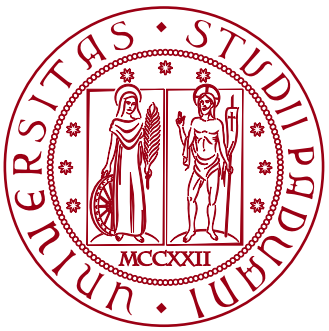


UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Leggi di scala Gaussiane nelle strutture di proteine

Relatore

Prof. Antonio Trovato

Laureando

Matteo Tessarolo

Anno Accademico 2018/2019

Abstract

Lo scopo di questa tesi è quello di analizzare un dataset di strutture di proteine globulari a singolo dominio alla luce della teoria delle catene polimeriche, così da dimostrare che l'interno di queste si comporta come un *polymer melt*.

L'elaborato si divide in 4 punti principali;

- Raffinazione, durante la quale si escludono le proteine "problematiche" che presentano lacune nei relativi files o il cui raggio giratore si discosta eccessivamente dall'andamento atteso delle proteine globulari;
- Analisi della distribuzione delle distanze end-to-end di frammenti proteici di diversa lunghezza e confronto con la distribuzione attesa, ovvero una Maxwell-Boltzmann, come previsto dalla teoria del Polymer Melt di Flory;
- Analisi della bontà dei fit di tali distribuzioni mediante calcolo del chi quadro al variare della lunghezza dei frammenti e di un parametro che varia la restrittività di selezione delle proteine idonee all'analisi.
- Considerazioni e riflessioni finali.

Il dataset a disposizione consiste in 21153 proteine globulari a singolo dominio estratte dal Protein Data Bank [6], un portale che ospita oltre 150000 strutture macromolecolari biologiche. L'analisi numerica è stato ad opera di un programma in C++ realizzato a tal scopo, mentre i plot invece sono stati prodotti con il software Matlab.

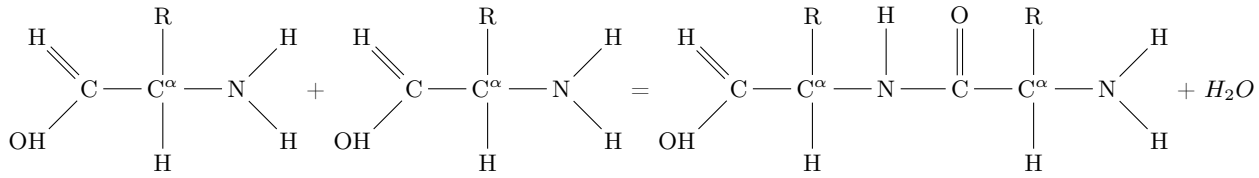
Indice

Capitolo 1

Introduzione

1.1 Proteine

Le proteine, i motori molecolari di ogni organismo vivente, sono macromolecole biologiche scoperte dal chimico tedesco Herman Emil Fischer agli innizi del 20° secolo [1]. Sono costituite da catene di amminoacidi, il cui ordine di successione, denominato struttura primaria, è determinato dall'informazione contenuta nel codice genetico. Gli amminoacidi sono 20 differenti molecole organiche che presentano 2 gruppi funzionali: quello amminico, detto N-terminus e quello carbossilico, detto C-terminus. Due gruppi funzionali C e N di amminoacidi consecutivi, mediante una reazione di condensazione formano un legame detto peptidico, che è la maglia di giunzione di ogni catena proteica, detta anche polipeptide. In seguito alla condensazione, gli amminoacidi cambiano il loro nome in residui, per via della perdita di una molecola di H_2O.



Le proteine, per quanto riguarda questo elaborato, sono polimeri. Come mostrato dal chimico britannico Frederick Sanger nel 1950, la loro struttura chimica è data da una spina dorsale, o *backbone*, chimicamente regolare e da una sequenza ben precisa di catene laterali, o *side chains R*. La distanza di legame virtuale è identificata dai C^{α}, i punti nodali degli amminoacidi, la cui distribuzione è particolarmente piccata attorno al valore di 3.8 Å, a causa della planarità del legame peptidico. Questa peculiarità permette di filtrare in maniera efficiente le proteine che, per via della complessità dei metodi di indagine, presentano delle lacune nella rappresentazione della catena.

Altra particolarità della catena polipeptidica è quella di assumere, in termini di angoli diedrali, certi orientamenti preferenziali permettendo la formazione delle cosiddette strutture secondarie presenti nelle proteine, quali le α-eliche e i β-foglietti. L'organizzazione spaziale di questi 2 tipi di strutture rappresenta la struttura terziaria delle proteine, seguita solo dalla struttura quaternaria che descrive la combinazione di più subunità proteiche a formare una proteina.

Fattore che gioca un ruolo fondamentale nella conformazione delle proteine è l'ambiente nel quale sono immerse; su questo è possibile fare una catalogazione:

- Proteine fibrose: formano grandi aggregati poveri di H_2O, sono unite tra loro da legami idrogeno e godono di grande regolarità;
- Proteine di membrana: risiedono nella membrana cellulare, ambiente internamente povero di H_2O ed esposto ad essa da un lato, inoltre sono proteine molto regolari e mantenute unite da legami idrogeno
- Proteine idrosolubili globulari: meno regolari, soprattutto quelle più piccole, e a differenza delle precedenti sono mantenute unite dalle auto-interazioni della catena con se stessa o al limite con alcuni cofattori

Queste ultime rappresentano l'oggetto dell'elaborato, e la trattazione teorica è completamente incentrata sulla teoria delle catene polimeriche di seguito esposta.

1.2 Polimeri

La teoria delle catene polimeriche è vasta e spazia tra fisica, matematica e chimica.

In questo elaborato se ne fa uso di una porzione limitata che include macroscopicamente i modelli di catena ideale, catena reale, teoria di Flory e polymer melt.

Il contenuto teorico necessario alla comprensione dell'analisi è descritto in dettaglio nel successivo capitolo 2, con particolare riferimento alle leggi di scala Gaussiane presenti sia nelle catene ideali che nei melt di polimeri.

1.3 Dataset

Il dataset a disposizione consiste di 21153 singoli domini di proteine globulari, estratti dal portale *Protein Data Bank* [6], lo stesso già adoperato in [8]. Un dominio e' una porzione di catena proteica in grado di raggiungere autonomamente la propria struttura terziaria. Ogni proteina è descritta da un file in cui ogni riga corrisponde ad un atomo e tra le colonne si trovano la specie atomica, l'amminoacido di appartenenza e le coordinate espresse in Å. Tale rappresentazione non è esente da problemi, data la complessità dei metodi di determinazione utilizzati, come cristallografia a raggi-X, laser a elettroni liberi Europeo "XFEL", risonanza magnetica nucleare "NMR" e microscopia elettronica 3D [5]. Tra questi problemi si trovano le lacune, ovvero atomi o interi amminoacidi mancanti, fenomeno da identificare accuratamente e isolare per evitare di invalidare la correttezza dell'analisi.

Per l'analisi svolta è stato sufficiente utilizzare solo le coordinate tridimensionali X, Y e Z del carbonio C^{α}, quindi tutte le altre informazioni contenute nei files sono sovrabbondanti. Per alleggerire l'analisi è stato quindi ricreato un dataset ridotto di soli 108 MB, a differenza dei 1,82 GB di spazio occupato su disco del dataset completo. Questo ha permesso un guadagno nel tempo di elaborazione di oltre 2 ordini di grandezza.

1.4 Sintesi dell'elaborazione dati

L'analisi condotta si articola sostanzialmente nelle seguenti fasi:

- Raffinazione dati, nella quale si rimuovono le proteine che contengono difformità, come lacune o caratteristiche dimensionali non idonee devianti dai prequisiti richiesti, ovvero di essere globulari.
- Analisi dell'andamento del raggio giratore al variare del grado di polimerizzazione N;
- Analisi della distribuzione delle distanze end-to-end, all'interno della proteina;

Tale analisi riprende per la maggior parte quella riportata in un articolo del 2005 redatto da Banavar, Hoang e Maritan [2].

Capitolo 2

Teoria dei polimeri

L'analisi effettuata fa uso di alcuni modelli tipici della teoria del polimeri[7], di cui si tratta in seguito.

2.1 Polimeri

Un polimero è un insieme di monomeri uniti consecutivamente da un legame chimico. Per monomero si intende una molecola relativamente semplice capace di unirsi ad altre simili, anche se non necessariamente uguali, a creare una catena. Il numero N di monomeri presenti in una molecola polimerica è detto grado di polimerizzazione.

Le caratteristiche dei legami chimici in gioco, solitamente covalenti, fanno si che la lunghezza e l'angolo formato tra legami consecutivi siano pressoché costanti, ovvero con fluttuazioni molto piccole che non influenzano la conformazione della catena.

