero di errori di predizione commessi dal primo: nel caso considerato in tabella, PALMO classifica correttamente 23 peptidi per i quali PASTA produce un esito errato, mentre la situazione inversa si verifica in soli 6 casi. Ciò che distingue i due metodi non sono tanto le regioni di massima amiloidogenicità relativa all'interno dei frammenti (che, come già visto, sono spesso quasi coincidenti), quanto i differenti punteggi attribuiti ad ogni peptide e dai quali dipende direttamente il diverso esito della classificazione di questo.
In particolare, il fatto che PASTA prenda in considerazione ciascuna coppia di residui singolarmente, trascurando il contesto circostante, fa sì che a *pattern* amminoacidici identici non possa che assegnare la medesima energia di aggregazione, con la conseguenza che a peptidi che condividono pattern uguali viene attribuito lo stesso punteggio e, quindi, la stessa classe, anche quando i dati sperimentali assegnino loro differenti gradi di amiloidogenicità. Si osservino, ad esempio, i peptidi HABP15, HABP3, AB4, HABP1 (sperimentalmente propensi all'aggregazione), HABP2, AB3 e AB2 (sperimentalmente non aggreganti), ricavati dalla proteina β-amiloide: PASTA ne isola il medesimo segmento (`VQIVY`), attribuendo a tutti un punteggio uguale e classificandoli indistintamente come amiloidogenici.
Al contrario, l'osservazione del contesto consente a PALMO di differenziare la propria predizione, tenendo conto del fatto che peptidi simili, nonostante condividano pattern identici, possono avere propensioni all'aggregazione differenti a causa delle differenze tra le sequenze al di fuori di tali pattern. Per esempio, i peptidi K19Gluc782 e K19Gluc41 contengono la medesima sottosequenza `...KVQIVYK`, seguita da un acido glutammico (`E`) nel primo caso, da una prolina (`P`) nel secondo: PALMO è in grado di tenere conto del diverso effetto che tali residui determinano sulla propensione all'aggregazione (negativo nel caso della prolina, come confermato dai dati sperimentali) e, pertanto, assegna al primo un punteggio maggiore rispetto al secondo (laddove PASTA calcolava la stessa energia per entrambi). Anche i frammenti K19 e PHF8 ricevono punteggi differenti, nonostante PALMO rilevi in entrambi lo stesso pattern. Quanto visto consente di affermare che, grazie all'influenza del contesto, i punteggi di aggregazione attribuiti da PALMO si distinguono per una granularità più fine; caratteristica che, in linea di principio, può tradursi in una migliore accuratezza di classificazione.

Peptide	Palmo				Pasta			
	Sequenza	**Allineam.**	**Punt.**	**Posiz.**	**Sequenza**	**Allineam.**	**Punt.**	**Posiz.**
NAC1-18s	TVNGVGEVTATAVQGVAV	Pir	155.99	24	TVNGVGEVTATAVQGVAV	Pir	−2.8665	105
NAC6-14	VGGAVVTGV	Pir	226.58	6	VGGAVVTGV	Pir	−4.1766	69
E1	DWSFYLLYYTEFTPTGKDEYA	Air	205.46	10	DWSFYLLYYTEFTPTGKDEYA	Air	−4.1847	68
E	DWSFYLLYYTEFT	Air	195.53	11	DWSFYLLYYTEFT	Air	−4.1847	67
NAC1-13	EQVTNVGGAVVTG	Pir	180.40	17	EQVTNVGGAVVTG	Pir	−4.1269	73
Pc-14	GETYVVTL	Pir	182.22	16	GETYVVTL	Pir	−4.6969	57
Pc-12	MPEEELLNAPGETYVVTL	Pir	182.22	15	MPEEELLNAPGETYVVTL	Pir	−4.6969	56
NAC1-18	EQVTNVGGAVVTGVTAVA	Pir	327.61	2	EQVTNVGGAVVTGVTAVA	Pir	−5.3222	39
NAC3-18	VTNVGGAVVTGVTAVA	Pir	326.61	3	VTNVGGAVVTGVTAVA	Pir	−5.3222	40
K19Gluc782	NLKHQPGGGKVQIVYKEVD	Air	177.59	18	NLKHQPGGGKVQIVYKEVD	Pir	−5.0845	49
K19Gluc41	NLKHQPGGGKVQIVYKPVDLSK...				NLKHQPGGGKVQIVYKPVDLSK...			
	...VTSKCGSLGNIHHKPGGGQVE	Air	166.05	23	...VTSKCGSLGNIHHKPGGGQVE	Pir	−5.0845	48
K19	PGGGKVQIVYKPV	Pir	166.05	22	PGGGKVQIVYKPV	Pir	−5.0845	47
PHF8	GKVQIVYK	Air	150.91	28	GKVQIVYK	Pir	−5.0845	50
B helix	SAPNLATLVKVTTNHFTHEEAMMD	Pir	150.80	29	SAPNLATLVKVTTNHFTHEEAMMD	Pir	−5.1636	42
HABP15	KKLVFFAED	Pir	153.13	25	KKLVFFAED	Pir	−5.3691	37
HABP3	VHHPKLVFFAEDVGS	Pir	152.55	26	VHHPKLVFFAEDVGS	Pir	−5.3691	31
34-53	GVVGWVKNTSKGTVTGQVQG	Pir	139.35	43	GVVGWVKNTSKGTVTGQVQG	Pir	−5.1176	43
AB4	HHQKLVFFAED	Pir	146.61	39	HHQKLVFFAED	Pir	−5.3691	38
HABP1	VPHQKLVFFAEDVGS	Pir	146.61	38	VPHQKLVFFAEDVGS	Pir	−5.3691	29
HABP2	VHPQKLVFFAEDVGS	Pir	146.35	40	VHPQKLVFFAEDVGS	Pir	−5.3691	30
Pc-9	PHNVVFDEDDEIP	Pir	127.86	53	PHNVVFDEDDEIP	Pir	−6.0276	22
D-helix	AKNVDYCKEWLVNHIK	Pir	65.13	89	AKNVDYCKEWLVNHIK	Pir	−6.0866	19
D-helix	AKNVDYCKEWLVNHIK	Pir	65.13	90	AKNVDYCKEWLVNHIK	Pir	−6.0866	20
E-Helix	SEDLKKHGVTVLTALGAILK	P	250.48	5	SEDLKKHGVTVLTALGAILK	Pir	−4.8666	54
K19Chym	QTAPVPMPDLKNVKSKIGSTEN...				QTAPVPMPDLKNVKSKIGSTEN...			
	...LKHQPGGGKVQIVY	Pir	182.96	13	...LKHQPGGGKVQIVY	Pir	−5.0845	46
AB3	HQKLVFFAE	Pir	145.48	41	HQKLVFFAE	Pir	−5.3691	28
AB2	QKLVFFA	Pir	140.82	42	QKLVFFA	Pir	−5.3691	27
Pc-6a	GEKIVFKNNAGFPHNVVFDE	Pir	129.74	52	GEKIVFKNNAGFPHNVVFDE	Pir	−6.4616	6
H2 Mt	FVNVEAVKAFLEAHGIAY	P	51.05	98	FVNVEAVKAFLEAHGIAY	Pir	−5.4917	24

Legenda: Vero positivo | Falso positivo | Vero negativo | Falso negativo

Air: antiparallelo in registro Pir: parallelo in registro A: antiparallelo (fuori registro) P: parallelo (fuori registro)

Tabella 3.2: La tabella riporta i peptidi del data set *Tango* per i quali la classificazione operata da PALMO differisce da quella effettuata da PASTA. Per ambo i metodi, si è considerata la classificazione caratterizzata da un tasso di falsi positivi $fpr = 5\%$. La porzione superiore della tabella mostra i peptidi che PALMO classifica correttamente (ossia, in accordo con i dati sperimentali) e PASTA in modo errato; viceversa la porzione inferiore della tabella. Per ogni peptide e per ciascuno dei due classificatori, è evidenziata in rosso la coppia di segmenti di massima propensione all'aggregazione (qualora i segmenti accoppiati siano fuori registro, quindi non coincidenti, questi sono distinti da due linee rosse), indicando il tipo di allineamento (parallelo o antiparallelo, in registro o meno) e il punteggio di aggregazione di questa, nonché l'esito della classificazione del frammento peptidico (rappresentato dal colore della riga) e la posizione all'interno del test set in ordine di punteggio non crescente. Gli identificatori dei peptidi fanno riferimento alla Tabella A.1, pag. 75.

Penalità non nulla

La classificazione del data set *Tango* è stata ripetuta con l'applicazione di una penalità unitaria $K = 11.4$ (valore ricavato tramite il processo di ottimizzazione descritto in Sezione 2.5.3, pag. 34) al punteggio delle coppie di segmenti di lunghezza superiore a 6 residui. La relativa curva ROC (Figura 3.3) mostra un sensibile peggioramento rispetto al caso

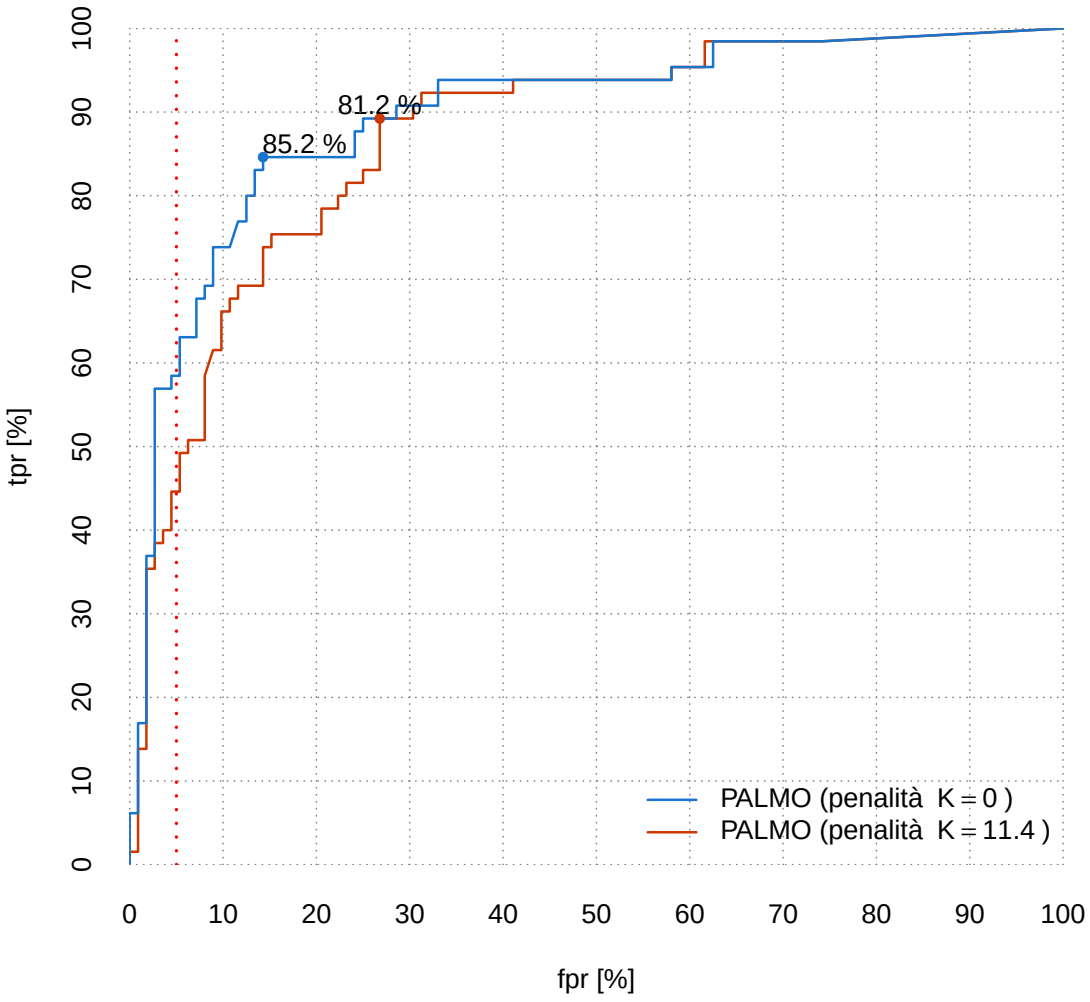


Figura 3.3: Confronto tra le curve ROC riferite alla classificazione del data set *Tango* tramite PALMO senza penalità ($K = 0$) e PALMO con penalità unitaria $K = 11.4$ per accoppiamenti di lunghezza superiore a 6.

