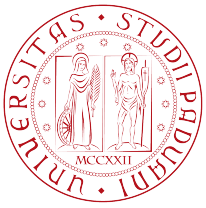


Università degli Studi di Padova

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Metodi Monte Carlo gran canonici per simulare modelli a grana grossa di proteine

Laureando:

Lorenzo Pantolini

Relatore:

Prof. Antonio Trovato

Anno accademico 2016/2017

Indice

Capitolo 1

Introduzione

1.1 Generalità sulle proteine

Le proteine sono polimeri, ovvero macromolecole costituite da un gran numero di gruppi molecolari uguali o simili tra loro, in questo caso da aminoacidi. In natura esistono 20 diversi tipi di aminoacidi, questi sono molecole organiche formate da un gruppo peptidico e da un residuo (*catena laterale*) che presentano la seguente struttura tridimensionale:

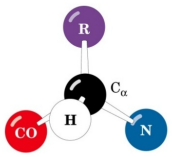


Figura 1.1: Struttura di tipo L

La presenza di una catena laterale più pesante di un atomo di idrogeno, fa si che ogni aminoacido, tranne la glicina abbia una chiralità definita. Gli aminoacidi in natura sono levogiri, questi si legano tra di loro tramite legami di tipo covalente, chiamati peptidici, tra gli atomi N e i gruppi CO andando a formare sequenze che possono andare da poche dozzine a diverse centinaia di aminoacidi. Le catene così formate consistono in una struttura principale chimicamente regolare (*backbone o catena principale*) dalla quale si distendono diverse catene laterali. Le funzioni biologiche svolte dalle proteine sono moltissime e dipendono strettamente dalla struttura tri-dimensionale assunta dalla catena proteica, questa viene chiamata struttura nativa. Infatti in una proteina "operativa" la catena presenta un'architettura molto specifica e piccoli cambiamenti in essa possono portare a pesanti cambiamenti nelle sue attività, se non addirittura cessarle completamente. Il premio Nobel per la chimica Christian Anfinsen postulò che, per quanto riguarda piccole proteine globulari[1], "lo stato nativo di una proteina è codificato nella sua sequenza primaria" [1]. Tale affermazione significa che la struttura spaziale è dovuta esclusivamente alla particolare sequenza di aminoacidi, detta sequenza primaria, che la compone; dunque sono essi a determinare le funzionalità della proteina. Tale principio è conosciuto come *ipotesi termodinamica di Anfinsen* ed ha una conseguenza molto importante: nelle condizioni ambientali in cui avviene il ripiegamento proteico il sistema proteina più solvente presenta un minimo stabile di energia libera in corrispondenza della struttura nativa. Le possibili architetture assumibili dalla catena proteica sono molto varie e complesse, tuttavia esistono alcuni motivi

[1] Le proteine globulari sono proteine nelle quali la catena polipeptidica tende ad avvolgersi su sé stessa.

ricorrenti. Vengono riconosciuti 4 livelli di struttura:

- *primaria:* ovvero la specifica sequenza di aminoacidi che compone il backbone.
- *secondaria:* sono strutture geometriche ordinate localizzate in diverse parti della proteina, tra queste riconosciamo: l'α-elica, struttura elicoidale e il β-foglietto, struttura planare.
- *terziaria:* corrisponde all'effettiva struttura tri-dimensionale assunata dalla proteina nello stato nativo.
- *quaternaria:* riguarda la disposizione spaziale e topologica di proteine molto grandi formate da più sub-unità.

1.2 Ripiegamento proteico

Il ripiegamento proteico (*protein folding*) corrisponde a quell'insieme di fenomeni attraverso cui la catena di aminoacidi arriva ad assumere lo stato nativo della proteina. Questo inizia già durante la sintesi della proteina ed è dovuto ad un vasto numero di interazioni non covalenti tra gli elementi della catena, tra le quali:

- legami ad idrogeno
- forze di Van der Waals
- interazioni π-π
- coordinazione di metalli
- volume escluso (gli orbitali elettronici di diversi atomi non possono sovrapporsi tra loro)
- forze idrofobiche (quando il processo avviene in solvente acquoso)

Sono proprio le forze idrofobiche che portano la proteina ad assumere una forma compatta globulare, infatti alcune delle catene laterali sono idrofobiche dunque la catena principale tende a disporsi in modo da minimizzare le interazioni tra queste e le molecole d'acuqa. Il processo di folding è estremamente rapido, per piccole proteine il tempo di ripiegamento può essere anche dell'ordine dei microsecondi. Il fenomeno contrario al protein folding, invece, è detto denaturazione, questa può avvenire tramite metodi sia chimici che fisici (*denaturazione termica*). Tale processo modifica la struttura tri-dimensionale della proteina facendole dunque perdere la sua funzione biologica, tuttavia non intacca i legami peptidici lasciando quindi invariato il backbone. Per quanto riguarda le proteine più piccole la denaturazione è una transizione cooperativa, sembra infatti che queste passino dalla struttura nativa allo stato denaturato senza passare per altri stati stabili parzialmente denaturati; si parla dunque di "all-or-none transition". Il fenomeno ricorda molto la coesistenza tra "fasi" nelle transizioni liquido vapore. In realtà la catena passa per tutti gli stati intermedi necessari, tuttavia il processo avviene così velocemente da rendere proibitiva l'identificazione degli stati parzialmente denaturati. Esistono dunque due stati stabili della catena proteica, questi sono separati da una barriera di energia libera:

$$F = E - TS \tag{1.1}$$

Dove E è l'enegia interna, T la temperatura e S l'entropia. Durante il ripiegamento tali componenti assumono tipicamente questo andamento:

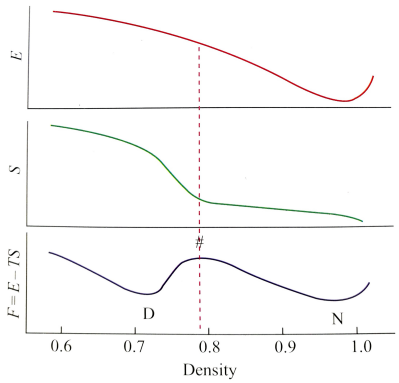


Figura 1.2: Andamento di Energia, Entropia ed Energia Libera preso da [2]

L'energia è espressa in funzione di una coordinata che descrive il passaggio da uno stato all'altro, questa vale 1 nello stato nativo. Come mostra il grafico la barriera tra i due stati è dovuta ad un repentino aumento di entropia durante il processo di denaturazione, questo porta l'energia libera ad avere due minimi, uno in corrispondenza dello stato nativo e l'altro di quello denaturato.