Ad esempio il polietilene ha una distanza di legame di 1.54 Å con fluttuazioni tipicamente di ± 0.05 Å, mentre l'angolo tra i legami consecutivi è $\theta = 68°$, detto angolo tetraedrico.

I legami hanno quindi un grado di libertà associato all'angolo diedrale di rotazione. L'interazione tra monomeri consecutivi da luogo ad un potenziale di interazione lungo l'angolo giro permesso all'angolo radiale, e solitamente si possono identificare 2 minimi relativi:

- Uno assoluto, al quale è associato l'angolo θ-trans. Essendo di equilibrio stabile è anche più probabile e inoltre un polimero con tutti i legami θ-trans gode della massima elongazione possibile R_{max};
- Uno relativo, al quale è associato l'angolo θ-gauche. E' un punto di equilibrio metastabile, e avendo energia maggiore è meno probabile del primo.

I fattori che influenzano la conformazione di una catena polimerica generalmente hanno origine energetica o entropica dovuta all'interazione di monomeri con altri monomeri o con un solvente. Data la complessità introdotta dal grande numero di monomeri che compongono una catena polimerica, è molto difficile, se non impossibile, trattarli uno ad uno. Per questo motivo la teoria trattata ha un impianto di tipo statistico/probabilistico e fa uso di leggi di scala o approssimazioni di campo medio. C'è sicuramente la possibilità di raffinare le approssimazioni introdotte, ma questo esula dallo scopo dell'elaborato.

Si passano quindi in rassegna i modelli teorici impiegati.

2.2 Catene ideali

Il modello di catena ideale è il più semplice modello possibile di catena polimerica. In tale modello si considera costante la distanza di legame b, ma completamente libera la sua direzione, trascurando quindi completamente le interazioni attrattive o repulsive tra i monomeri. La forza di questo modello è sicuramente la semplicità, compensata però dal fatto che quasi nessuna catena reale assume questo comportamento, fatta eccezione per alcuni casi, uno dei quali è esposto in seguito.

Un risultato interessante ricavabile con questo modello è la distribuzione della distanza *end-to-end*, ovvero la lunghezza del vettore che congiunge il primo monomero della catena con l'ultimo.

Per ottenere questa distribuzione si parte da un semplice *random-walk* binario, i cui passi monodimensionali possono essere solo +1 o -1 a partire da 0; Il numero di possibili cammini che portano alla posizione x in N passi sono dati da $W(N,x) = \frac{N!}{(\frac{N+x}{2})!(\frac{N-x}{2})!}$, e quindi la probabilità di arrivare in x in N passi è data da $W(N,x)$ diviso il numero totale di cammini 2^N;

$$P(N,x) = \frac{W(N,x)}{2^N} = \frac{N!}{2^N \left(\frac{N+x}{2}\right)! \left(\frac{N-x}{2}\right)!}$$

Si calcola quindi il limite per N molto grande di questa distribuzione discreta, rendendo la variabile x continua;

$$\log P_{1D}(N,x) = \log N! - N\log 2 - \underbrace{\log\left(\frac{N+x}{2}\right)!}_{=\log\left(\frac{N}{2}\right)!+\sum_{k=1}^{x}\log\frac{N+k}{2}} - \underbrace{\log\left(\frac{N-x}{2}\right)!}_{\log\left(\frac{N}{2}\right)!-\sum_{k=1}^{x}\frac{N-k+1}{2}}$$

$$= \log \underbrace{N!}_{stirling\sim\sqrt{2\pi N}\left(\frac{N}{e}\right)^N} - N\log 2 - 2\log \underbrace{\left(\frac{N}{2}\right)!}_{stirling\sim\sqrt{2\pi\frac{N}{2}}\left(\frac{N}{2e}\right)^{\frac{N}{2}}} - \sum_{k=1}^{x}\underbrace{\log\frac{N+k}{N-k+1}}_{\log\frac{N+k}{N}\log\cdot\frac{N-k+1}{N}}$$

$$\sim \log\sqrt{\frac{2}{\pi N}} - \sum_{k=1}^{N}\left[\underbrace{\log\left(1+\frac{k}{N}\right)}_{\sim\frac{k}{N}} - \underbrace{\log\left(1-\frac{k-1}{N}\right)}_{\sim-\frac{k-1}{N}}\right]$$

$$\sim \log\sqrt{\frac{1}{2\pi N}} - \sum_{k=1}^{N}\frac{2k-1}{N} = \log\left[\sqrt{\frac{1}{2\pi N}}\exp\left(-\frac{x^2}{2N}\right)\right]$$

Questa distribuzione continua approssima la probabilità discreta di trovarsi in x dopo un numero N, molto grande, di passi unitari del *random walk* binario unidimensionale in questione.

Si nota quindi che $\langle x^2\rangle = \int_{-\infty}^{+\infty} x^2 P_{1D}(N,x)dx = N$, e questo permette di riscrivere la distribuzione in termini di $\langle x^2\rangle$ anziché di N;

$$P_{1D}(x) = \sqrt{\frac{1}{2\pi\langle x^2\rangle}}\exp\left(-\frac{x^2}{2\langle x^2\rangle}\right)$$

A questo punto è naturale essere interessati ad ottenere la distribuzione tridimensionale delle distanze end-to-end a partire da quella unidimensionale. Così facendo si ricava la distribuzione della distanza end-to-end delle catene ideali tridimensionali, dette anche *freely rotating chain*, ovvero catene polimeriche caratterizzate da passo costante e direzione isotropica e uniformemente distribuita. Per fare questo è necessario sapere che per una catena ideale è possibile ricavare che $\langle R^2\rangle = Nb^2$ dove R è il vettore end-to-end, N è il grado di polimerizzazione e b la lunghezza media del legame. Dato che per ciascuna componente $\langle R_x\rangle = \langle R_y\rangle = \langle R_z\rangle = 0$, ed inoltre $\langle R^2\rangle = \langle R_x^2\rangle + \langle R_y^2\rangle + \langle R_z^2\rangle = Nb^2$, e di conseguenza $\langle R_x^2\rangle = \langle R_y^2\rangle = \langle R_z^2\rangle = \frac{Nb^2}{3}$.

La distribuzione end-to-end tridimensionale si ottiene mediante il seguente procedimento elementare:

$$P_{3D}(N,\overrightarrow{R})dR_xdR_ydR_z = P_{1D}(N,R_x)dR_x\,P_{1D}(N,R_y)dR_y\,P_{1D}(N,R_z)dR_z = P_{1D}^3(N,R^2)dR_xdR_ydR_z$$
$$= \left(\frac{1}{2\pi\langle R^2\rangle}\right)^{3/2}\exp\left(-\frac{R^2}{2\langle R^2\rangle}\right) = \left(\frac{3}{2\pi Nb^2}\right)^{3/2}\exp\left(-\frac{3R^2}{2Nb^2}\right)$$

Si nota infine che la densità di probabilità va integrata rispetto a $4\pi R^2 dR$, quindi in definitiva assume la forma:

$$P_{3D}(N,R)4\pi R^2 dR = 4\pi R^2\left(\frac{3}{2\pi Nb^2}\right)^{3/2}\exp\left(-\frac{3R^2}{2Nb^2}\right)dR$$

Si ottiene quindi che data una catena ideale di N monomeri con lunghezza di legame b, la probabilità di di avere una distanza end-to-end $\overrightarrow{R}$ giacente nel guscio sferico compreso tra i raggi R e R+dR è data da una distribuzione di Maxwell-Boltzmann. Questo risultato è fondamentale per le successive analisi.

Le catene ideali rappresentano un modello versatile e potente, ma non essendo particolarmente adatte a descrivere una vasta gamma di polimeri, è possibile definire, come proposto dal chimico svizzero Werner Kuhn, una catena ideale equivalente. Data una certa catena, è possibile associare ad essa una catena ideale equivalente avente stessa distanza massima di elongazione R_{max} e stessa distanza end-to-end media $\langle R^2\rangle$. Questo strumento è di grande importanza in quanto queste 2 quantità per la catena ideale sono $R_{max,id} = Nb$ e $\langle R^2\rangle_{id} = Nb^2$, e in questo modo è possibile trattare una catena qualsiasi tramite il modello ideale, introducendo per contro degli errori sistematici dovuti alla semplificazione.