$K = 0$. A parità di specificità, la sensibilità risulta quasi ovunque inferiore: in particolare, in corrispondenza di $fpr = 5\%$, il tasso di falsi positivi $tpr = 44.6\%$ fa registrare un decremento di quasi 14 punti percentuali (benché rimanga ancora superiore di oltre il 7% rispetto a PASTA e BETASCAN)[5]. Ne risente, com'è ovvio, l'accuratezza bilanciata, il cui valore massimo diminuisce all'81.2% (-4% rispetto al caso $K = 0$), pur mantenendo un leggero vantaggio su PASTA.
Quanto visto finora evidenzia come una penalità sul punteggio così concepita non si presti ad essere applicata al problema di classificare brevi peptidi. Essa, infatti, trae origine dall'osservazione che, all'interno di sequenze proteiche intere (nel caso di specie, quelle componenti il Top500 Database), le regioni coinvolte nella formazione di foglietti β si estendono raramente per più di sei residui ed è stata introdotta in PALMO con l'obiettivo di penalizzare quelle coppie di segmenti che, eccedendo tale soglia di lunghezza, hanno, nella realtà, scarsa probabilità di dare luogo a fenomeni di aggregazione. Pertanto, l'applicazione della penalità produce i risultati sperati nell'individuazione di regioni amiloidogeniche all'interno di proteine complete, come si vedrà in Sezione 3.3, pag. 57.
Di natura ben diversa è il problema della classificazione binaria: qui l'obiettivo consiste nell'assegnare un frammento peptidico, considerato nella sua interezza, ad una delle due classi *amiloidogenico* e *non amiloidogenico*, in base al confronto tra il suo punteggio di aggregazione complessivo e una soglia di classificazione prefissata. PALMO, in questo ambito, viene impiegato al solo scopo di ricavare questo *score* complessivo, che, di fatto, coincide con il punteggio dell'accoppiamento fra sottosequenze del peptide di massima propensione all'aggregazione. Si può ipotizzare che il degrado nell'accuratezza abbia origine dal fatto che la penalità agisce sul punteggio del singolo accoppiamento fra sottosequenze in base all'estensione di quest'ultimo, anziché dipendere dalla lunghezza dell'intero frammento peptidico oggetto della classificazione. In quest'ottica, la decisione se applicare o meno un qualche correttivo al punteggio complessivo e l'eventuale entità di questo dovrebbero essere stabilite in base a proprietà globali del peptide (ad esempio, la lunghezza totale).
Tuttavia, i più che soddisfacenti risultati ottenuti nella classificazione senza l'impiego di alcuna penalità hanno suggerito di lasciare la verifica di questa ipotesi ad una futura fase di affinamento del metodo.

3.2.2 Stima tramite test t di Student delle aree sottese alle curve ROC

La stima delle aree alle curve ROC tramite test t di Student (Sezione 2.7.1, pag. 39) fornisce un'ulteriore conferma alle precedenti considerazioni sull'accuratezza di PALMO rispetto agli altri classificatori in esame.
I valori riportati in Tabella 3.3 descrivono intervalli di confidenza $99\% = 100\,(1 - \alpha)\,\%$, con $\alpha = 0.01$. Poiché il valore p associato ai test t è risultato $p \approx 0 < \alpha$,[6] questi risultati

[5] Per un raffronto diretto tra la curva ROC relativa a PALMO con $K = 11.4$ e i classificatori considerati in precedenza, si rimanda all'Appendice A, Figura A.1, pag. 80.

[6] In realtà, la funzione `t.test` del software di analisi statistica R (`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/t.test.html`), impiegata per eseguire i test t di Student, ha restituito un

		[**0%**, **5%**]		[**0%**, **10%**]	
		$\bar{x}$ $[\times 10^{-2}]$	E $[\times 10^{-6}]$	$\bar{x}$ $[\times 10^{-2}]$	E $[\times 10^{-6}]$
Palmo	$K = 0$	1.89308	3.76	5.246887	5.48
	$K = 11.4$	1.431200	3.01	4.157387	5.30
Pasta		1.473254	2.11	4.177326	5.35
BetaScan		1.546491	1.80	3.516022	3.41
FoldAmyloid		0.5923699	1.225	1.630516	2.96

		[**0%**, **20%**]		[**0%**, **100%**]	
		$\bar{x}$ $[\times 10^{-2}]$	E $[\times 10^{-6}]$	$\bar{x}$ $[\times 10^{-2}]$	E $[\times 10^{-6}]$
Palmo	$K = 0$	13.39571	7.7	89.73134	13.9
	$K = 11.4$	11.35886	8.1	87.17668	15.0
Pasta		11.20818	7.9	86.17192	13.8
BetaScan		8.117919	6.58	77.25052	18.5
FoldAmyloid		6.284644	8.90	78.98623	17.4

Tabella 3.3: Confronto tra le AUC calcolate, rispettivamente, sugli intervalli di fpr $[0\%, fpr_{\max}]$, con $fpr_{\max} \in \{5\%, 10\%, 20\%, 100\%\}$, per i classificatori in esame. Il range $[\bar{x} - E, \bar{x} + E]$ rappresenta l'intervallo di confidenza 99% per il valore della AUC, ricavato tramite test t di Student. Un valore $p \approx 0$ conferisce significatività statistica ai test.

sono statisticamente significativi.
I valori complessivi (riferiti, cioè, alle intere curve ROC, per $fpr \in [0\%, 100\%]$) attribuiscono a Palmo (con $K = 0$, ossia senza l'applicazione di penalità sul punteggio) la massima capacità discriminante tra i classificatori in esame. In altri termini, estratti a caso un peptide amiloidogenico e uno non amiloidogenico da un insieme qualsiasi di frammenti proteici, il primo riceverà da Palmo un punteggio di aggregazione maggiore del secondo con probabilità vicina al 90%, con oltre 3.5 punti percentuali di vantaggio su Pasta e quasi 11 su FoldAmyloid.
Focalizzando l'attenzione sulle AUC parziali corrispondenti ad elevati valori di specificità, il divario di Palmo sui concorrenti si amplia ulteriormente.[7] Per tasso di falsi positivi $fpr \leq 5\%$, ovvero specificità $sp \geq 95\%$, la AUC di Palmo è di quasi il 7% maggiore del

valore $p < 2.2 \times 10^{-16}$: di fatto, essendo tale valore il più piccolo numero in virgola mobile rappresentabile, è possibile approssimare p con 0.

[7] Nel seguito, la grandezza base dei valori percentuali è l'estensione dell'area sottesa alla curva ROC del classificatore perfetto, pari a $fpr_{\max}$ per l'intervallo $[0, fpr_{\max}]$.

secondo migliore classificatore, che qui, eccezionalmente, non è PASTA, come in tutti gli altri casi, ma BETASCAN. Allentando il vincolo sulla specificità ($sp \geq 90\%$) e considerando un tasso di falsi positivi $fpr \leq 10\%$, il nostro metodo mostra oltre 10.5 punti percentuali di vantaggio rispetto a PASTA, che crescono a quasi 11 se si estende ancora l'analisi fino a $fpr_{\max} = 20\%$, cioè specificità $sp \geq 80\%$.
Un discorso a parte merita PALMO con penalità $K = 11.4$, che conferma un sensibile degrado rispetto al caso $K = 0$, esibendo valori di AUC di fatto in linea con quelli di PASTA, rispetto al quale mostra solo un modesto incremento in termini di capacità discriminante complessiva.
In definitiva, l'analisi delle AUC non fa altro che conferire significatività statistica a quanto affermato nell'osservare le curve ROC: PALMO ($K = 0$) mostra un'accuratezza di classificazione nettamente superiore agli altri predittori in esame, specialmente nelle situazioni in cui sia richiesta una specificità elevata.

3.3 Individuazione di regioni amiloidogeniche

Si considera, ora, il secondo problema cui PALMO ambisce a dare risposta: data la sequenza primaria completa di una proteina amiloide, isolare le regioni responsabili di innescare processi di aggregazione.
Le metodologie qui adottate e le misure in base a cui si sono valutati i risultati sono illustrate in Sezione 2.7.2, pag. 42. Il test set di proteine amiloidi è, invece, descritto in Sezione 2.3, pag. 21.
È importante notare che i risultati riportati nel seguito si riferiscono a PALMO con **penalità unitaria $\mathbf{K = 11.4}$** (Sezione 2.5.3, pag. 34): infatti, le verifiche condotte nel corso dello sviluppo hanno evidenziato come penalizzare il punteggio delle coppie di segmenti in relazione alla loro lunghezza garantisca una maggiore accuratezza in questo specifico problema, in accordo con quanto affermato in Sezione 3.2.1, pag. 54. D'altra parte, si ricorda che il valore unitario della penalità $K = 11.4$ è frutto di una ottimizzazione basata sul rilevamento, all'interno di sequenze proteiche complete, di strutture β stabilizzate da legami idrogeno: problema sostanzialmente affine, in base all'assunzione biochimica fondamentale di PALMO, al rilevamento di segmenti aggreganti. Tale analogia, oltre a motivare la procedura di ottimizzazione, giustifica l'impiego, qui, del valore che ne è risultato.

3.3.1 Verifica delle prestazioni di PALMO

In primo luogo, si focalizza l'attenzione su PALMO, verificando l'accuratezza delle predizioni corrispondenti agli n accoppiamenti non ridondanti di punteggio massimo, al variare di n tra 1 e 10. Una rappresentazione grafica dei risultati è mostrata in Figura 3.4, pag. 59, e 3.6, pag. 61, mentre una valutazione quantitativa dell'accuratezza di predizione in termini di punteggio di sovrapposizione SOV è riportata in Tabella 3.4, pag. 58.
Si osservi come, al crescere di n, la predizione si estenda su porzioni di sequenza via via più estese, il che consegue dalla scelta di scartare gli accoppiamenti ridondanti. In particolare, la predizione relativa a $n = 1$ corrisponde esattamente ai segmenti che compongono la

n	1	2	3	4	5	6	7	8	9	10
β-amiloide	66.0	66.7	67.5	72.5	72.7	71.1	72.7	61.4	58.9	51.9
HET-s	25.2	26.6	43.2	45.5	45.7	40.7	38.2	39.6	52.3	51.6
FgHET-s	28.3	30.3	27.6	28.4	40.2	41.0	48.1	47.6	42.8	44.7
Amilina	51.4	50.7	46.2	49.8	38.3	49.7	47.4	50.2	45.6	47.0
α-sinucleina	53.0	54.5	75.9	75.9	76.1	77.3	78.2	65.6	65.9	66.1
Proteina τ	44.1	59.4	59.3	65.2	65.6	65.9	66.0	60.9	61.1	61.1
SOV complessivo	44.0	53.0	57.8	61.7	62.2	62.6	63.0	58.3	58.6	58.5
SOV medio	44.7	48.0	53.3	56.2	56.4	57.6	58.4	54.2	54.4	53.7

Tabella 3.4: Punteggi SOV associati alle predizioni effettuate da PALMO per le sei proteine amiloidi del test set al variare di $n \in \{1, \ldots, 10\}$. *SOV medio* rappresenta la media aritmetica dei punteggi SOV relativi a ciascuna proteina per n fissato, mentre *SOV complessivo* coincide con la misura $\overline{\text{SOV}}$ definita al termine di Sezione 2.7.2, pag. 43. In rosso sono evidenziati i valori migliori al variare di n.

coppia con il massimo punteggio di aggregazione; segmenti che, a causa della già evidenziata prevalenza degli allineamenti in registro, molto frequentemente risultano coincidenti, ad individuare, di fatto, un'unica regione.

β-amiloide Fin da $n = 1$, PALMO rileva correttamente la seconda delle due regioni che i dati sperimentali qualificano come amiloidogeniche[8] (residui 30-41).[9] A partire da $n = 6$, anche la prima (residui 17-25) può ritenersi individuata, benché l'accuratezza sia penalizzata dal fatto che la predizione si estenda erroneamente sulla porzione di sequenza che separa i due segmenti sperimentali. In linea di massima, la predizione tende a concentrarsi sulla seconda metà della sequenza, quella effettivamente coinvolta nei fenomeni di aggregazione, come testimonia l'elevato punteggio SOV (72.7) ottenuto in corrispondenza di $n = 7$. Tra l'altro, PALMO riesce a prevedere l'allineamento parallelo in registro con cui copie corrispondenti dello stesso segmento amiloidogenico in catene distinte si assemblano fra loro nel generare gli aggregati fibrillari caratteristici di questo amiloide (Figura 2.3, pag. 22).