1.3 Obiettivo

Sono molteplici le ragioni che motivano lo studio del protein folding, ma in particolare due di queste rivestono notevole importanza. Da una parte si cerca di giungere a metodi e modelli che permettano di predire lo stato nativo di una proteina a partire dalla sequenza dei suoi aminoacidi. Questo permetterebbe di elaborare delle tecniche tramite cui progettare nuove proteine. D'altra parte giungeremmo alla comprensione dei fenomeni fisici essenziali alla base del processo di folding.
In questa tesi cercheremo di riprodurre il processo di ripiegamento con l'obiettivo di testare una nuova tecnica di simulazione Monte Carlo. In particolare andremo a studiare una proteina ampiamente conosciuta in modo da poter verificare i risultati ottenuti, la scelta è ricaduta sulla: *serine proteinase inhibitor* o $2CI2$. Questa è composta da 65 aminoacidi e nello stato nativo presenta la seguente struttura tri-dimensionale:

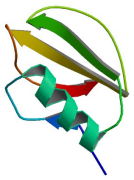


Figura 1.3: Le frecce rappresentano la presenza di β-foglietti, mentre la spirale rappresenta un'α-elica

Capitolo 2

Modelli

2.1 Modelli di Gō

Per implementare la simulazione è stato necessario definire un modello che descrivesse la proteina in modo da semplificare il sistema da studiare e ridurne i gradi di libertà. La scelta è ricaduta su un modello a grana-grossa che consiste nel sostituire la catena proteica principale con una di sfere dure, a cui ci riferiremo come pseudo-atomi o $C\alpha$, giacenti nelle posizioni occupate dai $C\alpha$ nella catena originale. Nella proteina la distanza tra due gruppi $C\alpha$ consecutivi è all'incirca costante e vale ≈ 3.8 Å, dunque potremo considerarla fissa. In questo modo, una volta fissati gli assi e l'origine di un sistema cartesiano, il modello presenta $2N + 1$ gradi di libertà, dove N è il numero di $C\alpha$ presenti nella proteina, nel nostro caso 65. Dato che a noi interessano solo le posizioni relative tra i vari aminoacidi e non la posizione che la proteina assume nello spazio come un corpo rigido, a questi dovremo sottrarre altri 6 gradi ottenendone $2N - 5$.

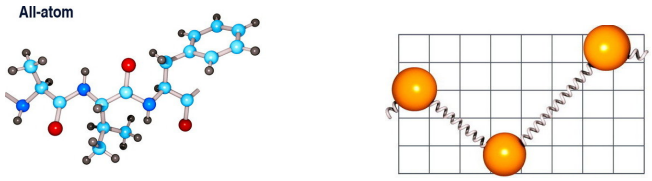


Figura 2.1: Proteina di tre aminoacidi rappresentata con tutti gli atomi e rispettivo modello a grana-grossa

Per indagare il processo di ripiegamento della proteina è stato necessario modellizzare anche il potenziale di interazione tra i vari aminoacidi. Infatti procedere con una ricerca casuale delle possibili configurazioni porterebbe al così detto paradosso di Levinthal [3]. Secondo questo una tale ricerca, anche per una proteina di solo un centinaio di aminoacidi, impiegherebbe un tempo superiore all'età dell'universo. Il modello che siamo andati ad utilizzare, noto come modello di $G\bar{o}$ [4], si basa sull'assunzione che i contatti nativi, ovvero i residui che interagiscono nella proteina completamente piegata, siano gli unici a contribuire al processo di ripiegamento. Questo genere di approccio è stato proposto per la prima volta da Nobuhiro Gō, dal quale il modello prende il nome, ed ha portato a numerosi risultati concordanti con quelli sperimentali [5]. A prescindere dai dettagli con cui viene definito il potenziale l'idea di base è quella di costruirne uno il cui minimo dovrà coincidere con la struttura nativa. Per questo motivo i modelli di tipo Gō si possono utilizzare solo se la struttura nativa è nota. Nel nostro potenziale vogliamo tener conto del così detto "volume escluso", infatti ogni pseudo-atomo genera una zona inaccessibile per gli altri, eccetto che per il precedente e il successivo. Questa zona equivale ad una sfera di

raggio 4 Å centrata nel suddetto $C\alpha$. Vogliamo inoltre tener conto delle interazioni attrattive che hanno luogo nello stato nativo. In particolare abbiamo considerato interagire due pseudo-atomi, non consecutivi, quando la loro distanza nella struttura nativa è inferiore a 6.5 Å.
Il potenziale effettivamente utilizzato assume la forma:

$$H = \sum_{i=1}^{N} \sum_{j=i+2}^{N} \left[V^{Nat}(r_{ij}) \Delta_{ij} + V^{NNat}(r_{ij})(1 - \Delta_{ij}) \right] \tag{2.1}$$

Dove:

- N è il numero totale di pseudo-atomi.
- r_{ij} è la distanza tra il $C\alpha$ i-esimo e il $C\alpha$ j-esimo.
- Δ è una matrice simmetrica conosciuta come mappa di contatto le cui entrate, Δ_{ij}, sono uguali a 1 se il $C\alpha$ i-esimo e il $C\alpha$ j-esimo interagiscono nella struttura nativa, e 0 altrimenti.
- $V^{Nat}(r_{ij})$ è il potenziale di interazione per residui in contatto nella struttura nativa, che vale:

$$V^{Nat}(r_{ij}) = \begin{cases} \infty & 0 \leq r_{ij} < 4.0 \text{ Å} \\ -1 & 4.0 \text{ Å} \leq r_{ij} \leq 6.5 \text{ Å} \\ 0 & r_{ij} > 6.5 \text{ Å} \end{cases} \tag{2.2}$$