2.3 Catene reali

Il modello di catena reale, a differenza del *toy model* delle catene ideali, prende in considerazione le interazioni tra i monomeri che costituiscono la catena. Prendere in considerazione le interazioni tra i monomeri risulta essere proibitivo vista la loro grande numerosità, quindi la strategia per analizzare una catena reale sfrutta un'approssimazione di campo medio, il vero punto cardine di questo modello. Effettuare in questo contesto un'approssimazione di campo medio significa sostituire la catena con un gas di N monomeri interagenti confinati in uno spazio tridimensionale caratterizzato dalla dimensione caratteristica R, verosimilmente confrontabile con la distanza end-to-end del polimero. In quanto gas, ha senso quindi di definire il corrispettivo della densità, detto fattore ϕ^* di sovrapposizione, o *overlap*. Limitandosi la trattazione ad un polimero tridimensionale, tale fattore è dato dal rapporto tra il volume caratteristico b^3 sommato sul numero di monomeri N rispetto al volume totale del gas R^3, quindi $\phi^* = \frac{Nb^3}{R^3}$.

Il fattore di *overlap* per una catena ideale, caratterizzata da $R = b\sqrt{N}$, è dato da $\phi^* = \frac{Nb^3}{N^{3/2}b^3} = \frac{1}{\sqrt{N}}$, e quindi tende ad annullarsi al crescere del numero di monomeri. La catena ideale è infatti un oggetto frattale. Il fattore di *overlap* è importante per la stima di campo medio del numero di contatti tra i monomeri, dal quale deriva il contributo energetico di interazione. Tale numero è stimabile quindi moltiplicando il fattore ϕ^* per il numero N di monomeri che costituiscono la catena. Sempre per un modello di catena ideale, il numero di contatti stimato è numericamente confrontabile con $\sim N\phi^* = N\frac{1}{\sqrt{N}} = \sqrt{N}$, crescente con il numero di monomeri, a differenza del fattore ϕ^*. E' importante riportare però che questa è una sottostima rispetto ai dati ottenuti tramite strumenti più sofisticati, che esulano dallo scopo dell'elaborato, i quali mostrano che il numero di contatti varia con N anziché con $\sqrt{N}$. Il numero di contatti tra monomeri produce comportamenti diversi a seconda che l'interazione netta sia attrattiva, nulla o repulsiva. L'interazione presente tra i costituenti della catena è della forma tipica del potenziale $U(r)$ di Lennard-Jones, che presenta classicamente una buca di potenziale ad una certa distanza insieme ad un potenziale di repulsione molto simile ad una sfera rigida con distanza di repulsione d. A titolo di esempio, se ne riporta un grafico in figura 2.1.

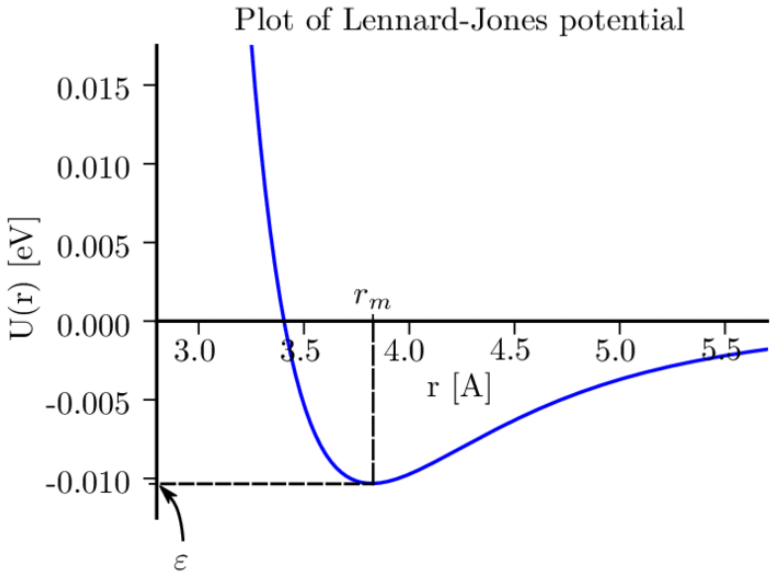


Figura 2.1: Potenziale di Lennard-Jones per gli atomi di gas Argon [3]

Dal potenziale $U(r)$ di interazione si ottiene la f-Function di Mayer $f(r) = \exp\left(-\frac{U(r)}{kT}\right) - 1$, mostrata in figura 2.2, che integrata fornisce l'opposto del volume escluso $v = -\int f(r)d^3r$. Quest'ultimo fornisce informazioni sulle condizioni di repulsività o attrattività dell'interazione in atto; se negativo, indica un'attrazione netta, se positivo una repulsione netta, altrimenti se nullo un'interazione netta nulla.

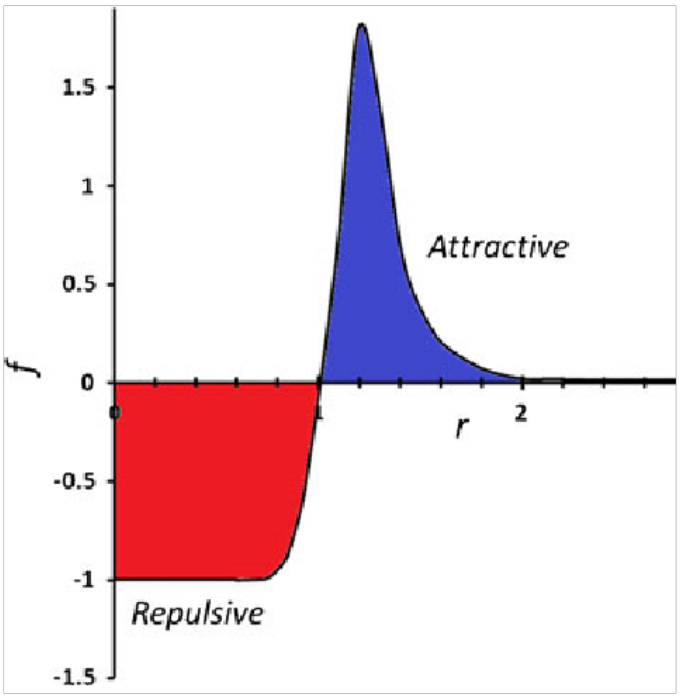


Figura 2.2: *f-function* di Mayer [4]

Fin'ora si è ritenuto valido l'assunto che i monomeri costituenti avessero una forma approssimativamente sferica. Questo spesso non si verifica, come nel caso eclatante delle proteine, e quindi la trattazione precedentemente fatta perde di validità. Per considere questa eventualità bisogna cambiare approccio, calcolando la densità di energia libera tramite un'espansione viriale;

$$\frac{F_{int}}{V} = \frac{kT}{2}\left(vc_n^2 + wc_n^3 + ...\right) \sim kT\left(v\frac{N^2}{R^6} + w\frac{N^3}{R^9} + ...\right)$$

Dove $c_n = \frac{N}{R^3}$ rappresenta la densità numerica di monomeri, mentre i coefficienti viriali sono rappresentati dal volume escluso v di cui sopra, e il coefficiente di interazione a tre corpi w. Per monomeri sferici e atermici valgono le seguenti stime di scala: $v \sim d^3$ e $w \sim d^6$. Per monomeri molto asimmetrici si ottengono invece stime molto più alte.

Il volume escluso è influenzato sensibilmente dalla tipologia di solvente nel quale un polimero è immerso;

- Solventi atermici: nel limite delle alte temperature, la *f-function* di Mayer ha come unico contributo la repulsione di sfera dura, quindi il volume escluso non dipende dalla temperatura, e può essere approssimato da $v \sim b^2 d$;
- Buoni solventi: nessuna distinzione energetica nell'interazione dei monomeri con loro stessi o tra monomeri e fluido. In questo caso è la buca di potenziale attrattiva a giocare un ruolo fondamentale, e questo riduce il volume escluso: $0 \leq v < b^2 d$;
- Solventi θ: ad una certa temperatura, detta $\theta - Temperature$, la buca di potenziale attrattiva annulla completamente il volume escluso, quindi $v = 0$. In questa condizione la conformazione del polimero è quasi quella di una catena ideale;
- Solventi poveri: a temperature più basse della $\theta - Temperature$ la buca di potenziale attrattiva diventa dominante e quindi il volume escluso diventa negativo: $-b^2 d \leq v < 0$. Questo è il caso trattato nell'elaborato, modello con il quale si affrontano delle proteine immerse in un fluido polare, come l'acqua.

2.4 Teoria di Flory

La teoria di Flory nasce per colmare l'esigenza di descrivere la conformazione di una catena reale di N polimeri gonfiata alla dimensione $R > R_0 = b\sqrt{N}$ considerando i fenomeni di attrazione e perdita entropica dovuta alla deformazione. Per fare questo, Paul John Flory, chimico statunitense attivo nel 20 ° secolo, ipotizzò che i monomeri della catena fossero distribuiti uniformemente e senza correlazioni tra di essi. Su questo punto si pone

una particolare attenzione perchè questo modello successivamente sarà applicato alle proteine, le quali presentando delle strutture secondarie non sono certamente immuni da correlazioni tra le posizioni degli amminoacidi, introducendo delle difficoltà nell'analisi.