HET-s Degli otto filamenti β che, innescando il processo di aggregazione, determinano la complessa struttura di questo amiloide (Figura 2.4, pag. 22), il nostro metodo prevede fin da subito la penultima (residui 54-59), cui si aggiunge, per $n \geq 3$, la seconda (residui 25-28) e, per $n \geq 8$, la sesta (residui 48-52). Si rivela, invece, un falso positivo il segmento individuato tra i residui 1-6. In definitiva, PALMO rileva 3 regioni amiloidogeniche su 8, più un falso positivo, con un punteggio SOV che raggiunge il valore di 52.3 per $n = 9$.

[8] D'ora in poi, in breve, "regioni sperimentali".

[9] Gli indici numerici, in base zero, indicano gli estremi della regione sperimentale.

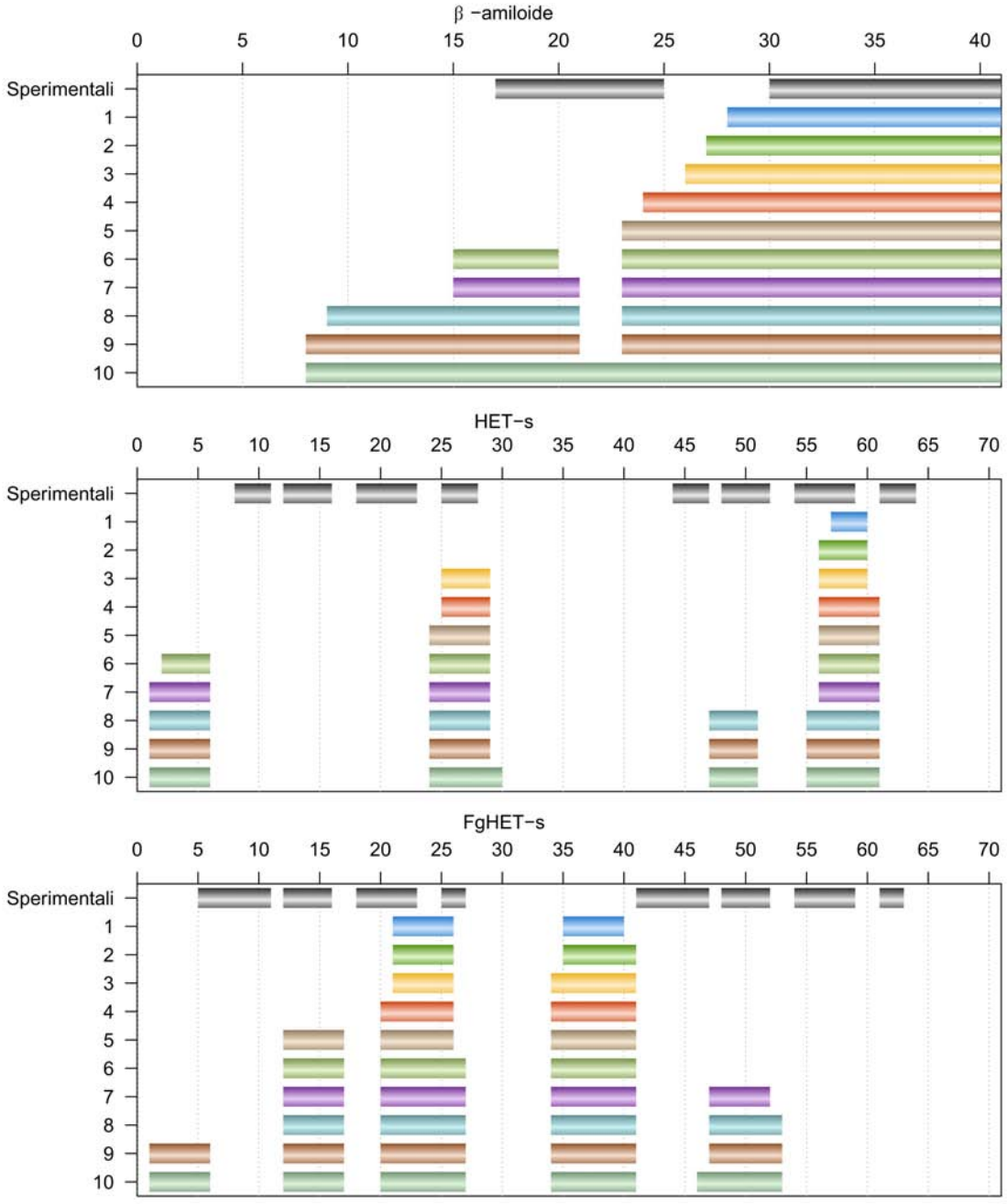


Figura 3.4: Rappresentazione grafica delle regioni amiloidogeniche predette da PALMO per le prime tre proteine del test set. La riga superiore mostra l'estensione dei segmenti di elevata propensione all'aggregazione in base ad evidenze sperimentali, mentre le dieci successive rappresentano le predizioni derivate dalle n migliori coppie non ridondanti restituite da PALMO, $n \in \{1, \ldots, 10\}$.

D'altra parte, si tenga conto che la particolare disposizione dei segmenti sperimentali, i quali, ad esclusione dell'ampia cesura centrale, sono intervallati da non più di un residuo, rende particolarmente arduo il compito a qualsiasi predittore.

FgHET-s Questa proteina, remota omologa della precedente, presenta una struttura tridimensionale molto simile, pur esibendo una bassa similarità di sequenza (38%): si tratta, per questi motivi, di un caso di particolare interesse. Insolitamente, l'accoppiamento di massimo punteggio di aggregazione restituito da PALMO ($n = 1$) è fuori registro ed individua due segmenti distinti, dei quali il primo copre simultaneamente due regioni sperimentali intervallate da un solo amminoacido (residui 18-23 e 25-27), il secondo provoca un falso positivo. I segmenti sperimentali di estremi 12-16 e 48-52, che nella conformazione 3D a solenoide β assunta dall'amiloide risultano reciprocamente legati, vengono entrambi predetti con notevole accuratezza, il primo per $n \geq 5$, il secondo per $n \geq 7$. Con $n = 9$, viene introdotto un nuovo falso positivo. Fermo restando quanto affermato al punto precedente sulla difficoltà nel predire una struttura di questa complessità, le regioni sperimentali individuate sono 4 su 8, con 2 falsi positivi ed un punteggio SOV che, a causa di questi ultimi, non raggiunge 50.

Amilina (IAPP) La corrispondenza tra la predizione di PALMO ed il modello cui si fa riferimento in Figura 3.6, pag. 61 [LYLT07], appare piuttosto scarsa: il nostro predittore sembra segnalare con sufficiente accuratezza solo il primo segmento sperimentale (residui 7-16), conseguendo un modesto 51.4 come punteggio SOV in corrispondenza di $n = 1$. Per questo amiloide, tuttavia, esiste un modello alternativo (Figura 3.5), discordante con il

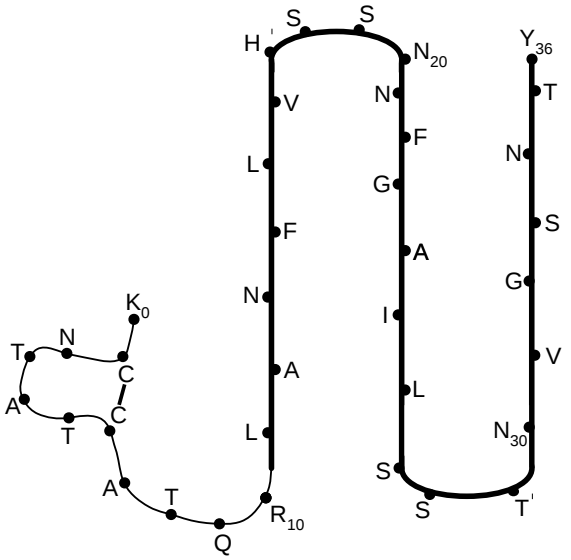


Figura 3.5: Modello alternativo per l'amilina, che propone un ripiegamento a serpentina composto da tre filamenti β. [KAS05]

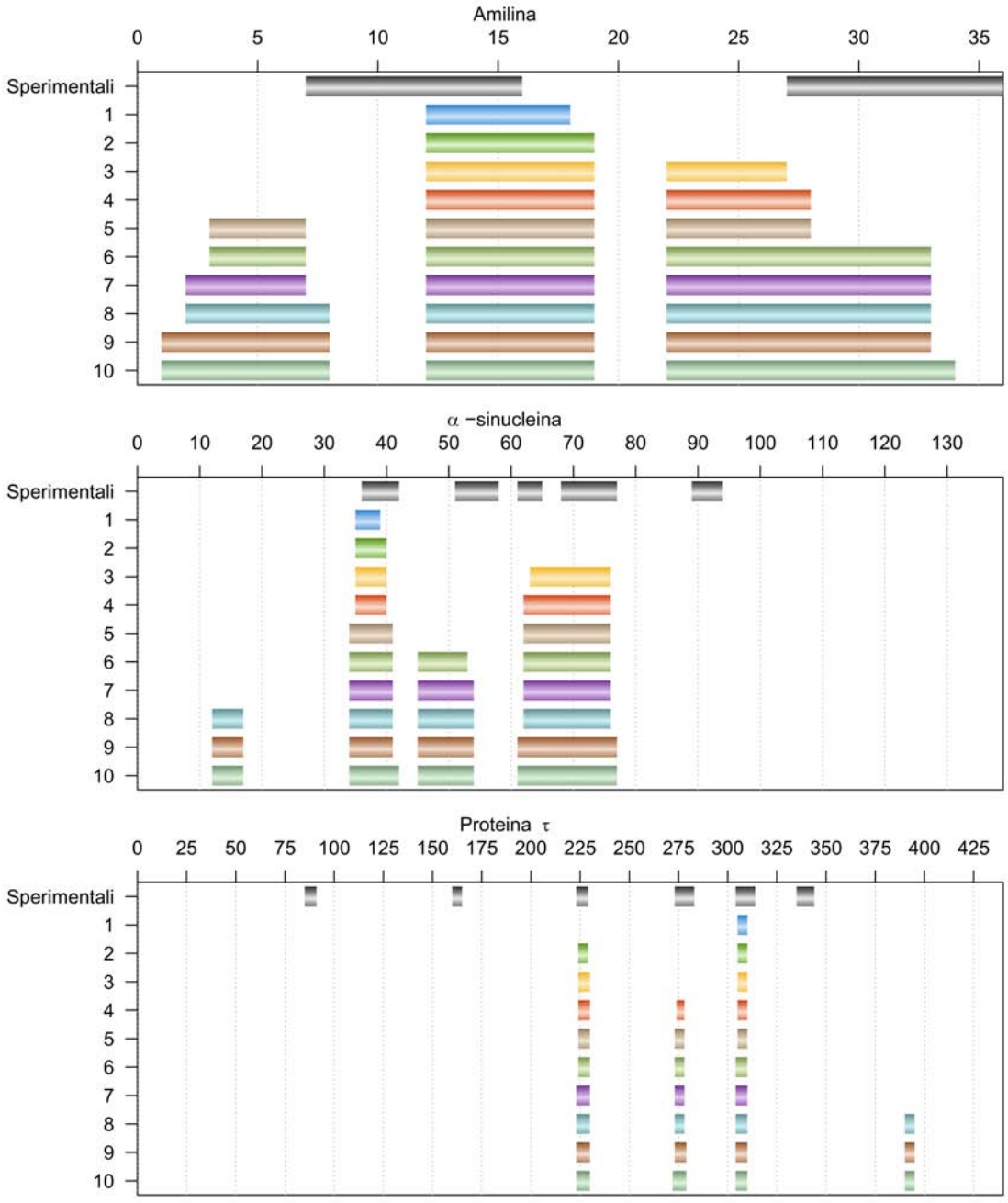


Figura 3.6: Rappresentazione grafica delle regioni amiloidogeniche predette da PALMO per le restanti proteine del test set.

primo, che descrive una conformazione a serpentina composta da tre filamenti β, coincidenti all'incirca con i segmenti di estremi 11-17, 20-27 e 30-36; la struttura tridimensionale della fibrilla è data dall'accumulo in registro di molteplici repliche della serpentina, dove filamenti adiacenti si legano a formare tre foglietti β paralleli in registro [KAS05]. Se posta a confronto con quest'ultimo modello, la predizione vede un sensibile miglioramento in accuratezza: infatti, i segmenti individuati da PALMO per $n = 1$ (12-18) e per $n = 3$ (22-27) ricalcano quasi alla perfezione i primi due filamenti β sperimentali (nel caso del primo, tra l'altro, il nostro metodo intuisce anche la disposizione parallela che le sue repliche assumono nel formare uno dei foglietti β componenti le fibrille); il falso positivo introdotto con $n = 5$ all'inizio della sequenza penalizza la specificità della predizione, ma, in compenso, per $n \geq 6$ PALMO riesce ad individuare con buona accuratezza anche il terzo filamento β sperimentale. Predendo a riferimento il secondo modello, dunque, PALMO rileva 3 regioni sperimentali su 3 con un falso positivo.