- $V^{NNat}(r_{ij})$ è il potenziale che descrive il volume escluso per i $C\alpha$ che non interagiscono nella struttura nativa:

$$V^{NNat}(r_{ij}) = \begin{cases} \infty & 0 \leq r_{ij} < 4.0 \text{ Å} \\ 0 & r_{ij} \geq 4.0 \text{ Å} \end{cases} \tag{2.3}$$

Inoltre è necessario un termine che induca nella struttura di minima energia la chiralità propria dello stato nativo[1], abbiamo dunque aggiunto il seguente potenziale angolare:

$$V^{Ang} = h \sum_{i=1}^{N-2} (\theta_{i,i+1,i+2} - \theta^{Nat}_{i,i+1,i+2})^2 + k \sum_{i=1}^{N-3} (1 - \cos(\phi_{i,i+1,i+2,i+3} - \phi^{Nat}_{i,i+1,i+2,i+3})) \tag{2.4}$$

Dove:

- $\theta_{i,j,k}$ è l'angolo formato tra gli pseudo-atomi i-esimo, j-esimo e k-esimo.

[1] Determinata dal volume escluso delle catene laterali e dalla chiralità definita dagli aminoacidi naturali.

- $\phi_{i,j,k,l}$ è l'angolo diedro formato tra gli pseudo-atomi i-esimo, j-esimo, k-esimo e l-esimo; la trasformazione $\phi \to -\phi$ realizza una struttura tridimensionale analoga con chiralità opposta.
- $\theta^{Nat}_{i,j,k}$ e $\phi^{Nat}_{i,j,k,l}$ sono i corrispondenti angoli formati nella struttura nativa.
- h e k sono due costanti che servono a determinare il peso di questo termine del potenziale. I valori utilizzati nella simulazione, in unità adimensionali, sono $h = 5$ e $k = 0.3$ [2].

Notiamo che il contributo del potenziale angolare è sempre positivo, e si annulla esclusivamente quando gli angoli della proteina simulata coincidono con quelli della struttura nativa. Questo contributo serve a facilitare le configurazioni con la corretta orientazione angolare e la corretta chiralità.

2.2 Cambio di coordinate

Dato che nel corso della simulazione vengono utilizzati due differenti sistemi di riferimento è utile introdurre il metodo che utilizzeremo per cambiare coordinate. Infatti durante il processo di costruzione della catena proteica la scelta più comoda è un sistema locale di coordinate sferico-polari centrato, di volta in volta, nell'ultimo pseudo-atomo aggiunto. Invece per calcolare il potenziale di interazione, dato che è in funzione delle distanze tra i $C\alpha$, è più naturale utilizzare un sistema di coordinate cartesiane.
Dati dunque due pseudo-atomi siano: $\vec{r}_{i-1}$ e $\vec{r}_i$ le loro coordinate in un determinato sistema di riferimento cartesiano. Siano invece $\theta \in [0, \pi]$ e $\phi \in [0, 2\pi]$ le coordinate sferico-polari centrate in $\vec{r}_i$ di un terzo pseudo-atomo. Per esprimere queste ultime in cartesiane è utile definire i tre assi del sistema locale:

$$\vec{z}_i = \frac{\vec{r}_i - \vec{r}_{i-1}}{|\vec{r}_i - \vec{r}_{i-1}|} \tag{2.5}$$

$$\vec{y}_i = \frac{\vec{r}_{i-1} \times \vec{r}_i}{|\vec{r}_{i-1} \times \vec{r}_i|} \tag{2.6}$$

$$\vec{x}_i = \vec{y}_i \times \vec{z}_i \tag{2.7}$$

Da queste possiamo ottenere il versore che porta da $\vec{r}_i$ a $\vec{r}_{i+1}$, questo sarà:

$$\vec{z}_{i+1} = cos(\phi)sin(\theta) \cdot \vec{x}_i + sin(\phi)sin(\theta) \cdot \vec{y}_i + cos(\theta) \cdot \vec{z}_i \tag{2.8}$$

Avremo in fine:

$$\vec{r}_{i+1} = \vec{r}_i + R \cdot \vec{z}_{i+1} \tag{2.9}$$

Dove R è la distanza tra due $C\alpha$ consecutivi.

[2] Tali costanti sono state prese da [6].

Capitolo 3

Simulazione

3.1 Implementazione del metodo di simulazione

Il processo di folding è stato riprodotto tramite una simulazione Monte-Carlo. La nozione di metodi, algoritmi o tecniche Monte-Carlo sintetizza una vasta area di metodi basati sul campionamento di numeri casuali [7]. Si tratta dunque di una simulazione stocastica la cui forza principale risiede nella grande efficacia computazionale. Infatti, come si può intuire dal paradosso di Levinthal, un approccio deterministico al nostro problema sarebbe proibitivo a causa del gran numero di possibili configurazioni[1] assumibili dalla proteina. Dato che lo scopo della simulazione è riprodurre il processo di ripiegamento sarà necessario un metodo per vagliare le diverse configurazioni assunte dalla catena proteica. In generale non è banale costruire algoritmi efficienti per campionare le strutture di polimeri (e quindi di proteine) compatti, a causa dell'effetto di volume escluso che porta a non considerare gran parte delle configurazioni campionate. Nel nostro caso l'idea è quella di aggirare il problema cambiando la forma della catena, "smontandola" e poi ricostruendola di volta in volta in maniera differente. Notiamo che in questo modo il sistema passerà per un gran numero di stati a lunghezza intermedia, mentre ovviamente noi saremo interessati solo a quelli di lunghezza massima (è interessante osservare che sono stati studiati modelli di Gō per indagare il ripiegamento di frammenti di proteina con diverse lunghezze, mentre questa viene assemblata sul ribosoma [8]). La mossa *Monte-Carlo* che vogliamo testare consiste dunque nell'aggiungere o togliere uno degli pseudo-atomi esterni della catena. La simulazione presenta le seguenti fasi:

1. Viene selezionata l'estremità della catena su cui lavorare.
2. Viene selezionato uno dei processi possibili tra: *aggiungere*, *togliere* e *non fare nulla*.
3. In dipendenza dal processo selezionato, viene modificata la catena, *aggiungendo/togliendo* uno degli pseudo-atomi.
4. Vengono eseguiti dei test per decidere se accettare o meno la mossa di prova. I due test effettuati sono: *Volume escluso* e *Metropolis*.