Come primo aspetto si considera l'energia dovuta all'interazione. Per ricavare questo risultato si nota che, dato un monomero, la probabilità di un secondo monomero di essere entro il volume escluso del primo è fornita dal prodotto del volume escluso v e della densità numerica di monomeri che compongono la catena $c_n = \frac{N}{R^3}$. Moltiplicando questa probabilità per il fattore termico kT si ottiene il contributo energetico di interazione per un singolo monomero, che se moltiplicato per il numero N di monomeri fornisce il contributo energetico di interazione totale F_{INT}.

$$F_{INT} = kTvc_nN = kTv\frac{N^2}{R^3}$$

Il contributo energetico di origine entropica è invece stimato da Flory esser pari all'energia necessaria a stirare una catena fino alla distanza end-to-end R;

$$F_{ENT} = kT\frac{R^2}{Nb^2}$$

L'energia libera totale risulta essere quindi $F = F_{INT} + F_{ENT} = kT\left(v\frac{N^2}{R^3} + \frac{R^2}{Nb^2}\right)$. E' facile notare che esiste un R ottimo che minimizza l'energia libera, detto R_F di Flory. La caratteristica di essere un minimo per l'energia libera fa della R_F una distanza end-to-end più probabile rispetto alle altre distanze.

$$\left.\frac{\partial F}{\partial R}\right|_{R_F} = kT\left[-3v\frac{N^2}{R_F^4} + 2\frac{R_F}{Nb^2}\right] = 0$$

$$R_F = \left(\frac{3}{2}\right)^{1/5} v^{1/5}b^{2/5}N^{3/5}$$

Le dimensioni di una catena di questo tipo sono molto maggiori di quelle di una catena ideale avente lo stesso numero di monomeri, avente raggio giratore $b\sqrt{N}$.

$$\frac{R_F}{b\sqrt{N}} \sim \frac{v^{1/5}b^{2/5}N^{3/5}}{bN^{1/2}} = v^{1/5}b^{-2/5}N^{1/10}$$

Le predizioni di Flory sono in buon accordo con esperimenti e teorie più sofisticate grazie ad una fortuita serie di cancellazione di errori. La teoria di Flory infatti è teoria semplice e versatile, ma che introduce delle approssimazioni consistenti.

Teoria di Flory in un solvente povero.

La teoria di Flory può essere specializzata a polimeri immersi in un solvente povero, che porta quindi il volume escluso ad essere negativo. Si analizzano quindi i passaggi che portano alla formulazione della legge di scala che lega la dimensione caratteristica R del polimero al numero di monomeri N. Si riprende l'energia libera di cui sopra, $F = kT\left(\frac{R^2}{Nb^2} + \underbrace{v}_{<0}\frac{N^2}{R^3}\right)$ e si nota che la distanza R che la minimizza per $v < 0$ è $R = 0$. Questo è fisicamente inaccettabile e significa che mancano nel conteggio dei contributi energetici forti abbastanza da impedire il collasso del polimero su se stesso. Si intende quindi cercare quali termini possono evitare questo collasso, portando il minimo dell'energia in corrispondenza a dimensioni R non nulle.

Primo tra questi è il contributo dovuto all'entropia di confinamento, ovvero la penalità entropica che una certa catena deve pagare per poter essere confinata in una cavità sferica di raggio $R < b\sqrt{N}$, stimata come $F_{CONF} \sim kT\frac{Nb^2}{R^2}$. Con questo contributo aggiuntivo l'energia libera diventa $F \sim kT\left[\frac{R^2}{Nb^2} + v\frac{N^2}{R^3} + \frac{Nb^2}{R^2}\right]$. Anche considerando questo contributo energetico, il minimo dell'energia rimane in corrispondenza di $R = 0$, e questo significa che esso non è abbastanza forte da scongiurare il collasso in un punto della catena.

Un ulteriore contributo da considerare riprende i versi dell'espansione viriale precedentemente affrontata per la densità di energia libera, introducendo il coefficiente di repulsione a 3 corpi w.

$$\frac{F_{INT}}{R^3} \sim kT\left(vc_n^2 + wc_n^3 + ...\right) = kT\left(v\frac{N^2}{R^6} + w\frac{N^3}{R^9}\right)$$

Considerando questo componente l'energia libera totale risulta essere:

$$F \sim kT \left(\underbrace{\frac{R^2}{Nb^2} + \frac{Nb^2}{R^3}}_{=o(N^2)} + v\frac{N^2}{R^3} + w\frac{N^3}{R^6} \right) \sim kT \left(\underbrace{v}_{<0} \frac{N^2}{R^3} + \underbrace{w}_{>0} \frac{N^3}{R^6} \right)$$

Si ottiene finalmente un'espressione dell'energia libera i cui minimi si abbiano in corrispondenza di valori non nulli di R, e in particolare tale valore ottimo varia qualitativamente con $R_{gl} \sim \left(\frac{wN}{|v|}\right)^{1/3} \sim bd\left(\frac{N}{|v|}\right)^{1/3}$, in cui $w \sim (bd)^3$. Quest'ultima è la relazione di scala cercata facente da legante tra la dimensione caratteristica R del polimero e il grado di polimerizzazione N.

2.5 Polymer melt

Si definisce *polymer melt*, o "concentrato di polimeri", un sistema costituito da una miscela binaria di catene chimicamente identiche, di catene di lunghezza N_A in piccola concentrazione immerse in un bagno di catene di lunghezza N_B. Questo modello deve anche rispettare le seguenti ipotesi:

- Il contributo energetico relativo alla miscelazione dei 2 costituenti deve essere nullo.
- Il volume escluso contribuisce entropicamente solo in una piccola parte, essendo $v = \frac{b^3}{N_B}$ piccolo per un polymer melt.

Peculiarità centrale di questo modello è che non si distingue se una catena stia interagendo con se stessa o con le catene circostanti.

Si riporta infine un risultato fondamentale originariamente indicato da Flory; in un polymer melt di catene molto lunghe il volume escluso v è approssimativamente nullo, e quindi la conformazione dei polimeri in questione è simile a quella di catene ideali, con le corrispondenti distanze end-to-end che seguono la distribuzione di Maxwell-Boltzmann.

Per riassumere, le leggi di scala che caratterizzano la distanza end-to-end al variare del solvente nel quale la catena polimerica è immersa sono:

- Per un buon solvente, ovvero con $v > 0$: $R \sim N^{3/5}$;
- Per un solvente al punto theta, catene ideali e polymer melt, ovvero con $v = 0$: $R \sim N^{1/2}$;
- Per un cattivo solvente, ovvero con $v < 0$: $R \sim N^{1/3}$.

Capitolo 3

Analisi dati

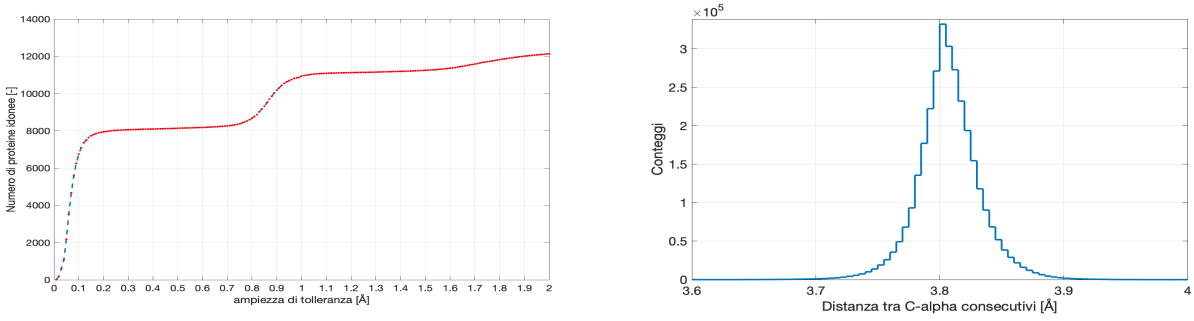
3.1 Filtraggio dei dati

In questa fase preliminare dell'analisi si vuole filtrare una porzione di proteine non adatte all'analisi sulla base di 2 criteri fondamentali, che sono:

- Presenza di lacune, analizzando le distanze tra tutti i C^{α} consecutivi e imponendo un limite di tolleranza oltre al quale scartare l'intera proteina problematica;
- Raggio giratore superiore ad un certo valore di tolleranza, fissato al variare della lunghezza della proteina, questo vincolo permette l'esclusione di proteine fibrose e di membrana dall'analisi, le quali hanno una forma più allungata e non sono quindi propriamente globulari.