α-sinucleina Il primo segmento sperimentale (residui 36-42) viene individuato con sufficiente accuratezza fin da $n = 1$, così come il suo orientamento parallelo in registro nella struttura degli aggregati fibrillari. Con $n \geq 3$, anche la terza (residui 61-65) e la quarta (residui 68-77) regione sperimentali possono ritenersi rilevate, così come l'allineamento parallelo in registro delle rispettive repliche, benché la predizione le comprenda in un unico segmento continuo. A partire da $n = 6$, e ancor più per $n \geq 7$, PALMO sembra cogliere un segnale dell'esistenza della seconda regione sperimentale (residui 51-58), mentre con $n = 8$ introduce un segmento che non trova riscontro nei dati sperimentali. Nel complesso, il nostro metodo individua chiaramente 3 regioni sperimentali su 5 e marginalmente una ulteriore, con un falso positivo; al netto di quest'ultimo, ovvero per $n = 7$, PALMO consegue un notevole punteggio SOV di 78.2.

Proteina τ Questa lunga proteina presenta sei regioni amiloidogeniche, ma gli studi sperimentali assegnano a due particolari esapeptidi un ruolo cruciale nell'innescare il processo di aggregazione: ^{274}`VQIINK`^{279} e ^{305}`VQIVYK`^{310} (nuclei centrali delle regioni sperimentali che si estendono, rispettivamente, tra i residui 273-283 e 304-314). PALMO individua alla perfezione fin da subito ($n = 1$) ^{305}`VQIVYK`^{310}, cui, insolitamente, assegna un allineamento sì in registro, ma antiparallelo; a partire da $n = 4$, anche l'esapeptide ^{274}`VQIINK`^{279} viene rilevato con grande accuratezza, ma con un orientamento parallelo in registro. Con $n \geq 2$, il nostro metodo segnala la presenza della terza regione sperimentale (residui 223-229), mentre per $n = 8$ introduce un falso positivo nella porzione finale della sequenza. PALMO, dunque, rileva con notevole accuratezza 3 segmenti sperimentali su 5, con un falso positivo, trascurando il quale (prendendo, cioè, $n = 7$), ottiene un punteggio SOV pari a 66.

Sia la media aritmetica dei punteggi SOV (pari a 63) sia il valore complessivo pesato per la lunghezza di ciascuna proteina (oltre 58) portano a concludere che PALMO raggiunge la massima accuratezza di predizione per $n = 7$.

3.3.2 Confronto con i predittori concorrenti

Una volta verificate le prestazioni assolute di PALMO, si procede ad un raffronto con gli analoghi metodi già presi a riferimento in precedenza (PASTA, BETASCAN, FoldAmyloid). Per garantire la massima equità, le predizioni di questi sono state ricavate secondo le medesime modalità impiegate per PALMO nella sezione precedente, ovvero considerando, per ogni amiloide del test set, gli $n = 7$ segmenti proteici non ridondanti cui ciascun metodo assegna punteggi di aggregazione massimi. L'unica eccezione è rappresentata da FoldAmyloid, che non si limita a restituire un insieme di segmenti contraddistinti da un punteggio numerico, ma riporta esplicitamente, in termini assoluti, la predizione delle regioni amiloidogeniche ritenuta più plausibile, cosicché si è ritenuto più opportuno basare la valutazione su quest'ultima.
Ancora, i risultati sono presentati sia in forma grafica (Figura 3.7, pag. 64) sia in forma tabulare, sotto forma di punteggio di sovrapposizione SOV (Tabella 3.5). La Tabella 3.6, pag. 66, dà conto dei risultati di predizione in termini di numero di veri/falsi positivi/negativi, conteggiati residuo per residuo.

	SOV			
	PALMO	PASTA	BETASCAN	FoldAmyloid
β-amiloide	72.7	46.5	49.8	90.6
HET-s	38.2	40.1	44.2	18.0
FgHET-s	48.1	31.9	46.8	18.6
Amilina	57.4	45.5	44.4	58.2
α-sinucleina	78.2	79.9	82.7	26.4
Proteina τ	66.0	56.8	40.5	43.6
SOV complessivo	63.0	56.0	47.1	38.7
SOV medio	58.4	50.1	51.4	42.6

Tabella 3.5: Confronto tra i punteggi SOV ottenuti da PALMO, PASTA, BETASCAN, FoldAmyloid nella predizione di regioni amiloidogeniche sulle sei proteine del test set per $n = 7$. Per ciascuna proteina, è evidenziato in rosso il valore più elevato.

β-amiloide Tutti i metodi presi in esame segnalano l'alta amiloidogenicità di entrambe le regioni sperimentali. PALMO, in realtà, rileva erroneamente il segmento non aggregante che le separa, ma la tendenza alla sovrappredizione è ancora più accentuata nel caso di PASTA e di BETASCAN, con quest'ultimo, in particolare, che predice un segmento in una porzione della proteina, quella iniziale, del tutto estranea a fenomeni di aggregazione. Ciò vale a PALMO un vantaggio di oltre 20 punti SOV rispetto a questi ultimi. Un discorso a parte merita FoldAmyloid, che, in quest'unico caso, si rivela nettamente più preciso rispetto al nostro metodo, tanto da ottenere un punteggio SOV superiore a 90.

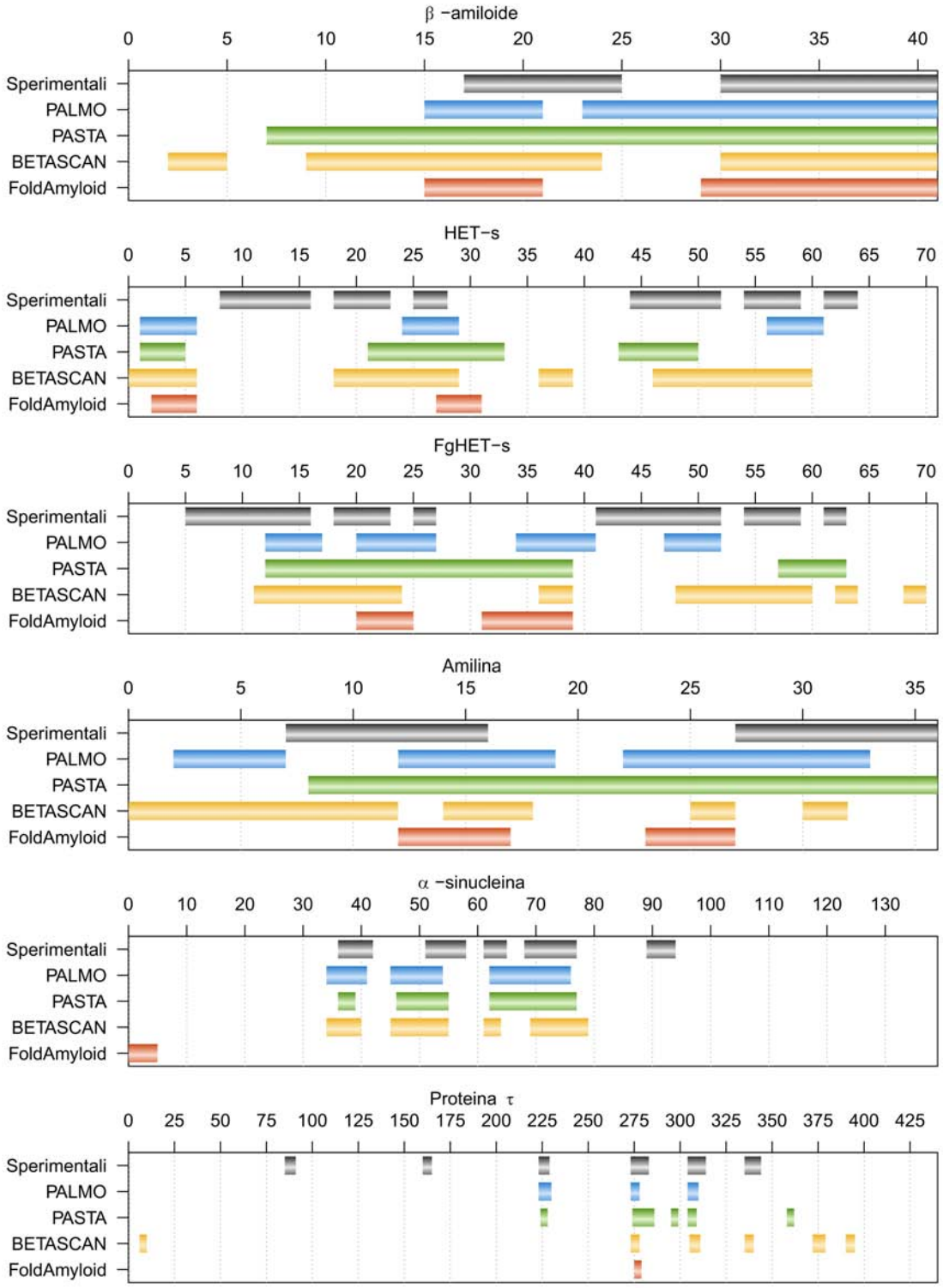


Figura 3.7: Confronto grafico tra i risultati di predizione delle regioni amiloidogeniche secondo Palmo, Pasta, BetaScan e FoldAmyloid.

HET-s Curiosamente, tutti i metodi incorrono in un falso positivo all'inizio della sequenza. PASTA e BETASCAN rilevano, rispettivamente, 3 e 4 regioni sperimentali, ma ancora al prezzo di una marcata sovrappredizione. PALMO, al contrario, prevede 2 soli segmenti sperimentali, ma si distingue per una maggiore specificità. I punteggi SOV risultanti, inferiori a 45 punti per tutti i predittori in esame, testimoniano la difficoltà nel prevedere una struttura complessa, quale quella di questo amiloide, per mezzo di tecniche basate esclusivamente sull'analisi della sequenza primaria.

FgHET-s Considerazioni analoghe valgono per questo amiloide, omologo del precedente e anch'esso contraddistinto da un struttura particolarmente articolata. Anche in questo caso, i predittori in esame sono accomunati da una falsa predizione positiva, in corrispondenza della porzione centrale della proteina, sperimentalmente non amiloidogenica. Al netto di questo errore, PALMO (4 regioni sperimentali rilevate su 8) e BETASCAN (5 su 8) si equivalgono in termini di precisione, ma il nostro metodo prevale ancora per specificità. Anche il punteggio SOV vede prevalere PALMO, benché anche in questo caso non arrivi ai 50 punti.

Amilina (IAPP) Se si prende a riferimento il modello illustrato in Figura 3.7, solo PALMO e PASTA sembrano cogliere un chiaro segnale di entrambe le regioni sperimentali, ma con una netta tendenza alla sovrappredizione. Se, invece, si considera il modello alternativo, il nostro metodo è l'unico, nonostante un falso positivo, a rilevare tutti e tre i filamento β, tenendo al contempo conto del segmento non amiloidogenico corrispondente alla prima curva; PASTA, invece, li comprende tutti in un unico segmento contiguo, mentre PASTA ne rileva solo uno su tre.

α-sinucleina Escludendo la pessima prestazione di FoldAmyloid, i restanti tre predittori si equivalgono su un buon livello di accuratezza: tutti rilevano 4 regioni amiloidogeniche su 5. Il vantaggio di BETASCAN in termini di SOV, per quanto molto ridotto, è giustificato dalla sua capacità di rilevare il breve segmento non aggregante che si frappone tra la terza e la quarta regione.