La probabilità di lavorare su un lato piuttosto che un altro è fissa e vale $\frac{1}{2}$. Invece quella di aggiungere p_+ è uno dei parametri cruciali della simulazione e va scelto con grande cura. Gli elementi aggiunti si troveranno nella sfera di raggio 3.8 $\mathring{\text{A}}$ centrata nello pseudo-atomo

[1] Per configurazione si intende la particolare struttura spaziale assunta dalla catena.

più esterno. Per generare tali punti in maniera uniforme nella sfera è necessario distribuire opportunamente le coordinare sferiche θ e ϕ di tale sistema locale. Dato che l'elemento di angolo solido di una sfera vale:

$$d\Omega = \sin(\theta) d\theta d\phi \tag{3.1}$$

Definendo $\lambda = \cos(\theta)$ avremo $d\lambda = \sin(\theta)d\theta$ e potremo dunque riscrivere:

$$d\Omega = d\lambda d\phi \tag{3.2}$$

In tal modo potremmo ottenere la distribuzione cercata distribuendo ϕ uniformemente tra $[0, 2\pi]$ e λ uniformemente tra $[-1, 1]$. L'elemento di angolo solido $d\Omega$ viene quindi selezionato con probabilità $d\Omega/4\pi$, vale a dire con densità di probabilità uniforme pari a $1/4\pi$. Una volta campionati gli angoli si procede con il cambio di coordinate descritto in precedenza per ottenere le coordinate cartesiane dello pseudo-atomo aggiunto.

3.2 Catene di Markov

Un processo aleatorio nel quale la probabilità di transizione, che determina il passaggio da uno stato del sisitema ad un altro, dipende esclusivamente dallo stato immediatamente precedente è detto processo di Markov. Nel caso in cui questo processo sia definito per tempi discreti si parla invece di *catena* di Markov.
La nostra simulazione è un metodo Monte Carlo basato su una catena di Markov, questa infatti genera una sequenza di configurazioni $S^{(n)}$:

$$S^{(1)} \to S^{(2)} \to ... \to S^{(n)} \to ... \tag{3.3}$$

ognuna delle quali dipende esclusivamente da quella precedente ed ha una ben definita probabilità di transizione $P(S^{(n-1)} \to S^{(n)})$ che soddisfa le seguenti proprietà:[2]

$$P(S \to S') \geq 0 \qquad \sum_{S'} P(S \to S') = 1 \tag{3.4}$$

Si può dimostrare che una catena di Markov converge ad una determinata distribuzione stazionaria $P(S)$ se, oltre ad una proprietà di ergodicità, vale la proprietà nota come *bilancio dettagliato* [7]:

$$P(S)P(S \to S') = P(S')P(S' \to S) \tag{3.5}$$

Le probabilità introdotte nelle equazioni (3.4) e (3.5) possono essere sia probabilità discrete vere e proprie sia densità di probabilità continue. É comodo riscrivere la probabilità di transizione da uno stato all'altro in tale maniera:

$$P(S \to S') = P_s(S \to S')P_a(S \to S') \tag{3.6}$$

Dove P_s è la probabilità di selezionare una determinata configurazione S' mentre P_a è quella di accettarla. Nel caso in cui si scelga di aggiungere $P(S \to S') = \frac{p_+}{4\pi}$ ed è una densità di probabilità, mentre nel caso in cui si tolga $P(S \to S') = 1 - p_+$ che è una probabilità.
Lo scopo della simulazione è quello di portare il sistema all'equilibrio termodinamico, dunque la distribuzione di probabilità continua P verso la quale vorremmo far tendere il sistema sarà

[2] Useremo il formalismo per gli stati discreti, anche se nel nostro caso le variabili angolari utilizzate sono continue.

quella di Boltzmann:

$$P_B(S) = \frac{e^{-\frac{E(S)}{K_B T}}}{\sum_{S'} e^{-\frac{E(S')}{K_B T}}} \cdot \frac{z^{N(S)}}{z^{N(S')}} \tag{3.7}$$

Dove E è l'energia del sistema descritta dal potenziale modellizzato nel precedente capitolo, T è la temperatura, K_B è la costante di Boltzmann, $N(S)$ è il numero di aminoacidi dello stato S e z è la fugacità del sistema. Quest'ultima, in pratica, viene determinata dalla scelta di p_+, con l'obiettivo di ottenere una simulazione che campiona catene di lunghezze diverse in maniera simile:

$$z = \frac{p_+}{4\pi(1 - p_+)} \tag{3.8}$$

Notiamo che nella simulazione implementata P_s equivale alla probabilità di aggiungere/togliere uno pseudo-atomo, e dunque di passare da una configurazione con n elementi ad una con $n + 1$ / $n - 1$ elementi. Invece la probabilità P_a di accettare o meno tale cambio di stato è fornita dai test menzionati nel punto 4. Questi dovranno rispettare la condizione di bilancio dettagliato. La simulazione è stata costruita partendo da un "Random walk", successivamente è stato implementato il test per il volume escluso ed infine è stato aggiunto il test di Metropolis in modo da ottenere la distribuzione cercata.

3.3 Random walk

Per "random walk" si intende un'evoluzione comletamente casuale del sistema. In pratica un macrostato S ha la stessa probabilità di passare in ogni macrostato S' a lui accessibile. Ricordiamo che la catena assume diverse lunghezze nel corso della simulazione, in particolare queste andranno da 3 a 65 pseudo-atomi[3]. Un buon modo per vagliare lo spazio delle configurazioni della proteina completa è quello di campionare stati di lunghezza diversa in maniera simile. Implementando un "random walk" nello spazio delle lunghezze otteniamo proprio questo risultato. Perchè ciò avvenga è necessario definire opportunamente le probabilità p_+ e P_a. Nel nostro caso $P_a = 1$, questo comporta che ogni cambiamento proposto viene accettato (non viene dunque eseguito alcun test). Dovremmo inoltre avere la stessa probabilità di aggiungere uno pseudo-atomo che di toglierlo, $p_+ = \frac{1}{2}$ e quindi $z = \frac{1}{4\pi}$. Sia dunque S^n il macrostato corrispondente a tutti gli stati con lunghezza n, questo si potrà raggiungere dal macrostato S^{n-1} e da quello S^{n+1}, in entrambi casi con probabilità $p_+ = p_- = \frac{1}{2}$.