Rimozione delle proteine che presentano lacune

Stabilito che la distanza tra due C^{α} consecutivi è intorno a 3.8 Å, è stato realizzato il grafico in figura 3.1, nel quale si riporta l'andamento del numero di proteine i cui C^{α} consecutivi hanno tutti distanze comprese nella finestra di tolleranza $[3.8\ \text{Å} - \delta\ ;\ 3.8\ \text{Å} + \delta]$ al variare dell'ampiezza δ. A lato invece è riportato un semplice istogramma di tutte le 3159265 distanze analizzate, con un particolare focus attorno a 3.8 Å. Inoltre da questo istogramma si calcola facilmente una stima della distanza del legame peptidico di 3.809 ± 0.277 Å, quindi ottimamente compatibile con il valore approssimato di 3.8 Å.



(a) Proteine idonee

(b) Istogramma delle distanze

Figura 3.1: Andamento del numero di proteine idonee al variare dell'ampiezza della finestra di tolleranza δ e istogramma delle distanze tra C^{α} consecutivi

Dal grafico è evidente che una finestra di ampiezza 0.4 Å è sufficientemente stretta per riuscire a scartare in modo efficiente tutte le distanze non compatibili con la distanza attesa di 3.8 Å. Si procede quindi ad effettuare una selezione delle proteine i cui C^{α} consecutivi distano tra loro nell'intervallo 3.8 ± 0.4 Å. Tale selezione riduce il numero di proteine idonee da 21153 a 7948, ovvero circa il 38%

Rimozione di proteine con limitazione al raggio giratore

In questa sezione si introduce il concetto di raggio giratore di una proteina, definito come:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\overrightarrow{R}_i - \overrightarrow{R}_{cm}\right]^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\overrightarrow{R}_i - \frac{1}{N}\sum_{i=1}^{N}\overrightarrow{R}_i\right]^2}$$

dove gli $\overrightarrow{R_i}$ rappresentano le coordinate dei punti C^α degli amminoacidi.

Questa grandezza quantifica la dispersione degli amminoacidi rispetto al centro di massa della proteina, analogamente allo scarto quadratico medio per un set di misure. Vista la grande densità delle proteine globulari in questione, è lecito aspettarsi che il raggio che descrive la dispersione della proteina vari con la radice cubica del numero di amminoacidi: $R_g = m \cdot N^{1/3}$. Si ottiene quindi la seconda condizione di scarto delle proteine non idonee all'analisi: se la coppia $(N\,;\,R_g)$ discosta eccessivamente dalla retta di interpolazione di tutte le coppie, si considera la proteina non idonea. Inoltre, un raggio giratore eccessivamente alto rispetto al numero di amminoacidi che la compongono, permette di dubitare che si tratti di una proteina globulare, in favore invece di una proteina fibrosa o di membrana, caratterizzate da strutture elongate, per le quali $R_g \sim N$.

Nel grafico in figura 3.2 si riportano le coppie $(N\,;\,R_g)$ in scala log-log, e da questa rappresentazione si può apprezzare come la maggior parte delle proteine che hanno superato il primo test si dispongano lungo una nuvola compresa tra 2 rette estreme di pendenza 1/3 e 1; la prima rappresenta l'andamento rispettato dalle proteine globulari, mentre la seconda è l'andamento delle proteine fibrose e di membrana. Si decide quindi di tagliare le coppie sovrastanti alla retta identificata da una pendenza 1/3 e passante per il punto (128,16). Questo taglio introduce un ulteriore scarto di proteine che ne salva 6546, ovvero l' 82% delle 7948 analizzate, il 31% delle 21153 totali. Il grafico delle coppie $(N\,;\,R_g)$ in scala log-log è riportato nel grafico in figura 3.3, e i parametri della retta di interpolazione $logR_g = a + b \cdot logN$ sono: $a = 0.991 \pm 0.007$, $b = 0.3414 \pm 0.001$. La pendenza, come atteso, è vicino al valore 1/3.

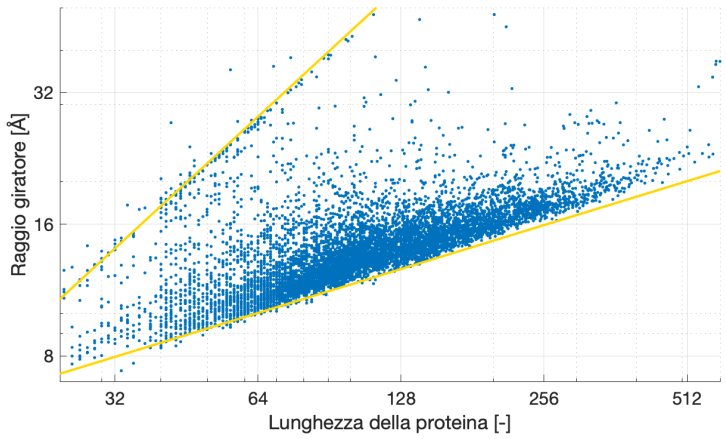


Figura 3.2: Andamento del raggio giratore al variare della lunghezza della proteina

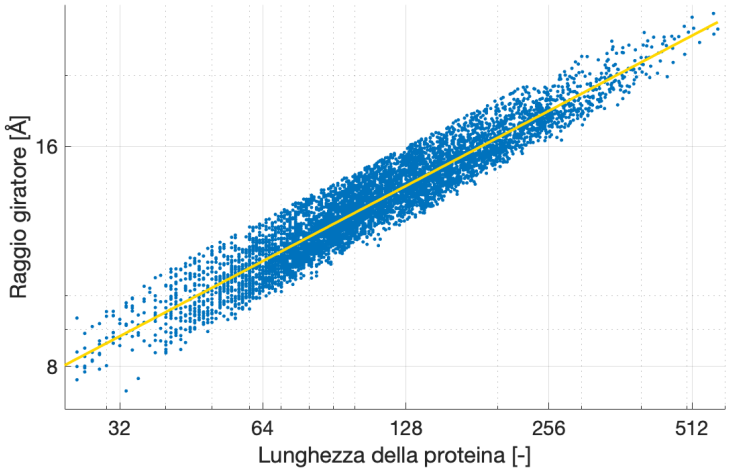


Figura 3.3: Dati raffinati e retta di interpolazione dei raggi giratori

Si considera quindi così conclusa la fase di filtraggio dei dati.

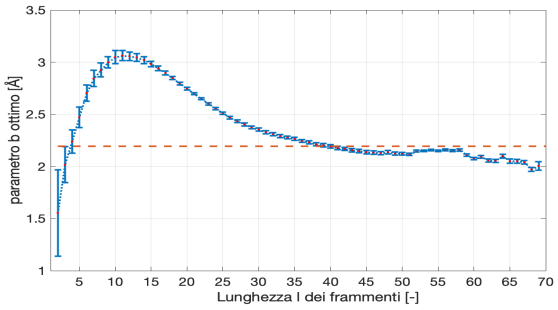
3.2 Analisi della distanza end-to-end su frammenti

Scopo di questa fase dell'analisi è quello di analizzare le distanze end-to-end di frammenti proteici di data lunghezza l, ovvero di sottoinsiemi connessi delle catene proteiche a disposizione composti da un certo numero l di amminoacidi. Detti quindi $\overrightarrow{R_i}$, $i = 1...N$ i raggi vettori degli N amminoacidi che compongono una proteina, si potranno identificare $N - l + 1$ frammenti di lunghezza l, ciascuno avente distanza end-to-end pari a $d_i = |\overrightarrow{R}_{i+l-1} - \overrightarrow{R}_i|$, $i = 1...N - l + 1$. Si ricorda che lo scopo di questa fase dell'analisi è quella di verificare che l'interno delle proteine globulari si comporta come un polymer melt, nel quale le catene di amminoacidi seguono localmente un modello a *free chain*, la cui distanza end-to-end divisa per $\sqrt{l}$, da qui in poi denominata come "distanza end-to-end riscalata", segue una distribuzione di Maxwell-Boltzmann con media $b\sqrt{\frac{8}{\pi}}$ e varianza $b^2\frac{3\pi-8}{\pi}$ determinate dal parametro b: $f_b(x) = \sqrt{\frac{2}{\pi}}\frac{x^2}{b^3}\exp\left[-\frac{1}{2}\left(\frac{x}{b}\right)^2\right]$. La lunghezza di legame risulta quindi $\sqrt{3}b$ in questo modello di catena ideale E' dunque chiaro che i frammenti analizzati devono essere interni alla proteina per poter essere fruibili allo scopo dell'analisi, in modo che frammenti diversi si possano pensare come appartenenti a diverse catene; da questa condizione nasce la disuguaglianza $N > a \cdot l^{3/2}$, dove N è la lunghezza della proteina, a è un parametro che regola la rigidezza della selezione e l è la lunghezza dei frammenti analizzati. Se si ricorda che $R_g \sim N^{1/3}$, se ne conviene che, a meno di un fattore moltiplicativo, deve essere $l^{1/2} < R_g$. Questa condizione stabilisce come la distanza end-to-end dei frammenti (proporzonale a $l^{1/2}$) debba essere minore del raggio giratore della proteina. Se la proteina e' pensata come una sfera di raggio pari al raggio giratore, questo equivale proprio a chiedere che il frammento di l residui sia interno alla proteina e che quindi non "senta" la presenza della superficie del globulo.