Proteina τ Sull'amiloide di gran lunga più esteso del training set, il nostro metodo è l'unico in grado di rilevare la maggior parte delle regioni sperimentali (3 su 5), compresi i già citati esapeptidi ^{274}`VQIINK`^{279} e ^{305}`VQIVYK`^{310}, senza incorrere in alcun falso positivo. Anche PASTA e BETASCAN individuano 3 segmenti sperimentali su 5, ma con precisione e specificità decisamente inferiori a causa dei falsi positivi in cui incappano entrambi; FoldAmyloid, invece, prevede solo il primo dei due esapeptidi. Ciò permette a PALMO di ottenere quasi 10 punti SOV di vantaggio sul secondo miglior predittore.

Nel complesso, una visione di insieme della Tabella 3.5, pag. 63, consente di notare in modo inequivocabile come, anche nel problema della rilevazione di regioni amiloidogeniche, sia stato raggiunto in pieno l'obiettivo con cui si era inaugurato lo sviluppo di PALMO,

ovvero migliorare le prestazioni predittive di PASTA, pur partendo dalle medesime assunzioni fondamentali: il nostro metodo, infatti, supera di gran lunga il predecessore in 4 casi su 6, mentre nei rimanenti è solo lievemente inferiore.

	PALMO				PASTA			
	TP	FP	TN	FN	TP	FP	TN	FN
β-amiloide	20	6	15	1	21	14	7	0
HET-s	9	9	25	29	14	12	22	24
FgHET-s	19	9	21	23	20	15	15	22
Amilina	13	13	4	7	19	10	7	1
α-sinucleina	23	10	94	13	23	7	97	13
Proteina τ	20	1	388	32	21	13	376	31
Totale	104	48	547	105	118	71	524	91
MCC [%]			46.7				46.1	
Precisione [%]			68.4				62.4	
Specificità [%]			91.9				88.1	

	BETASCAN				FoldAmyloid			
	TP	FP	TN	FN	TP	FP	TN	FN
β-amiloide	20	12	9	1	17	3	18	4
HET-s	23	15	19	15	2	8	26	36
FgHET-s	25	12	18	17	5	10	20	37
Amilina	13	11	6	7	6	5	12	14
α-sinucleina	23	10	94	13	0	6	98	36
Proteina τ	19	19	370	33	5	0	389	47
Totale	123	79	516	86	35	32	563	174
MCC [%]			46.1				18.0	
Precisione [%]			60.9				52.2	
Specificità [%]			86.7				94.6	

Tabella 3.6: La tabella contiene, per ciascun amiloide del test set, i conteggi, residuo per residuo, dei veri/falsi positivi/negativi conseguiti dai metodi in esame. Per ogni predittore, inoltre, sono riportati i valori complessivi del coefficiente di correlazione di Matthews (MCC), della precisione e della specificità.

Allargando lo sguardo agli altri predittori in esame, PALMO si distingue per il **miglior livello medio di prestazioni** sugli amiloidi del test set: tanto la media aritmetica dei punteggi SOV, quanto il SOV complessivo (che, di fatto, pesa i singoli punteggi in base

all'estensione delle sequenze proteiche) mostrano ben 7 punti di vantaggio sui secondi migliori.
Se si analizzano i risultati di predizione residuo per residuo, mostrati in Tabella 3.6, pag. 66, PALMO si segnala per un **numero di falsi positivi particolarmente ridotto**.[10] Ciò influisce positivamente sulla **precisione**, che, infatti, vede PALMO prevalere nettamente (+6% su PASTA), rendendo conto dell'ottima affidabilità di predizione del nostro metodo. In aggiunta, l'alto numero di veri negativi rispetto all'ammontare totale dei residui non amiloidogenici determina un'elevata **specificità**.[11]
In altri termini, PALMO si distingue dai metodi concorrenti per due aspetti:

1. è in grado di identificare buona parte dei residui non aggreganti (è **specifico**);
2. se riconosce un residuo come aggregante, con alta probabilità questo è effettivamente amiloidogenico (è **preciso**).

Non deve stupire, dunque, che anche una misura riassuntiva, quale il coefficiente di Matthews, premi PALMO rispetto ai predittori esaminati, attribuendo al nostro metodo la più alta correlazione tra l'esito della predizione e le evidenze sperimentali.

[10] Solo FoldAmyloid totalizza un numero inferiore di falsi positivi, ma la scarsa qualità complessiva di predizione conseguita da questo metodo rende tale cifra poco significativa.

[11] Ancora, FoldAmyloid ottiene un valore di specificità maggiore, ma questo non è altro che un'ovvia conseguenza del ridotto numero di predizioni positive, di gran lunga inferiore rispetto agli altri predittori.

Capitolo 4

Conclusioni

In anni recenti, l'aggregazione proteica è stata oggetto di un profondo interesse da parte della comunità medica, cresciuto di pari passo con l'emergere di prove scientifiche del coinvolgimento di questo fenomeno in una gran varietà di malattie neurodegenerative, quali il tristemente noto morbo di Alzheimer. Grandi sforzi sperimentali sono stati profusi nel tentativo di comprenderne le cause profonde e i meccanismi di innesco, con la consapevolezza che solo la conoscenza di questi aspetti possa aprire la strada alla cura delle patologie ad essa connesse.
In quest'ottica, può rivelarsi cruciale il supporto di adeguati strumenti software, in grado di affiancare e coadiuvare, se non sostituire, le pratiche sperimentali, dispendiose sia in termini di risorse economiche sia di tempo, nell'attività di ricerca. Compito della comunità bioinformatica è ideare e mettere a disposizione dei ricercatori tali strumenti software.
Palmo (*Protein Aggregation Likelihood and Mutation Optimization*), frutto del lungo lavoro di sviluppo di cui la presente tesi di laurea rappresenta il resoconto finale, si inserisce nell'affollato filone dei metodi computazionali che, avendo a disposizione la sola sequenza di amminoacidi caratteristica di una struttura proteica, ambiscono a prevederne la propensione ad innescare fenomeni di amiloidogenesi. Due sono, in particolare, i problemi cui questi approcci algoritmici tentano di dare risposta: il riconoscimento, da un insieme di brevi frammenti peptidici, dei responsabili del processo di aggregazione e, in secondo luogo, il rilevamento di regioni amiloidogeniche in sequenze proteiche complete.
Il processo di progettazione e realizzazione del nuovo metodo, costruito sulle solide fondamenta del già valido predittore Pasta (adottato, non a caso, da una tra le prime case farmaceutiche a livello mondiale), si è posto il traguardo di superarne le capacità predittive attraverso l'adozione di tecniche algoritmiche più raffinate ed efficienti, nonché di un complesso modulo di apprendimento automatico.
I risultati ottenuti dimostrano come gli obiettivi siano stati raggiunti: in entrambi i problemi di interesse, Palmo non solo esibisce un sostanziale miglioramento in accuratezza rispetto a Pasta, ma raggiunge prestazioni di rilievo anche nel panorama dei predittori di aggregazione oggi disponibili.
D'altra parte, i margini di miglioramento sono ancora ampi, specialmente nell'individuazione di regioni aggreganti in strutture proteiche complete. In quest'ambito, lo sfruttamento

di informazioni sulla struttura secondaria delle proteine, ricavabili con elevata affidabilità a partire dalla sequenza amminoacidica grazie ad appositi software, dovrebbe garantire una predizione più accurata.
Un'ulteriore percorso di sviluppo, del tutto innovativo, potrebbe riguardare l'ampliamento di Palmo con un modulo per l'analisi mutazionale (cui, non a caso, fa riferimento l'ultima parte dell'acronimo). Tale modulo si porrebbe l'obiettivo di prevedere quali mutazioni della sequenza peptidica (cioè, inserzioni o cancellazioni di singoli amminoacidi) siano in grado di provocarne variazioni del potenziale amiloidogenico. Si tratta di un progetto non privo di ambizione né di difficoltà, per il quale, tuttavia, gli eventuali sforzi futuri, ne siamo fiduciosi, daranno i loro frutti.

Appendice A

Dati supplementari

Peptide	Sequenza	Amiloidogenicità sperimentale
	τ-protein	
K19	PGGGKVQIVYKPV	+
K19d	PGGGKVYKPV	-
Mut1	PGGGKNAEVYKPV	-
Mut2	PGGGKVQIVEKPV	-
K19Chym	QTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVY	-
K19Chym1	KPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDF	-
K19Chym2	KDRVQSKIGSLDNITHVPGGGN	-
K19Gluc4	QTAPVPMPDLKNVKSKIGSTE	-
K19Gluc41	NLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE	+
K19Gluc42	VKSE	-
K19Gluc43	KLDFKDRVQSKIGSLDNITHVPGGGN	-
K19Gluc78	QTAPVPMPD	-
K19Gluc781	LKNVKSKIGSTE	-
K19Gluc782	NLKHQPGGGKVQIVYKEVD	+
K19Gluc783	LSKVTSKCGSLGNIHHKPGGGQVE	-
K19Gluc784	VKSEKLDFKDRVQSKIGSLDNITHVPGGGN	-
PHF8	GKVQIVYK	+
PHF6	VQIVYK	+
V313-K321	VDLSKVTSK	-
V318-G335	VTSKCGSLGNIHHKPGGG	-
V335-E342	GQVEVSKE	-
	Amyloid beta Aβ peptide (1-40)	
Whole	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV	+
HABP1	VPHQKLVFFAEDVGS	+
HABP2	VHPQKLVFFAEDVGS	+
HABP3	VHHPKLVFFAEDVGS	+
HABP4	VHHQPLVFFAEDVGS	+
HABP5	KKPVFFAED	-
HABP6	KKLPFFAED	-
HABP7	KKLVPFAED	-
HABP8	VHHQKLVPFAEDVGS	-

Prosegue...

HABP9	KKLVFPAED	–
HABP10	KKLVFFPED	+
HABP11	VHHQEKLVFFAPDVGS	–
HABP12	VHHQEKLVFFAEPVGS	+
HABP13	VHHQEKLVFFAEDPGS	+
HABP14	VHHQEKLVFFAEDVPS	+
HABP15	KKLVFFAED	+
HABP16	VHHQKLVFFAEDVGS	+
AB1	KLVFF	–
AB2	QKLVFFA	–
AB3	HQKLVFFAE	–
AB4	HHQKLVFFAED	+
AB5	VHHQKLVFFAEDV	+
AB6	EVHHQKLVFFAEDVG	+
AB7	YEVHHQKLVFFAEDVGS	+
AB8	GYEVHHQKLVFFAEDVGSN	+
AB9	SGYEVHHQKLVFFAEDVGSNK	+
AB10	DSGYEVHHQKLVFFAEDVGSNKG	+
AB11	HDSGYEVHHQKLVFFAEDVGSNKGA	+
	α-synuclein	
NAC1-18	EQVTNVGGAVVTGVTAVA	+
NAC1-18s	TVNGVGEVTATAVQGVAV	+
NAC3-18	VTNVGGAVVTGVTAVA	+
NAC1-13	EQVTNVGGAVVTG	+
NAC6-14	VGGAVVTGV	+
	Acyl phosphatase	
1-17	STAQSLKSVDYEVFGRV	–
18-33	QGVSFRMYTEDEARKI	–
34-53	GVVGWVKNTSKGTVTGQVQG	+
54-68	PEDKVNSMKSWLSKV	–
69-85	GSPSSRIDRTNFSNEKT	–
86-98	ISKLEYSNFSVRY	+
	β2-microglobulin	
A	IQRTPKIQVYSRHPAE	–
B	NGKSNFLNCYVSG	–
C	FHPSDIEVDLLK	–
D	NGERIEKVEHSDLSFSKD	–
E	DWSFYLLYYTEFT	+
E1	DWSFYLLYYTEFTPTGKDEYA	+
F	PTGKDEYACRVNHVT	–
G	LSQPKIVKWDRDM	–
	434 Cro repressor	
1Cro	MQTLSERLKKRRIALKY	–
2Cro	YKMTQTELATKAGVK	–
3Cro	YKQQSIQLIEAGVTKR	–
4Cro	TKRPRFLYEIAMALNSD	+
5Cro	AMALNCDPVWLQYGTKRGKA	–

Prosegue...