$$S^{n-1} \longleftarrow S^n \longrightarrow S^{n+1} \tag{3.9}$$

Questo tuttavia non è sempre possibile, infatti ci sono particolari situazioni in cui non si può passare a macrostati con determinate lunghezze. Per queste dovranno essere implementate particolari condizioni a contorno. Ci si ritrova in tali situazioni quando:

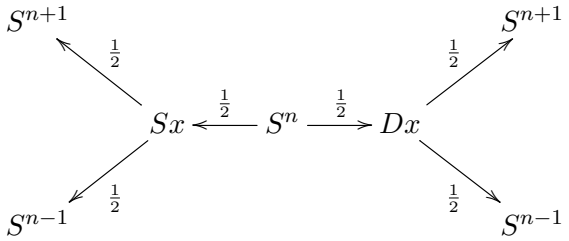
- Si cerca di *aggiungere* a lunghezza massima.
- Si cerca di *rimuovere* a lunghezza minima.
- Uno dei $C\alpha$ si trova in una delle estremità della catena, impedendo così di aggiungere da uno dei due lati.

[3]Il limite massimo è 65 perchè non avrebbe senso superare la lunghezza della proteina, quello minimo invece è una scelta.

- Si ha una combinazione tra le situazioni elencate.

I macrostati di lunghezza massima e minima, al contrario degli altri, sono raggiungibili da un solo macrostato, è dunque necessario che abbiano la possibilità di "tornare" in loro stessi sempre con probabilità $\frac{1}{2}$, in modo che la proabilità di lasciarli sia sempre $\frac{1}{2}$:

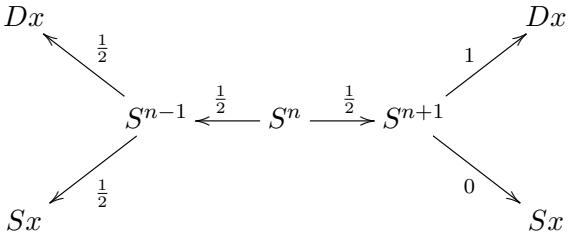
Per spiegare come risolvere le altre condizioni è utile uno schema semplificativo, mostriamo la situazione generale:



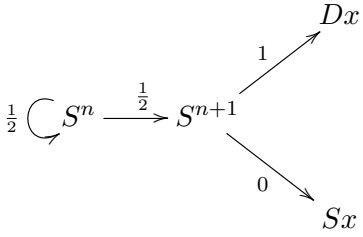
Dato un macrostato S^n lo schema mostra il processo di selezione del macrostato in cui esso evolve. Il numero sopra la freccia indica la probabilità di effettuare quella "scelta" mentre con Dx e Sx si intende il lato su cui agire. In questo modo possiamo calcolare semplicemente la probabilità di transizione tramite una somma, nel caso in questione avremo:

$$P_s(S^n \to S^{n+1}) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = P_s(S^n \to S^{n-1}) \tag{3.10}$$

come ci aspettiamo. Valutiamo ora il caso in cui uno pseudo atomo si trovi in prima posizione impedendoci dunque di aggiungere a sinistra e che la lunghezza della catena sia maggiore di 3. In questo caso per mantenere le caratteristiche del processo sarà necessario cambiare approccio, altrimenti sarebbe più probabile passare ad un macrostato di lunghezza inferiore. La strategia utlizzata è la seguente:



Per completezza mostriamo il caso analogo ma con lunghezza 3:



Ometteremo le situazioni con uno pseudo-atomo in ultima posizione, con lunghezza massima e tutte le altre combinazioni possibili in quanto sono molto simili a quelle viste.

Andando a calcolare il bilancio dettagliato con questa scelta di P_a e p_+ troviamo che, per i macrostati:

$$P(S^n) = P(S^{n+1}) \tag{3.11}$$

Questo significa che avremo la stessa probabilità di trovare ogni macrostato con diversa lunghezza, e dunque che il random walk fornisce la distribuzione cercata, ovvero quella uniforme[4]. Mostriamo ora i risultati ottenuti con la simulazione:

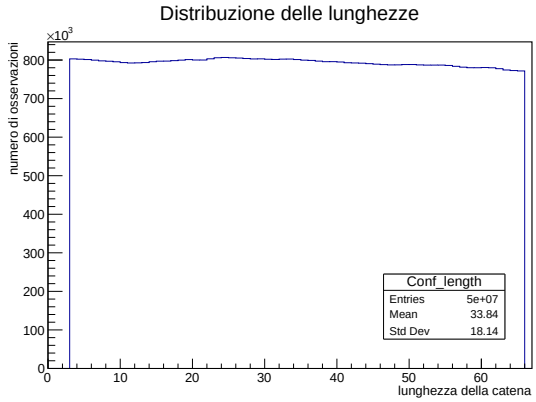


Figura 3.1: Istogramma i cui canali contengono le lunghezze delle configurazioni ottenute

3.4 Volume escluso

Il passo successivo è stato implementare l'elemento di potenziale responsabile del volume escluso. In pratica ogni volta che si tenta di aggiungere un $C\alpha$ si controlla se questo occupa una zona inaccessibile e in caso la mossa viene rifiutata. Questo test va dunque a modificare la probabilità P_a nel caso in cui si aggiunga uno pseudo-atomo, tale probabilità è difficile da stimare a priori è però possibile farne una stima tramite i risultati di una simulazione. Consideriamo la condizione di bilancio detagliato per il passaggio dal macrostato S^3 a S^4:

$$P(S^3)P_s(S^3 \to S^4)P_a(S^3 \to S^4) = P(S^4)P_s(S^4 \to S^3)P_a(S^4 \to S^3) \tag{3.12}$$

Dato che la simulazione è stata costruita in modo che $p_+ = p_-$ e che $P_a(S^{n+1} \to S^n) = 1$ potremo riscrivere la condizione in questo modo:

$$P(S^4) = P(S^3)P_a(S^3 \to S^4) \tag{3.13}$$

Ipotizzando di fare n transizioni e definendo $\gamma = \frac{1}{P_a(S^n \to S^{n+1})}$ otterremmo:

$$P(S^{n+3}) = \left(\frac{1}{\gamma}\right)^n P(S^3) \tag{3.14}$$

Riscrivendola poi come esponenziale si avrà:

$$P(S^{n+3}) = P(S^3)e^{-n\ln(\gamma)} \tag{3.15}$$

[4]Le condizioni a contorno utilizzate sono riflettenti e portano ad una distribuzione uniforme.