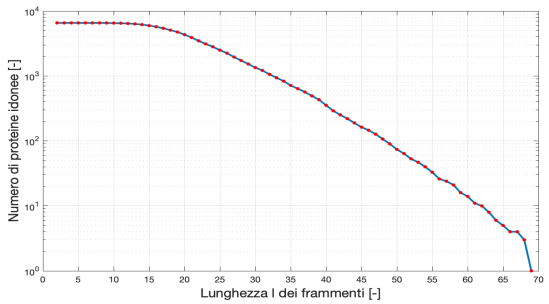
Analisi delle distanze end-to-end di frammenti con a=1

Durante questa prima fase dell'analisi si sceglie a=1. Si calcola quindi al variare della lunghezza dei frammenti analizzati, il parametro b della distribuzione di Maxwell-Boltzmann che meglio interpola l'istogramma della distanza end-to-end riscalata tramite la formula $b_{opt} = \sqrt{\frac{\langle x^2\rangle}{3}}$.

Questa stima di b si ottiene massimizzando la verosimiglianza. Si riporta quindi in figura 3.4 (a) l'andamento del parametro b al variare della lunghezza dei frammenti analizzati, in 3.4 (b) invece l'andamento del numero di proteine idonee all'analisi, al variare del numero di residui l dei frammenti analizzati.

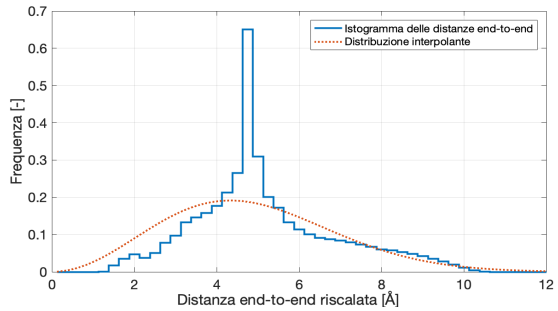


(a) Andamento di b al variare di l per a=1

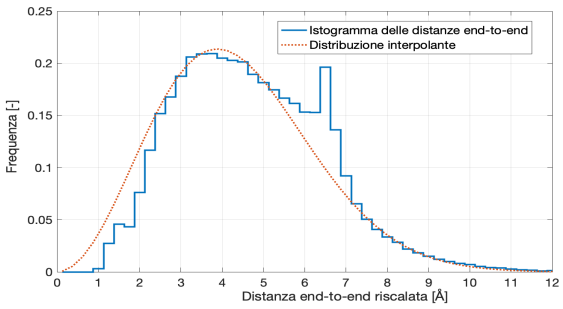


(b) Proteine idonee alla condizione $N > a \cdot l^{3/2}$ al variare di l con a=1

Figura 3.4

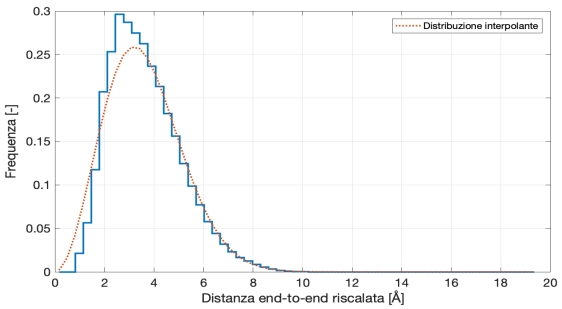


(a) Interpolazione per l=11

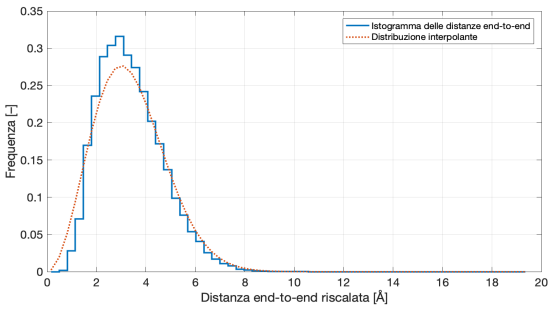


(b) Interpolazione per l=20

Figura 3.5: Interpolazioni a titolo di esempio delle distanze end-to-end riscalate.



(a) Interpolazione per l=35



(b) Interpolazione per l=50

Figura 3.6: Interpolazioni a titolo di esempio delle distanze end-to-end riscalate.

Dai grafici in figura 3.4, 3.5 e 3.6 si notano alcuni aspetti interessanti;

- la stima del parametro b presenta un massimo attorno al valore l=11; questo non è un caso, in quanto è proprio la lunghezza media delle α-eliche; molti di questi frammenti sono quindi elongati, causando un aumento della lunghezza di legame ($b\sqrt{3}$) stimate per mezzo del modello di catena ideale;

- per frammenti lunghi, la curva tende a stabilizzarsi attorno al valore 2.2 Å$\sim \frac{3.8}{\sqrt{3}}$ Å; e' interessante osservare che questa stima, ottenuta nell'ambito del modello di catena ideale, e' consistente con la lunghezza di legame "virtuale" fra C_α consecutivi, pari a 3.8 Å;

- per a=1, attorno a l=70 si esauriscono le proteine idonee all'analisi;

- l'istogramma delle distanze end-to-end riscalate presenta uno scostamento dalla distribuzione di Maxwell-Boltzmann che si ripete su tutti gli altri valori di l, consistente in uno scarto negativo per distanze inferiori alla moda, positivo per distanze superiori. Come si vedrà in seguito, questo effetto è mitigato da un aumento del parametro a, mentre inasprito da un suo decremento.

Da una prima analisi le distanze end-to-end riscalate sembrano adattarsi correttamente a distribuzioni di Maxwell-Boltzmann, perlomeno per l sufficientemente grandi (si confrontino fig. 3.5 e 3.6), suggerendo la correttezza dell'ipotesi dell'applicabilità del regime di Flory per le proteine globulari.

Si attende però la successiva fase dell'analisi, che quantifica la bontà del fit al variare del parametro a, quindi con diversi valori di rigidezza nella selezione delle proteine idonee.

3.3 Analisi della bontà dei fit al variare del parametro a

Si analizzano i casi con parametro $a = 2/3$ e $a = 3/2$, ovvero un caso inferiore all'unità e uno superiore.

Si nota immediatamente che riducendo il parametro a si riduce la selettività della condizione $N > al^{3/2}$, arricchendo quindi la statistica ma inquinando gli istogrammi con frammenti che potrebbero fuoriuscire dall'interno della proteina o distribuirvisi sulla superficie, uscendo quindi dalla categoria di frammenti immersi in un *polymer melt*.

Se il parametro a invece viene incrementato, la selettività si fa più severa impoverendo la statistica, ma assicurandosi con maggiore confidenza che i frammenti appartengano al *polymer melt*.

L'obiettivo di questa sezione è di quantificare come queste variazioni del parametro a influiscano nella bontà dei fit della distribuzione di Maxwell-Boltzmann per gli istogrammi della distanza end-to-end riscalata, e a questo scopo si utilizzano 2 strumenti:

- Calcolo della variabile chi quadro al variare di l e del parametro a; l'andamento è mostrato dal grafico in figura 3.7;

- Analisi visiva dei grafici di livello degli scarti tra istogramma della distanza end-to-end riscalata e relativa Maxwell-Boltzmann interpolante in figure 3.8, 3.9 e 3.10, dai quali, alla luce di quanto detto nella precedente sezione ci si aspetta osservare "valli" e "creste" andare via via a levigarsi all'aumentare di a.