Sperm whale myoglobin		
A-Helix	VLSEGEWQLVLHVWAKVEA	+
AB-Domain	EGEWQLVLHVWAKVEADVAGHGQDILIRLFK	+
B-Helix	DVAGHGQDILIRLFKS	+
BC-Turn	KSHPET	-
CCD-Domain	HPETLEKFDRFKHLK	-
D-Helix	TEAEMKA	-
E-Helix	SEDLKKHGVTVLTALGAILK	-
EF-Turn	KKGHHEAE	-
F-Helix	ELKPLAQSHA	-
FG-Turn	ATKHKIP	-
Myohemerithrin		
N-terminal	GWEIPEPYVWDESFRVFY	-
C-terminal	GTDFKYKGKL	-
A helix	YEQLDEEHKKIFKGIFDCIRD	-
AB loop	RDNSA	-
B helix	SAPNLATLVKVTTNHFTHEEAMMD	+
BC loop	DAAKYSEV	-
C helix	EVVPHKKMHKDFLEKIGGL	+
CD loop	GLSAPVD	-
D helix	AKNVDYCKEWLVNHIK	-
D helix	AKNVDYCKEWLVNHIK	-
French bean plastocyanin		
Pc-1	LEVLLGSG	-
Pc-2	LEVLLGSGDGSLVFV	+
Pc-2a	SGDGSL	-
Pc-3	SLVFVPSEFS	-
Pc-4	SEFSV	-
Pc-5	SEFSVPSGEK	-
Pc-6	KIVFKNNA	-
Pc-6a	GEKIVFKNNAGFPHNVVFDE	+
Pc-7	KIVFKNNAGFPH	-
Pc-8	KNNAGFPHNV	-
Pc-9	PHNVVFDEDDEIP	-
Pc-10	IPAGVDAVKISM	+
Pc-10a	EIPAGV	-
Pc-10b	DAVKIS	-
Pc-11	MPEEELL	-
Pc-12	MPEEELLNAPGETYVVTL	+
Pc-13	ELLNAPGETY	-
Pc-13a	NAPGETY	-
Pc-13b	APGET	-
Pc-14	GETYVVTL	+
Pc-14a	ETYVVT	-
Pc-15	VTLDTKGTY	-
Pc-16	GTYSFYT	+
Pc-16a	TYSFYC	-
Pc-17	YTSPHQGAGMV	-
Pc-18	MVGKVTVN	-

Prosegue...

Pc-19	GTVSFVTSPHQGAGMVGKVTVN	+
Bovine pancreatic trypsin inhibitor (BPTI)		
P1-15	RPDFSLEPPYTGPSK	-
P29-44	LSQTFVYGGSRAKRNN	+
P13-21	PSKARIIRY	-
P41-51	KRNNFKSAEDS	-
P16-28	ARIIRYFYNAKAG	-
P24-32	NAKAGLSQT	-
N-terminal domain of ribosomal protein L9		
Beta 1	MKVIFLKDVKG	+
Beta 2	KGKKGEIKNVAD	-
Alpha 1	GYANNFLFKQG	+
Beta 3	LAIEATPA	-
Alpha 2	TPANLKALEAQKQKEQR	-
Glutathione S transeferase P domain II (Glutex)		
Alpha 4	DQKEAALVDMVNDGVEDLRCKYATLIYT	-
Alpha 5	YEAGKEKYVKELPEHLKPFETLLSQ	-
Alpha 6	QISFADYNLLDLLRIHQVLN	+
Alpha 7	PLLSAYVARLSA	-
Alpha 8	PKIKAFLA	-
Spectrin SH3		
M 2	AYVKKLDSGTGKELVLAL	-
M 4	YDYQEKSPREVTMKKGD	-
M 8	DILTLLNSTNKDWWKVEVND	+
M C	GGKDWWKVGG	-
M 6	DWWKVEVNDRQGFVPA	+
M 68	DILTLLNSTNKDWWKVEVNDRQGFVPA	+
M 681	DILTLLNSTNKDWWKVEVNDRQGFVPA	-
Ada-2h		
H1 Wt	VPSNEEQIKNLLQLEAQEHLQY	-
H1 Mt	VPSNEEQIKKLLELEAKKHLQY	-
H2 WT	FVNVQAVKVFLESQGIAY	+
H2 Mt	FVNVEAVKAFLEAHGIAY	+
Ara		
Ara1	AVGKSNLLSRYARNEFSA	-
Ara2	RFRAVTSAYYRGAVG	-
Ara3	TRRTTFESVGRWLDELKIHSD	-
Ara4	AVSVEEGKALAEEEGLF	-
Ara5	STNVKTAFEMVILDIYNNV	+
Com-A		
ComA1	DHPAVMEGTKTILETDSNLS	-
ComA2	EPSEQFIKQHDFSSY	-
ComA3	VNGMELSKQILQENPH	-
ComA4	EVEDYFEEAIRAGLH	-
ComA5	TESKEKITQYIYHVLNGEIL	+

Prosegue...

Che Y		
Che Y1	DFSTMRRIVRNLLKELGYN	-
Che Y2	EDGVDALNKLQAGGY	-
Che Y3	MDGLELLKTIRADSAY	-
Che Y4	AKKENIIAAAQAGASGY	+
Che Y5	PFTAATLEEKLNKIFEKLGMY	+
Flavodoxin		
FXN1	GTGNTEKMAELIAKGIIESGKDY	-
FXN3	EESEFEPFIEEISTKISY	-
FXN4	GDGKWMRDFEQRMNGYGSV	-
FXN5	EPDEAEQDSIEFGKKIANIY	-
P21-ras		
P21A	GVGKSALTIQLIQNHFVY	+
P21B	EYSAMRDQYMRTGEG	-
P21C	INNTKSFEDIHQYREQIKRVKDS	-
P21D	ARTVESRQAQDLARSYGIP	-
P21E	RQGVEDAFYTLVREIRQHK	+
PL B1 protein		
PL B1 95-114 pH 4.1	VTIKANLIFANGFTQTAEFKG	+
PL B1 114-138 pH 2.4	KGTFEKATSEAYAYADTLKKDNGEY	+
PL B1 136-155D pH 6.1	GEYTVDVADKGYTLNIKFAGD	+
Protein G		
ProteinG2-19	TYKLINGKTLKGETTTEA	-
ProteinG21-40	GDAATAEKVFKQYANDNGVD	-
ProteinG41-56	GEWTYDDATKTFTVTE	+

Tabella A.1: Elenco completo dei peptidi ricavati dal data set *Tango*. Si tratta di 177 frammenti peptidici, estratti da 21 proteine, la cui propensione all'aggregazione è stata accertata attraverso i seguenti metodi sperimentali: dicroismo circolare (CD, *circular dichroism*), risonanza magnetica nucleare (NMR, *nuclear magnetic resonance*), spettroscopia infrarossa in trasformata di Fourier (FTIR, *Fourier transform infrared spectroscopy*), fluorescenza della tioflavina T (ThT) o della tioflavina S (ThS), cromatografia liquida ad alta prestazione in fase inversa (RF-HPLC, *reverse-phase high-performance liquid chromatography*). In totale, 65 peptidi sono risultati amiloidogenici, i restanti 112 non amiloidogenici.

	Sequenza	**Punteggio**
1	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV	355.25
2	EQVTNVGGAVVTGVTAVA	327.61
3	VTNVGGAVVTGVTAVA	326.61
4	STNVKTAFEMVILDIYNNV	269.68

Legenda: Vero positivo | Falso positivo | Vero negativo | Falso negativo

	Sequenza	**Punteggio**
5	SEDLKKHGVTVLTALGAILK	250.48
6	VGGAVVTGV	226.58
7	TESKEKITQYIYHVLNGEIL	221.37
8	EGEWQLVLHVWAKVEADVAGHGQDILIRLFK	208.02
9	VLSEGEWQLVLHVWAKVEA	208.02
10	DWSFYLLYYTEFTPTGKDEYA	205.46
11	DWSFYLLYYTEFT	195.53
12	MKVIFLKDVKG	186.82
13	QTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVY	182.96
14	FVNVQAVKVFLESQGIAY	182.62
15	MPEEELLNAPGETYVVTL	182.22
16	GETYVVTL	182.22
17	EQVTNVGGAVVTG	180.40
18	NLKHQPGGGKVQIVYKEVD	177.59
19	VTIKANLIFANGFTQTAEFKG	173.01
20	LEVLLGSGDGSLVFV	169.80
21	GVGKSALTIQLIQNHFVY	168.37
22	PGGGKVQIVYKPV	166.05
23	NLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE	166.05
24	TVNGVGEVTATAVQGVAV	155.99
25	KKLVFFAED	153.13
26	VHHPKLVFFAEDVGS	152.55
27	SLVFVPSEFS	152.39
28	GKVQIVYK	150.91
29	SAPNLATLVKVTTNHFTHEEAMMD	150.80
30	VHHQKLVFFAEDVGS	146.61
31	VHHQKLVFFAEDV	146.61
32	EVHHQKLVFFAEDVG	146.61
33	YEVHHQKLVFFAEDVGS	146.61
34	GYEVHHQKLVFFAEDVGSN	146.61
35	SGYEVHHQKLVFFAEDVGSNK	146.61
36	DSGYEVHHQKLVFFAEDVGSNKG	146.61
37	HDSGYEVHHQKLVFFAEDVGSNKGA	146.61
38	VPHQKLVFFAEDVGS	146.61
39	HHQKLVFFAED	146.61
40	VHPQKLVFFAEDVGS	146.35
41	HQKLVFFAE	145.48
42	QKLVFFA	140.82
43	GVVGWVKNTSKGTVTGQVQG	139.35
44	VHHQEKLVFFAPDVGS	137.50
45	VHHQEKLVFFAEDPGS	136.59
46	VHHQEKLVFFAEDVPS	136.59
47	VHHQEKLVFFAEPVGS	135.46
48	DQKEAALVDMVNDGVEDLRCKYATLIYT	135.40
49	ETYVVT	133.63
50	VHHQPLVFFAEDVGS	133.22
51	DVAGHGQDILIRLFKS	129.99
52	GEKIVFKNNAGFPHNVVFDE	129.74
53	PHNVVFDEDDEIP	127.86
54	LSQTFVYGGSRAKRNN	127.01
55	KLVFF	126.23

Legenda: Vero positivo | Falso positivo | Vero negativo | Falso negativo

	Sequenza	Punteggio
56	GTVSFVTSPHQGAGMVGKVTVN	125.77
57	KKLVFFPED	122.85
58	VQIVYK	118.37
59	ARIIRYFYNAKAG	116.75
60	MVGKVTVN	112.93
61	DILTLLNSTNKDWWKVEVNDRQGFVPA	111.17
62	DILTLLNSTNKDWWKVEVND	111.17
63	DILTLLNSTNKDWWKVEVNDRQGFVPA	111.17
64	YEQLDEEHKKIFKGIFDCIRD	110.61
65	GEWTYDDATKTFTVTE	109.98
66	QISFADYNLLDLLRIHQVLN	106.90
67	LEVLLGSG	93.98
68	RQGVEDAFYTLVREIRQHK	93.59
69	GTYSFYT	93.02
70	PGGGKVQIVEKPV	92.87
71	GEYTVDVADKGYTLNIKFAGD	90.79
72	AYVKKLDSGTGKELVLAL	88.83
73	GWEIPEPYVWDESFRVFY	84.83
74	KIVFKNNA	80.20
75	KIVFKNNAGFPH	80.20
76	TYSFYC	76.74
77	KDRVQSKIGSLDNITHVPGGGN	76.61
78	KLDFKDRVQSKIGSLDNITHVPGGGN	76.61
79	VKSEKLDFKDRVQSKIGSLDNITHVPGGGN	76.61
80	NGKSNFLNCYVSG	75.52
81	KKPVFFAED	71.33
82	IQRTPKIQVYSRHPAE	70.67
83	AKKENIIAAAQAGASGY	70.00
84	IPAGVDAVKISM	69.87
85	PSKARIIRY	69.02
86	DWWKVEVNDRQGFVPA	67.89
87	GTGNTEKMAELIAKGIIESGKDY	66.18
88	PTGKDEYACRVNHVT	65.27
89	AKNVDYCKEWLVNHIK	65.13
90	AKNVDYCKEWLVNHIK	65.13
91	ISKLEYSNFSVRY	63.56
92	STAQSLKSVDYEVFGRV	56.03
93	GQVEVSKE	54.97
94	KPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDF	54.47
95	EPDEAEQDSIEFGKKIANIY	54.31
96	QGVSFRMYTEDEARKI	54.28
97	TKRPRFLYEIAMALNSD	51.42
98	FVNVEAVKAFLEAHGIAY	51.05
99	EESEFEPFIEEISTKISY	48.14
100	AMALNCDPVWLQYGTKRGKA	47.90
101	PGGGKVYKPV	43.60
102	DFSTMRRIVRNLLKELGYN	43.57
103	LSQPKIVKWDRDM	43.49
104	KKLVFPAED	42.74
105	YKQQSIQLIEAGVTKR	40.65
106	DHPAVMEGTKTILETDSNLS	40.29