Questo significa che in queste condizioni la probabilità di ottenere un determinato macrostato decresce esponenzialmente con la sua lunghezza. Questo risultato viene confermato dai risultati simulati:

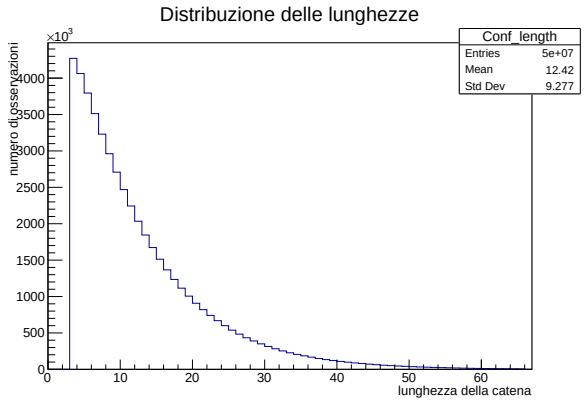


Figura 3.2: Distribuzione delle lunghezze con "*Volume escluso*" e $p_+ = 0.5$

Tramite un fit esponenziale è poi possibile ricavare γ, stimando dunque P_a. Sarà inoltre possibile trovare il modo di modificare la probabilità p_+ in modo da ottenere nuovamente una distribuzione uniforme. Dato che in questo caso $p_+ \neq \frac{1}{2}$ si dovrà riscrivere il bilancio dettagliato:

$$\frac{P(S^n)}{P(S^{n+1})} = \gamma \frac{p_+}{1-p_+} \tag{3.16}$$

Per ottenere una distribuzione uniforme si impone $P(S^n) = P(S^{n+1})$ e dunque:

$$P_s(S^n \to S^{n+1}) = p_+ = \frac{\gamma}{1+\gamma} \tag{3.17}$$

Dai risultati del fit otteniamo $\gamma = 1.12$. Avremo dunque $P_a(S^n \to S^{n+1}) = 0.89$ e $p_+ = 0.528$. Ecco la distribuzione con la nuova probabilità di selezione:

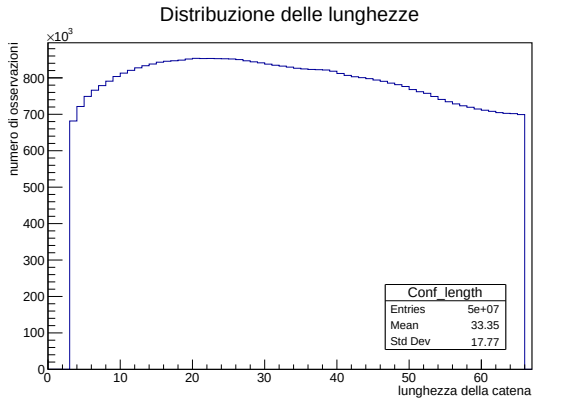


Figura 3.3: Distribuzione delle lunghezze con "*Volume escluso*" e $p_+ = 0.528$

Possiamo attribuire la leggerera deviazione dalla uniformità alla grande sensibilità che presenta la distribuzione rispetto alla probabilità p_+. Dato che la stima di γ presenterà sicuramente qualche errore (o nel fit o in un'approssimazione) è naturale aspettarsi un andamento non perfettamente uniforme.

3.5 Metodo di Metropolis

L'ultimo test implementato è quello di *Metropolis*. Questo ha lo scopo di definire la probabilità P_a in modo che il bilancio dettagliato sia soddisfatto per la distribuzione di Boltzmann. Data una distribuzione $P(S)$ la probabilà di accettare o meno un cambio di stato viene definita come:

$$P_a(S \to S') = min\left[1, \frac{P(S')}{P(S)}\right] \tag{3.18}$$

Si può dimostrare che tale scelta soddisfa la condizione di bilancio dettagliato per la densità di probabilità P(S) se $P_s(S \to S') = P_s(S' \to S)$ [7]. Nel nostro caso la condizione su P_s per gli stati non è rispettata e si può ricorrere all'algoritmo di *Metropolis-Hastings* per soddisfare il bilancio dettagliato [7]:

$$P_a(S \to S') = min\left[1, \frac{P(S')P_s(S' \to S)}{P(S)P_s(S \to S')}\right] = min\left[1, e^{\frac{-\Delta E(S,S')}{K_B T}}\right] \tag{3.19}$$

dove è stata usata la definizione (3.8) per la fugacità e si riottiene, per il test di accettazione usato, la forma standard del test di Metropolis.