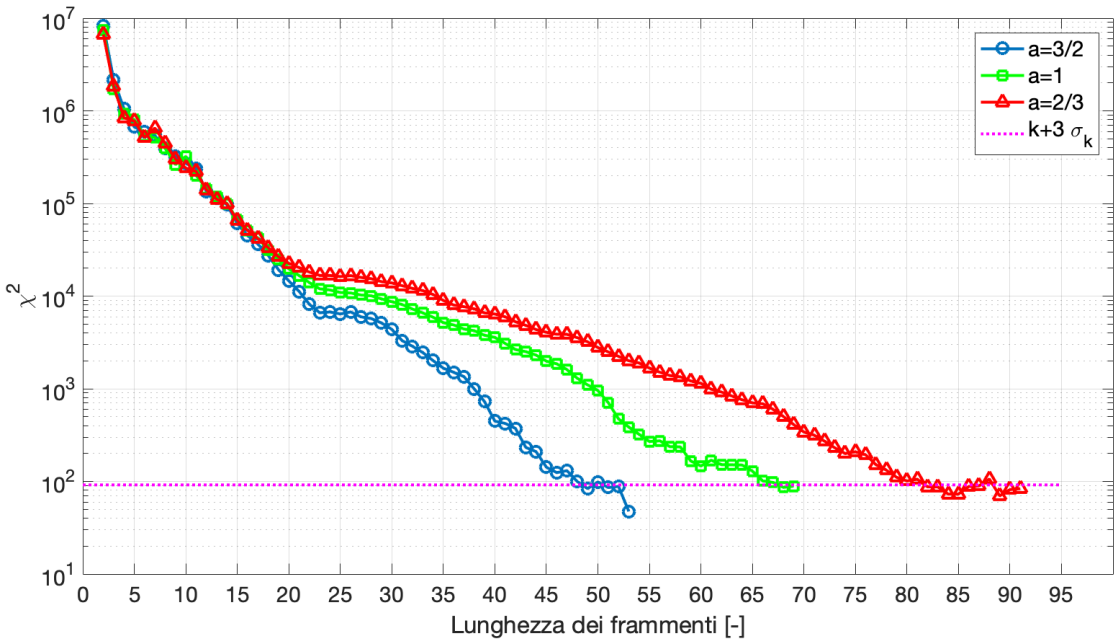


Figura 3.7: Andamento della variabile χ^2 al variare di l per diversi valori del parametro a; il segmento orizzontale punteggiato rappresenta la soglia massima del chi quadro $k + 3\sqrt{2k}$ con k=59 gradi di libertà.

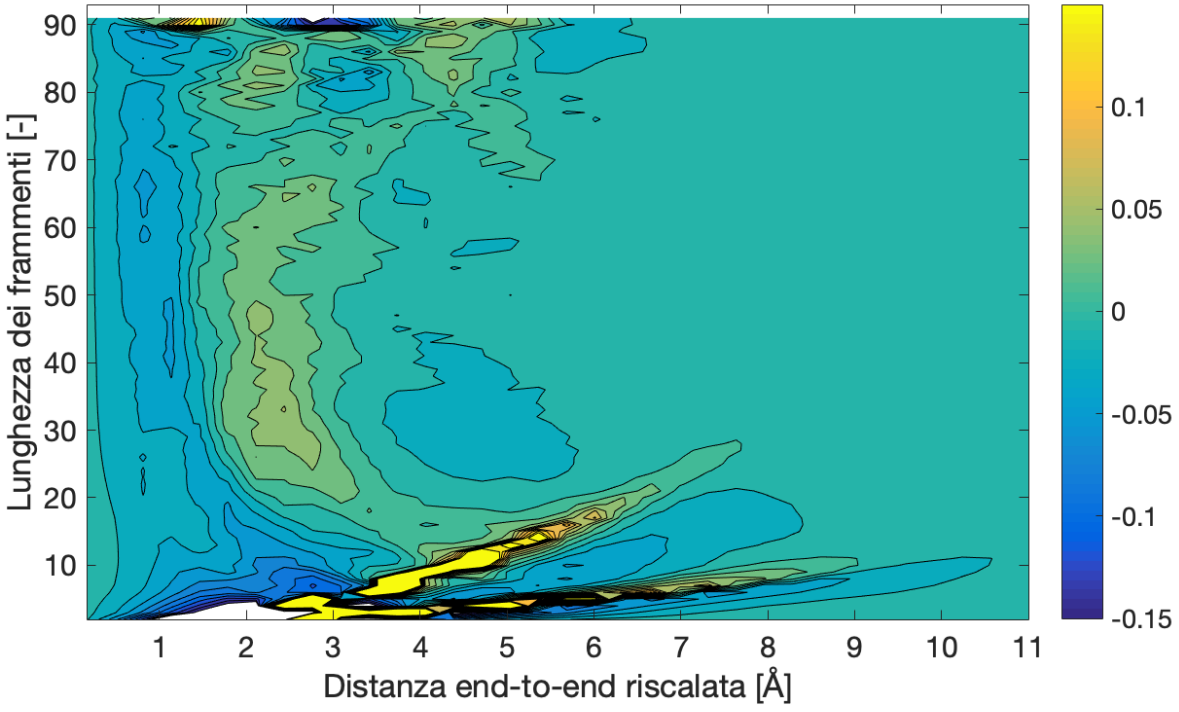


Figura 3.8: Superficie degli scarti per a=2/3

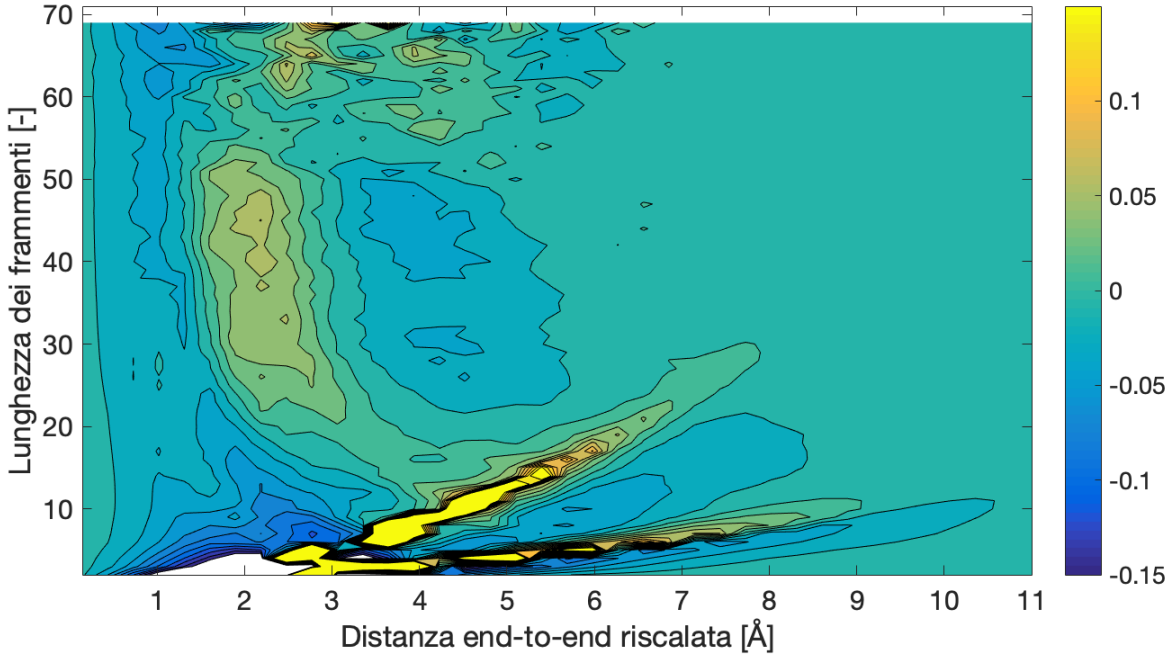


Figura 3.9: Superficie degli scarti per a=1

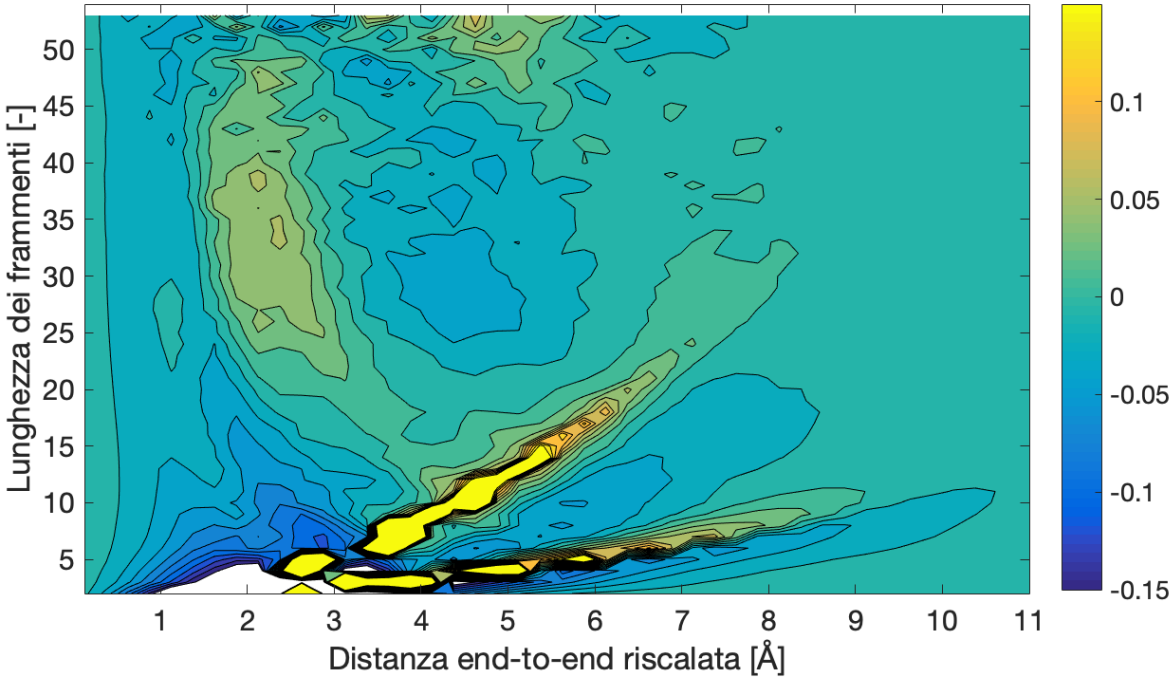


Figura 3.10: Superficie degli scarti per a=3/2

Da questi grafici si possono fare alcune osservazioni importanti;