Legenda: Vero positivo | Falso positivo | Vero negativo | Falso negativo

	Sequenza	Punteggio
107	TYKLINGKTLKGETTTEA	32.64
108	PLLSAYVARLSA	32.61
109	PKIKAFLA	32.23
110	LAIEATPA	29.73
111	RFRAVTSAYYRGAVG	29.21
112	DAVKIS	28.98
113	MQTLSERLKKRRIALKY	24.73
114	VTLDTKGTY	24.05
115	TRRTTFESVGRWLDELKIHSD	22.42
116	YEAGKEKYVKELPEHLKPFETLLSQ	20.68
117	GDGKWMRDFEQRMNGYGSV	20.29
118	MDGLELLKTIRADSAY	19.52
119	VTSKCGSLGNIHHKPGGG	18.90
120	GDAATAEKVFKQYANDNGVD	18.71
121	SEFSV	18.48
122	SEFSVPSGEK	18.48
123	PGGGKNAEVYKPV	18.47
124	FHPSDIEVDLLK	18.10
125	AVSVEEGKALAEEEGLF	16.85
126	LSKVTSKCGSLGNIHHKPGGGQVE	16.58
127	EVVPHKKMHKDFLEKIGGL	16.02
128	GGKDWWKVGG	16.01
129	KGKKGEIKNVAD	14.50
130	VDLSKVTSK	13.76
131	YDYQEKSPREVTMKKGD	13.60
132	KKLVPFAED	13.13
133	GYANNFLFKQG	12.56
134	PFTAATLEEKLNKIFEKLGMY	9.89
135	EVEDYFEEAIRAGLH	9.85
136	PEDKVNSMKSWLSKV	9.85
137	VHHQKLVPFAEDVGS	9.34
138	VPSNEEQIKKLLELEAKKHLQY	9.24
139	INNTKSFEDIHQYREQIKRVKDS	9.11
140	QTAPVPMPDLKNVKSKIGSTE	8.97
141	LKNVKSKIGSTE	8.97
142	ARTVESRQAQDLARSYGIP	7.09
143	VPSNEEQIKNLLQLEAQEHLQY	6.82
144	YKMTQTELATKAGVK	5.38
145	EPSEQFIKQHDFSSY	4.07
146	YTSPHQGAGMV	4.04
147	EDGVDALNKLQAGGY	2.09
148	DAAKYSEV	–
149	NGERIEKVEHSDLSFSKD	–
150	KGTFEKATSEAYAYADTLKKDNGEY	–
151	HPETLEKFDRFKHLK	–
152	EYSAMRDQYMRTGEG	–
153	AVGKSNLLSRYARNEFSA	–
154	GTDFKYKGKL	–
155	GSPSSRIDRTNFSNEKT	–
156	VNGMELSKQILQENPH	–
157	ELLNAPGETY	–

Legenda: Vero positivo | Falso positivo | Vero negativo | Falso negativo

	Sequenza	Punteggio
158	KNNAGFPHNV	–
159	KKGHHEAE	–
160	KKLPFFAED	–
161	ATKHKIP	–
162	TEAEMKA	–
163	KRNNFKSAEDS	–
164	RPDFSLEPPYTGPSK	–
165	TPANLKALEAQKQKEQR	–
166	NAPGETY	–
167	VKSE	–
168	ELKPLAQSHA	–
169	NAKAGLSQT	–
170	GLSAPVD	–
171	EIPAGV	–
172	MPEEELL	–
173	QTAPVPMPD	–
174	KSHPET	–
175	RDNSA	–
176	SGDGSL	–
177	APGET	–

Tabella A.2: Risultati completi della classificazione tramite PALMO (penalità $K = 0$) del data set *Tango*. I peptidi sono riportati in ordine di punteggio di aggregazione non crescente e classificati secondo il punteggio-soglia che produce un tasso di falsi positivi $tpr = 5\%$ (ovvero, 139.35).

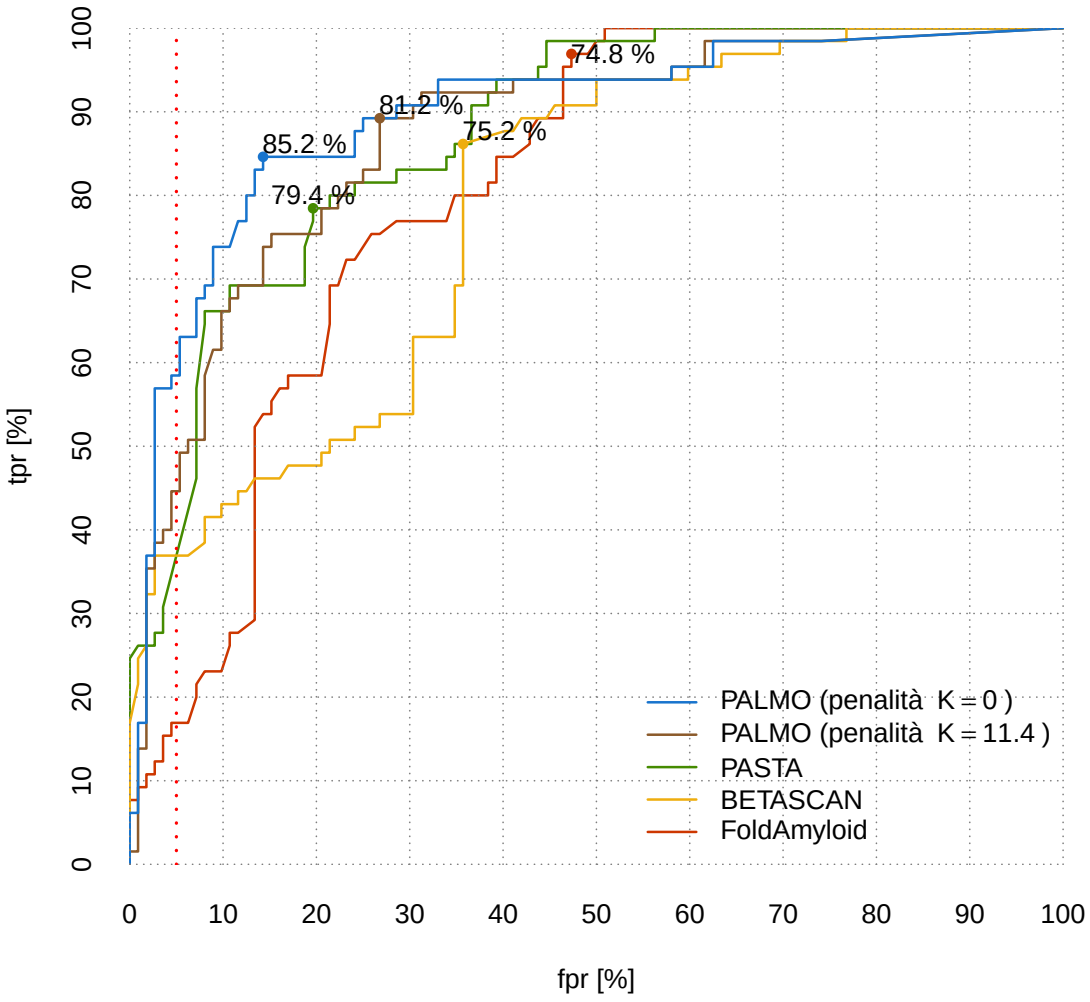


Figura A.1: Curve ROC relative alla classificazione del data set *Tango* attraverso PALMO, sia senza penalità ($K = 0$) sia con penalità unitaria $K = 11.4$ per coppie di segmenti di lunghezza superiore a 6, e gli altri classificatori in esame (PASTA, BETASCAN, FoldAmyloid). La linea rossa punteggiata evidenzia i punti di ascissa $fpr = 5\%$. I valori percentuali riportati nel grafico indicano la massima accuratezza bilanciata ottenuta da ciascun classificatore.

β-amiloide ($A\beta_{1-42}$)

```
0                   10                  20                  30                  40
D A E F R H D S G Y E V H H Q K L V F F A E D V G S N K G A I I G L M V G G V V I A
```

HET-s

```
0                   10                  20                  30                  40
K I D A I V G R N S A K D I R T E E R A R V Q L G N V V T A A A L H G G I R I S D Q T T N S V E T V ...
50                  60                  70
V G K G E S R V L I G N E Y G G K G F W D N
```

FgHET-s

```
0                   10                  20                  30                  40
K L N M I E G H N S A E F V N L E G S A K F L V G N V F S E K F L Q R D V L L N D R T K N S M R T V ...
50                  60                  70
S A T N Q S R L Q V G N V Y G G R G I W E D
```

Amilina (IAPP)

```
0                   10                  20                  30
K C N T A T C A T Q R L A N F L V H S S N N F G A I L S S T N V G S N T Y
```

α-sinucleina

```
0                   10                  20                  30                  40
M D V F M K G L S K A K E G V V A A A E K T K Q G V A E A A G K T K E G V L Y V G S K T K E G V V H ...
50                  60                  70                  80                  90
G V A T V A E K T K E Q V T N V G G A V V T G V T A V A Q K T V E G A G S I A A A T G F V K K D Q L ...
100                 110                 120                 130
G K N E E G A P Q E G I L E D M P V D P D N E A Y E M P S E E G Y Q D Y E P E A
```

Proteina τ

```
0                   10                  20                  30                  40
M A E P R Q E F E V M E D H A G T Y G L G D R K D Q G G Y T M H Q D Q E G D T D A G L K E S P L Q T ...
50                  60                  70                  80                  90
P T E D G S E E P G S E T S D A K S T P T A E D V T A P L V D E G A P G K Q A A A Q P H T E I P E G ...
100                 110                 120                 130                 140
T T A E E A G I G D T P S L E D E A A G H V T Q A R M V S K S K D G T G S D D K K A K G A D G K T K ...
150                 160                 170                 180                 190
I A T P R G A A P P G Q K G Q A N A T R I P A K T P P A P K T P P S S G E P P K S G D R S G Y S S P ...
200                 210                 220                 230                 240
G S P G T P G S R S R T P S L P T P P T R E P K K V A V V R T P P K S P S S A K S R L Q T A P V P M ...
250                 260                 270                 280                 290
P D L K N V K S K I G S T E N L K H Q P G G G K V Q I I N K K L D L S N V Q S K C G S K D N I K H V ...
300                 310                 320                 330                 340
P G G G S V Q I V Y K P V D L S K V T S K C G S L G N I H H K P G G G Q V E V K S E K L D F K D R V ...
350                 360                 370                 380                 390
Q S K I G S L D N I T H V P G G G N K K I E T H K L T F R E N A K A K T D H G A E I V Y K S P V V S ...
400                 410                 420                 430                 440
G D T S P R H L S N V S S T G S I D M V D S P Q L A T L A D E V S A S L A K Q G L
```

Tabella A.3: Sequenze amminoacidiche delle sei proteine amiloidi utilizzate per la verifica dell'accuratezza nell'individuazione di regioni amiloidogeniche (Sezione 2.3, pag. 21) [OWL$^+$11]. I segmenti evidenziati in rosso rappresentano le regioni che studi sperimentali hanno riconosciuto come propense all'aggregazione.

Bibliografia

[AK12] Ahmed, Abdullah e Andrey V Kajava: *Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence.* FEBS letters, 2012.

[Anf73] Anfinsen, C.B.: *Principles that govern the folding of protein chains.* Science, 181(4096):223–230, 1973.

[BJMC+09] Bryan Jr, Allen W, Matthew Menke, Lenore J Cowen, Susan L Lindquist e Bonnie Berger: *BETASCAN: probable β-amyloids identified by pairwise probabilistic analysis.* PLoS computational biology, 5(3):e1000333, 2009.

[BSTC02] Berg, J.M., L. Stryer, J.L. Tymoczko e N.D. Clarke: *Biochemistry.* W. H. Freeman, 2002, ISBN 9780716740148.