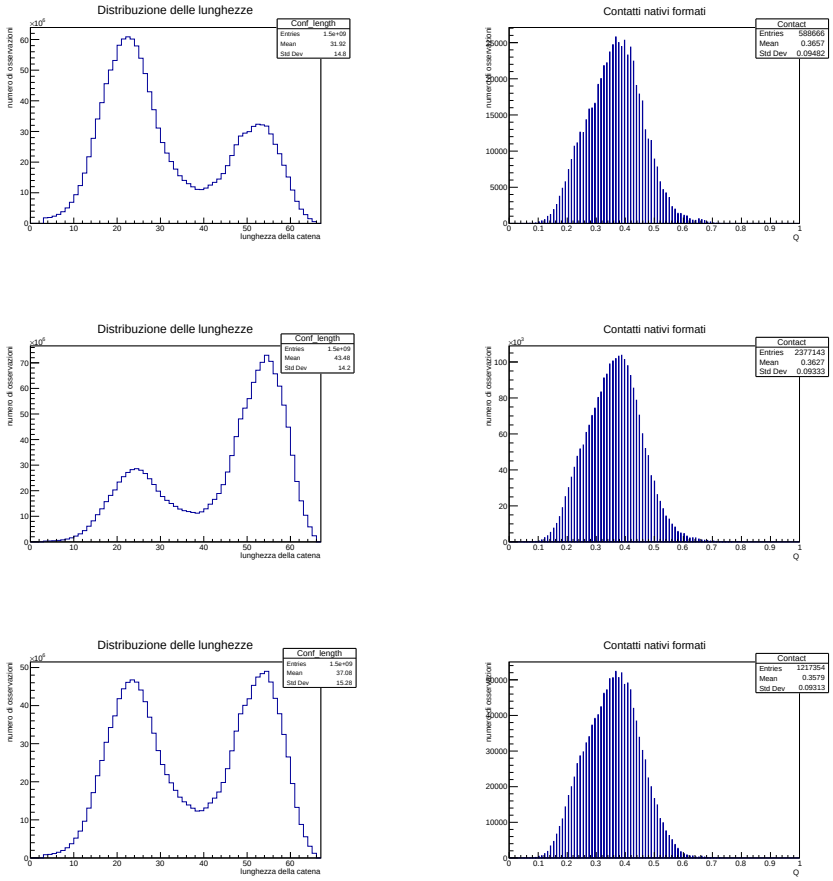
Capitolo 4

Risultati

In questo capitolo presentiamo i risultati ottenuti al variare di p_+ e della temperatura T[1]. Cominciamo col definire la grandezza:

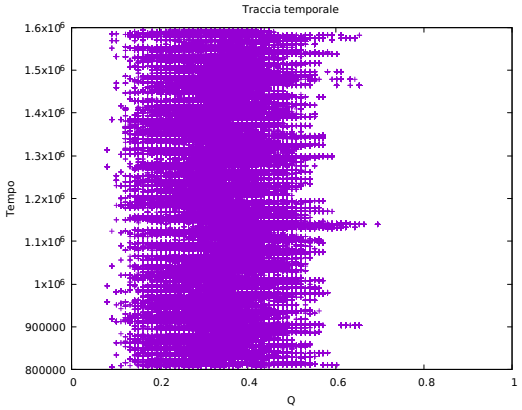
$$Q = \frac{\textit{Contatti nativi riprodotti}}{\textit{Contatti nativi totali}} \tag{4.1}$$

Ha senso definire tale grandezza solo quando la catena proteica è completa ovvero quando ha lunghezza massima. Notiamo inoltre che $Q \in [0, 1]$ e assume il valore 1 quando la proteina è nel suo stato nativo. Mostriamo ora i risultati ottenuti per le distribuzioni delle lunghezze e di Q fissando la temperatura a $T = 0.9$ e variando p_+:

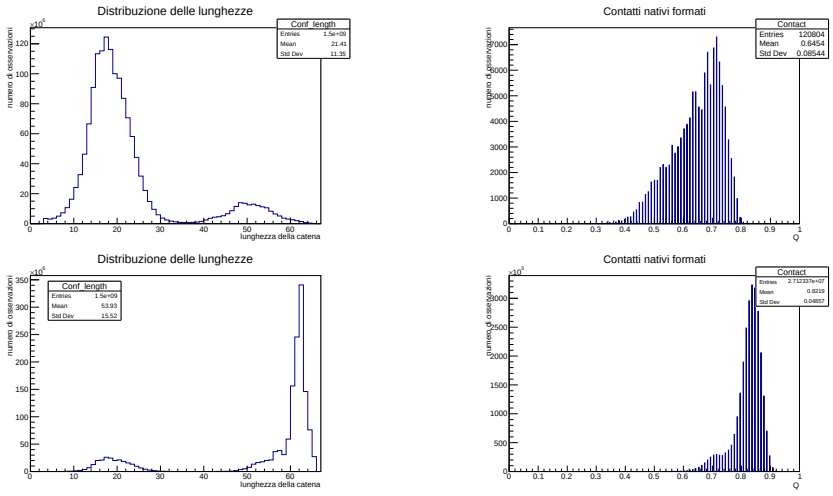


[1]Questa verrà misurata in unità nelle quali $K_B = 1$.

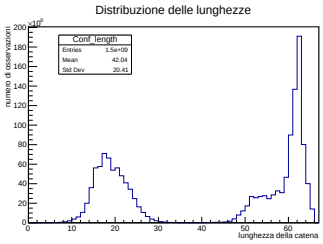
Nelle configuarzioni delle lunghezze notiamo la presenza di due massimi, esistono dunque due macrostati "privilegiati" rispetto agli altri. Notiamo inoltre che mentre le distribuzioni delle lunghezze variano al variare di p_+ quelle dei contatti rimangono all'incirca costanti. Mostriamo ora la *traccia temporale* di una delle simulazioni, questa rappresenta l'evoluzione della frazione di contatti nativi nel tempo. La traccia risulta stazionaria, ovvero si presenta sempre uguale in tutte le fasi della simulazione[2]:

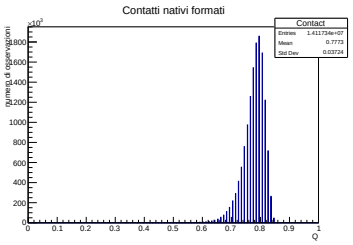


Il grafico mostra che il numero di contatti oscilla attorno a quello che dovrebbe essere il minimo corrispondendente allo stato denaturato. Ci aspettiamo che superato un certo valore di Q l'energia del sistema oltrepassi la barriera di energia libera situata tra lo stato denaturato e quello nativo, e dunque che il numero di contatti inizi ad oscillare intorno ad un nuovo minimo, quello in corrispondenza dello stato nativo. Questo comportamento è mostrato nell'articolo [9]. Tuttavia nel nostro caso il sistema non sembra in grado di superare tale barriera, ciò è dovuto al fatto che ci troviamo sopra la *temperatura di folding*.
Eseguendo nuovamente le simulazioni e abbassando la temperatura iniziano a sorgere problemi di campionamento, infatti la distribuzione di contatti nativi effettuati comincia a modificarsi al variare di p_+. Tale effetto è tanto più evidente quanto più è bassa la temperatura. Mostriamo i risultati ottenuti per T=0.6:

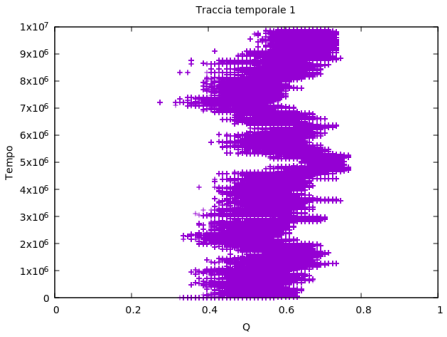


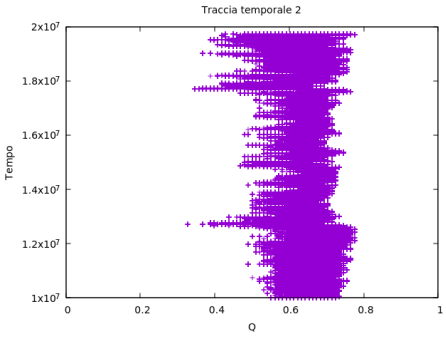
[2]La simulazione è stata divisa in 4 parti per esigenze tecniche, e tutte e quattro le tracce presentano la stessa forma. In grafico è mostrata la seconda.

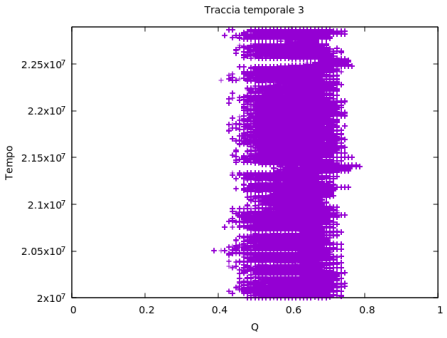


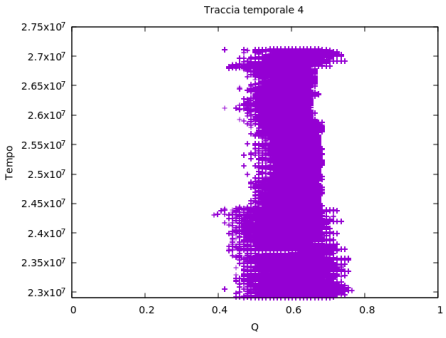


Andiamo ora ad analizzare la traccia temporale a questa temperatura, mostrando come si comporta in tutte e quattro le fasi della simulazione:









Notiamo che la traccia si presenta in maniera differente in ogni fase della simulazione, la distribuzione perde dunque di stazionarietà.

Capitolo 5

Conclusioni

Analizzando i risultati ottenuti si nota che il metodo funziona per $T = 0.9$, infatti la distribuzione del numero di contatti formati, Q, non varia al variare di p_+. Tuttavia in nessuna delle conformazioni ottenute si è riusciti a riprodurre tutti e 98 i contatti nativi (ovvero non ho nessuna configurazione con $Q = 1$), dunque la proteina non si piega completamente e non raggiunge lo stato nativo. Questo è dovuto al fatto che la temperatura utilizzata per la simulazione è superiore alla temperatura di folding. Andando però a diminuire la temperatura iniziano ad insorgere problemi di campionamento, infatti già a $T = 0.8$ la distribuzione di Q mostra una dipendenza dalla scelta di p_+. Inoltre continuando con la riduzione della temperatura tale dipendenza risulta ancor più evidente. I problemi di campionamento sono causati dall'eterogeneità introdotta dalla funzione energia (2.1). Infatti gli pseudo-atomi nello stato nativo presentano un diverso numero di contatti, questo può andare da 0 a 7. Ciò comporta che alcuni pseudo-atomi (quelli che presentano il maggior numero di contatti) contribuiranno maggiormente all'abbassamento di energia e saranno dunque privilegiati rispetto agli altri. Tale effetto viene enfatizzato con l'abbassarsi della temperatura che rende più difficile il superamento del test di Metropolis. È questa la causa dei problemi di campionamento a basse temperature. Il problema può essere risolto utilizzando una probabilità di aggiungere p_+ che dipenda dalla posizione in cui ci troviamo nella catena quando andiamo a modificarla. In tal modo potremo bilanciare la situazione per gli pseudo-atomi con più contatti e dunque ottimizzare la diffusione del Random Walk attraverso le varie lunghezze della catena, una tecnica del genere è stata eseguita da [10].
Alternativamente è possible utilizzare il metodo Wang-Landau. Questo rinuncia ad eseguire l'algoritmo di Metropolis a temperatura costante e va a stimare la densità degli stati tramite un Random Walk nello spazio delle energie. Tale densità verrebbe poi utilizzata per ricostruire tutta la termodinamica del sistema per una temperatura arbitraria [11].

Bibliografia

[1] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 1973.

[2] Alexei V. Finkelstein and Oleg B. Ptitsyn. *Protein Physics*. Academic Press, first edition, 2002.

[3] Cyrus Levinthal. How to fold graciously. *DeBrunner JTP*, 1969.

[4] Nobhuiro Gō and H. Abe. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers*, 1981.

[5] Sebastian Kmiecik, Michal Kolinski, Andrzej Kolinski, et al. Coarse-grained protein models and their applications. *Chemical Reviews*, **116**:7898 – 7936, 2016.

[6] Giovanni Settanni, Trinh Xuan Hoang, Cristian Micheletti, and Amos Maritan. Folding pathways of prion and doppel. *Biophysical Journal*, **83**:3533 – 3541, 2002.

[7] Benjamin A. Stickler and Ewald Schachinger. *Basic Concepts in Computational Physics*. Springer, second edition, 2015.

[8] Fabio Trovato and Edward P. O'Brien. Insights into cotranslational nascent protein behavior from computer simulations. *Annual Review of Biophysics*, **45**:345 – 369, 2016.

[9] Cecilia Clementi, Hugh Nymeyer, and José Nelson Onuchic. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of Molecular Biology*, **298**:937 – 953, 2000.

[10] Pengfei Tian, Sergei V. Krivov, Kresten Lindorff-Larsen, et al. Robust estimation of diffusion-optimized ensembles for enhanced sampling. *Journal of Chemical Theory and Computation*, **10**:543 – 553, 2014.

[11] Fugao Wang and David P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, **86**:2050 – 2053, 2001.