- Il test del χ^2 risulta non passato in tutti i casi analizzati. Questo e' giustificabile alla luce della presenza di correlazioni fra i diversi frammenti presenti nel data set analizzato estratti dalla stessa proteina. Il numero di frammenti estratti dalla stessa proteina cala all'aumentare di l, spiegando la corrispondente diminuzione del chi quadro. L'utilizzo del test del χ^2 infatti richiede che le misure siano statisticamente indipendenti.
- Le curve del χ^2 al variare di l mostrano un andamento decrescente all'aumentare di a, e questo fenomeno è visibile chiaramente per frammenti lunghi. Anche in questo caso l'effetto e' presumibilmente dovuto al diminuire del numero di frammenti estratti dalla stessa proteina con l'aumentare di a.

- Le curve del χ^2 si interrompono a lunghezze sempre più piccole all'aumentare del parametro a; questo è banalmente giustificato dal fatto che un incremento del parametro a porta una maggiore restrittività nella selezione delle proteine idonee all'analisi tramite la condizione $N > al^{3/2}$, fino a scartare tutte le proteine oltre un certo valore di l.

- Le superfici degli scarti mostrate nei *contour plot* mostrano delle variazioni qualitative all'aumentare del parametro a riassumibili nei seguenti aspetti; La valle, costituita da scarti negativi, corrispondente alla fascia delle distanze riscalate comprese tra 0 Å e 2 Å, mostra un appiattimento all'aumentare di a. Le creste invece non sembrano abbassarsi uniformemente.

Come ultimo punto dell'analisi si riporta in figura 3.11 l'andamento delle stime del parametro b della distribuzione di Maxwell-Boltzmann interpolante in funzione della lunghezza l dei frammenti, per i 3 diversi valori del parametro a analizzati in questo lavoro.

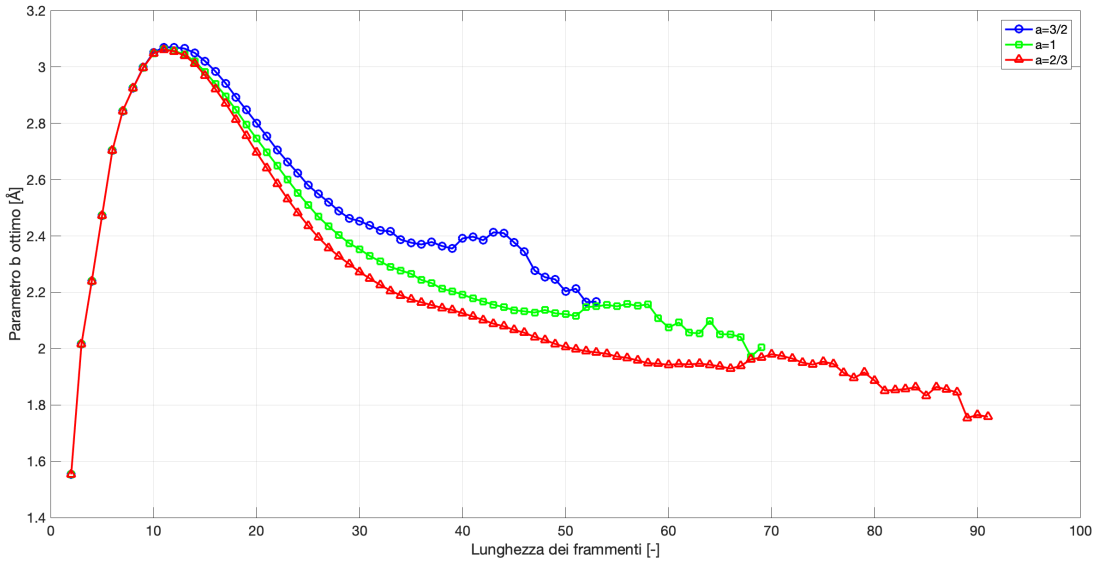


Figura 3.11: Grafico dell'andamento di b al variare di l per diversi parametri a

Alcune note sulle 3 curve di quest'ultimo ultimo grafico;

- sono molto vicine tra loro per l piccole, mentre si differenziano per frammenti più grandi al variare di a. In particolare, il parametro b si incrementa all'aumentare di a.

- Tutte le curve sembrano mostrare un plateaux per valori intermedi di l, per poi decrescere all'aumentare di l; i valori di b in corrispondenza dei plateaux aumentano all'aumentare di a, mentre i corrispondenti valori di l diminuiscono. Solo nel caso a=1 si ottiene il valore di plateaux confrontabile con $3.8/\sqrt{3}$ Å discusso in precedenza.

Conclusioni

Lo scopo dell'analisi è quello di verificare che la regione interna di singoli domini di proteine globulari si comporti come un *polymer melt* nel regime di Flory, rispettando quindi una statistica delle distanze end-to-end propria di una catena ideale, ovvero una distribuzione di Maxwell-Boltzmann.

Questo fatto è stato verificato alla luce di tutte le interpolazioni fatte su ogni istogramma delle distanze end-to-end riscalate, ottenendo dei fit apparentemente soddisfacenti. Il valore molto alto del χ^2 e' presumibilmente dovuto alla presenza di correlazione fra i diversi frammenti estratti dalla stessa proteina. Questa ipotesi si potrebbe validare provando per esempio a costruire un data set ridotto tramite un vincolo sul numero di frammenti estratti dalla stessa proteina.

La presenza di una zona di saturazione per i valori della lunghezza di legame della catena ideale stimati tramite fit con la distribuzione di Maxwell-Boltzmann e' un altro indizio a favore dell'esistenza del regime di Flory all'interno delle proteine globulari.

E' stato poi studiato come i risultati ottenuti dipendono dal grado di selettivita' adoperato nell'imporre che i frammenti analizzati appartengano all'interno del globulo. Il risultato interessante e' che la stima della lunghezza di legame della catena ideale nella zona di saturazione dipende dipende dal grado di selettività

Infine e' stato evidenziato uno scostamento sistematico fra la distribuzione interpolante di Maxwell-Boltzmann e l'istogramma empirico. Anche l'entita' di questo scostamento dipende dipende dal grado di selettività.

Il regime di Flory in un melt di polimeri dovrebbe essere caratterizzato da un valore uniforme della lunghezza di legame. I risultati ottenuti in questo lavoro di tesi mostrano che per validare l'effettiva presenza del regime di Flory all'interno delle proteine globulari e' necessaria una comprensione piu' accurata di come le leggi di scala Gaussiane osservate dipendano dal grado di selettivita' dei frammenti analizzati.

Bibliografia

[1] A.V. Finkelstein e O. Ptitsyn. *Protein Physics, A course of lectures*. Lectures 1, 2, 7, 13, 14. Academic Press, An Imprint of Elsevier Science, 2002.

[2] A. Maritan J. R. Banavar T. X. Hoang. «Proteins and polymers». In: *The Journal of Chemical Physics* 122 (2005).

[3] *Plot of the Lennard-Jones potential*. URL: `https://www.researchgate.net/figure/Plot-of-the-Lennard-Jones-potential-as-stated-in-eq-11-Using-the-parameters_fig1_281032446`.

[4] *Plot of the Mayer f-function*. URL: `https://www.semanticscholar.org/paper/Soft-interactions-and-crowding-Sarkar-Li/82bfc8d99b78231fdc061974d95831f5f72c0324`.

[5] *Protein Data Bank - methods for determinig structure*. URL: `http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure`.

[6] *Protein Data Bank homepage*. URL: `https://www.rcsb.org`.

[7] M. Rubinstein e R.H. Colby. *Polymer Physics*. Cap.1, 2, 3, 4. Oxford University Press, 2003.

[8] M. Baiesi E. Orlandini F. Seno A. Trovato. «Sequence and structural patterns detected in entangled proteins reveal the importance of cotranslational folding». In: *Scientific Reports* 9 (2019).