[CSdGA+07] Conchillo-Sole, Oscar, Natalia de Groot, Francesc Aviles, Josep Vendrell, Xavier Daura e Salvador Ventura: *AGGRESCAN: a server for the prediction and evaluation of hot spots of aggregation in polypeptides.* BMC Bioinformatics, 8(1):65, 2007, ISSN 1471-2105.

[CWT+99] Chiti, Fabrizio, Paul Webster, Niccolò Taddei, Anne Clark, Massimo Stefani, Giampietro Ramponi e Christopher M Dobson: *Designing conditions for in vitro formation of amyloid protofilaments and fibrils.* Proceedings of the National Academy of Sciences, 96(7):3590–3594, 1999.

[dlPS04] Paz, Manuela López de la e Luis Serrano: *Sequence determinants of amyloid fibril formation.* Proceedings of the National Academy of Sciences, 101(1):87–92, 2004.

[Edd04] Eddy, Sean R: *What is dynamic programming?* Nature biotechnology, 22(7):909–910, 2004.

[EG68] Eanes, ED e GG Glenner: *X-ray diffraction studies on amyloid filaments.* Journal of Histochemistry & Cytochemistry, 16(11):673–677, 1968.

[FERSS04] Fernandez-Escamilla, Ana Maria, Frederic Rousseau, Joost Schymkowitz e Luis Serrano: *Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins.* Nature biotechnology, 22(10):1302–1306, 2004.

[FFD01] Fändrich, Marcus, Matthew A Fletcher e Christopher M Dobson: *Amyloid fibrils from muscle myoglobin.* Nature, 410(6825):165–166, 2001.

[FP09] Fei, Li e Sarah Perrett: *Disulfide bond formation significantly accelerates the assembly of Ure2p fibrils because of the proximity of a potential amyloid stretch.* Journal of Biological Chemistry, 284(17):11134–11141, 2009.

[GLG10] Garbuzynskiy, Sergiy O, Michail Yu Lobanov e Oxana V Galzitskaya: *FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence.* Bioinformatics, 26(3):326–332, 2010.

[Hay94] Haykin, Simon: *Neural networks: a comprehensive foundation.* Prentice Hall PTR, 1994.

[HHB+05] Heise, Henrike, Wolfgang Hoyer, Stefan Becker, Ovidiu C Andronesi, Dietmar Riedel e Marc Baldus: *Molecular-level secondary structure, polymorphism, and dynamics of full-length α-synuclein fibrils studied by solid-state NMR.* Proceedings of the National Academy of Sciences of the United States of America, 102(44):15871–15876, 2005.

[HJH+02] Hammarström, Per, Xin Jiang, Amy R Hurshman, Evan T Powers e Jeffery W Kelly: *Sequence-dependent denaturation energetics: A major determinant in amyloid disease diversity.* Proceedings of the National Academy of Sciences, 99(suppl 4):16427–16432, 2002.

[HSW89] Hornik, Kurt, Maxwell Stinchcombe e Halbert White: *Multilayer feedforward networks are universal approximators.* Neural networks, 2(5):359–366, 1989.

[JñGO+99] Jimenez, Jose L, JI ñaki Guijarro, Elena Orlova, Jesús Zurdo, Christopher M Dobson, Margaret Sunde e Helen R Saibil: *Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing.* The EMBO journal, 18(4):815–821, 1999.

[Jon99] Jones, David T: *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of molecular biology, 292(2):195–202, 1999.

[JtBK+11] Joosten, Robbie P, Tim AH te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander e Gert Vriend: *A series of PDB related databases for everyday needs.* Nucleic acids research, 39(suppl 1):D411–D419, 2011.

[KAS05] Kajava, Andrey V, Ueli Aebi e Alasdair C Steven: *The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin.* Journal of molecular biology, 348(2):247–252, 2005.

[KS83] Kabsch, Wolfgang e Christian Sander: *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 22(12):2577–2637, 1983.

[LRA+05] Lührs, Thorsten, Christiane Ritter, Marc Adrian, Dominique Riek-Loher, Bernd Bohrmann, Heinz Döbeli, David Schubert e Roland Riek: *3D structure of Alzheimer's amyloid-β (1–42) fibrils.* Proceedings of the National Academy of Sciences of the United States of America, 102(48):17342–17347, 2005.

[LRSC01] Leiserson, Charles E, Ronald L Rivest, Clifford Stein e Thomas H Cormen: *Introduction to algorithms.* The MIT press, 2001.

[Lus71] Lusted, Lee B: *Signal detectability and medical decision-making.* Science, 171(3977):1217–1219, 1971.

[LYLT07] Luca, Sorin, Wai Ming Yau, Richard Leapman e Robert Tycko: *Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR.* Biochemistry, 46(47):13505–13522, 2007.

[MBK+09] Mukrasch, Marco D, Stefan Bibow, Jegannath Korukottu, Sadasivam Jeganathan, Jacek Biernat, Christian Griesinger, Eckhard Mandelkow e Markus Zweckstetter: *Structural polymorphism of 441-residue tau at single residue resolution.* PLoS biology, 7(2):e1000034, 2009.

[MSDK+10] Maurer-Stroh, Sebastian, Maja Debulpaep, Nico Kuemmerer, Manuela Lopez de la Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano *et al.*: *Exploring the sequence determinants of amyloid structure using position-specific scoring matrices.* Nature methods, 7(3):237–242, 2010.

[NE06] Nelson, Rebecca e David Eisenberg: *Recent atomic models of amyloid fibril structure.* Current opinion in structural biology, 16(2):260–265, 2006.

[NSB+05] Nelson, Rebecca, Michael R Sawaya, Melinda Balbirnie, Anders Ø Madsen, Christian Riekel, Robert Grothe e David Eisenberg: *Structure of the cross-β spine of amyloid-like fibrils.* Nature, 435(7043):773–778, 2005.

[NW70] Needleman, Saul B e Christian D Wunsch: *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of molecular biology, 48(3):443–453, 1970.

[ON08] Otzen, Daniel e Per Halkjær Nielsen: *We find them here, we find them there: functional bacterial amyloid.* Cellular and Molecular Life Sciences, 65(6):910–927, 2008.

[OWL+11] O'Donnell, Charles W, Jérôme Waldispühl, Mieszko Lis, Randal Halfmann, Srinivas Devadas, Susan Lindquist e Bonnie Berger: *A method for probing the mutational landscape of amyloid structure.* Bioinformatics, 27(13):i34–i42, 2011.

[PECS07] Pastor, M Teresa, Alexandra Esteras-Chopo e Luis Serrano: *Hacking the code of amyloid formation: the amyloid stretch hypothesis.* Prion, 1(1):9–14, 2007.

[PIB+02] Petkova, Aneta T, Yoshitaka Ishii, John J Balbach, Oleg N Antzutkin, Richard D Leapman, Frank Delaglio e Robert Tycko: *A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR.* Proceedings of the National Academy of Sciences, 99(26):16742–16747, 2002.

[RHW02] Rumelhart, David E, Geoffrey E Hinton e Ronald J Williams: *Learning representations by back-propagating errors.* Cognitive modeling, 1:213, 2002.

[Ros09] Ross, Sheldon M: *Introduction to probability and statistics for engineers and scientists.* Academic Press, 2009.

[RS93] Rost, Burkhard e Chris Sander: *Prediction of protein secondary structure at better than 70% accuracy.* Journal of molecular biology, 232(2):584–599, 1993.

[Ser00] Serpell, Louise C: *Alzheimer?s amyloid fibrils: structure and assembly.* Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1502(1):16–30, 2000.

[SW81] Smith, Temple F e Michael S Waterman: *Identification of common molecular subsequences.* Journal of molecular biology, 147(1):195–197, 1981.

[TCMS06] Trovato, Antonio, Fabrizio Chiti, Amos Maritan e Flavio Seno: *Insight into the structure of amyloid fibrils from the analysis of globular proteins.* PLoS computational biology, 2(12):e170, 2006.

[TE09] Teng, Poh K e David Eisenberg: *Short protein segments can drive a non-fibrillizing protein into the amyloid state.* Protein Engineering Design and Selection, 22(8):531–536, 2009.

[TSK+06] Thompson, Michael J, Stuart A Sievers, John Karanicolas, Magdalena I Ivanova, David Baker e David Eisenberg: *The 3D profile method for identifying fibril-forming segments of proteins.* Proceedings of the National Academy of Sciences of the United States of America, 103(11):4074–4078, 2006.

[TV08] Tartaglia, Gian Gaetano e Michele Vendruscolo: *The Zyggregator method for predicting protein aggregation propensities.* Chemical Society Reviews, 37(7):1395–1401, 2008.

[VBFB+00] Von Bergen, M, P Friedhoff, J Biernat, J Heberle, E M Mandelkow e E Mandelkow: *Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (306VQIVYK311) forming β structure.* Proceedings of the National Academy of Sciences, 97(10):5129–5134, 2000.

[VCL+08] Vilar, Marçal, Hui Ting Chou, Thorsten Lührs, Samir K Maji, Dominique Riek-Loher, Rene Verel, Gerard Manning, Henning Stahlberg e Roland Riek: *The fold of α-synuclein fibrils.* Proceedings of the National Academy of Sciences, 105(25):8637–8642, 2008.

[VV11] Voet, D. e J.G. Voet: *Biochemistry.* John Wiley & Sons, 2011, ISBN 9781118139929.

[VZN+04] Ventura, Salvador, Jesús Zurdo, Saravanakumar Narayanan, Matilde Parreño, Ramón Mangues, Bernd Reif, Fabrizio Chiti, Elisa Giannoni, Christopher M Dobson, Francesc X Aviles *et al.*: *Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case.* Proceedings of the National Academy of Sciences of the United States of America, 101(19):7258–7263, 2004.

[WBB+05] Westermark, Per, Merrill D Benson, Joel N Buxbaum, Alan S Cohen, Blas Frangione, Shu Ichi Ikeda, Colin L Masters, Giampaolo Merlini, Maria J Saraiva e Jean D Sipe: *Amyloid: Toward terminology clarification report from the nomenclature committee of the international society of amyloidosis.* Amyloid, 12(1):1–4, 2005.

[Wik13a] Wikipedia: *Alzheimer's disease — Wikipedia, The Free Encyclopedia*, 2013. `http://en.wikipedia.org/w/index.php?title=Alzheimer%27s_disease&oldid=575468828`, [Online; controllata il 3 ottobre 2013].

[Wik13b] Wikipedia: *Amminoacido — Wikipedia, L'enciclopedia libera*, 2013. `http://it.wikipedia.org/w/index.php?title=Amminoacido&oldid=61087738`, [Online; controllata il 29 settembre 2013].

[Wik13c] Wikipedia: *Beta sheet — Wikipedia, The Free Encyclopedia*, 2013. `http://en.wikipedia.org/w/index.php?title=Beta_sheet&oldid=560318633`, [Online; controllata il 2 ottobre 2013].

[Wik13d] Wikipedia: *Pearson product-moment correlation coefficient — Wikipedia, The Free Encyclopedia*, 2013. `http://en.wikipedia.org/w/index.php?title=Pearson_product-moment_correlation_coefficient&oldid=570726650`, [Online; controllata il 10 settembre 2013].

[Wik13e] Wikipedia: *Proteina — Wikipedia, L'enciclopedia libera*, 2013. `http://it.wikipedia.org/w/index.php?title=Proteine&oldid=61148267`, [Online; controllata il 30 settembre 2013].

[WLVM+08] Wasmer, Christian, Adam Lange, Hélène Van Melckebeke, Ansgar B Siemer, Roland Riek e Beat H Meier: *Amyloid fibrils of the HET-s (218–289) prion form a β solenoid with a triangular hydrophobic core.* Science, 319(5869):1523–1526, 2008.

[WWP+99] West, Michael W, Weixun Wang, Jennifer Patterson, Joseph D Mancias, James R Beasley e Michael H Hecht: *De novo amyloid proteins from designed combinatorial libraries.* Proceedings of the National Academy of Sciences, 96(20):11211–11216, 1999.

[WZS+10] Wasmer, Christian, Agnes Zimmer, Raimon Sabaté, Alice Soragni, Sven J Saupe, Christiane Ritter e Beat H Meier: *Structural similarity between the prion domain of HET-s and a homologue can explain amyloid cross-seeding in spite of limited sequence identity.* Journal of molecular biology, 402(2):311–325, 2010.

[ZVFR99] Zemla, Adam, Česlovas Venclovas, Krzysztof Fidelis e Burkhard Rost: *A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment.* Proteins: Structure, Function, and Bioinformatics, 34(2):220–223, 1999.