



UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Cinetica di ripiegamento su stati nativi topologicamente complessi

Relatore

Prof. Antonio Trovato

Laureando

Andrea Camerra

Anno Accademico 2019/2020

Indice

1	La struttura delle proteine	1
2	Il ripiegamento delle proteine	5
2.1	La denaturazione	5
2.2	Cinetica del ripiegamento	7
3	L'entanglement topologico	11
4	Modelli di ripiegamento	13
5	Simulazione del ripiegamento di tre proteine	16
6	Conclusioni	20
	Bibliografia	21

Capitolo 1

La struttura delle proteine

Le proteine sono macromolecole biologiche costituite da sequenze di amminoacidi, legati tra loro da un legame peptidico, che le distinguono l'una dall'altra. Ogni catena ripiega in una struttura tridimensionale specifica che determina la sua attività. Costituiscono una parte essenziale degli organismi viventi e partecipano a ogni processo che avviene all'interno delle cellule. Molte di loro sono enzimi, la cui funzione è catalizzare le reazioni chimiche, altre hanno funzioni strutturali o meccaniche, altre ancora sono fondamentali per la trasmissione di segnali inter- ed intracellulari, nella risposta immunitaria, e per il ciclo cellulare.

Le proteine sono strutture complesse organizzate gerarchicamente. Il primo livello di organizzazione è detto struttura primaria, la sequenza ordinata delle componenti fondamentali delle proteine, gli amminoacidi. La struttura secondaria è costituita dalle strutture più elementari in cui la catena si ripiega, ad esempio eliche e foglietti. L'intera catena, infine, assume una configurazione tridimensionale compatta, detta struttura terziaria. Le proteine più complesse possono avere una struttura quaternaria, composta da unità proteiche più piccole, dette domini, unite tra loro da legami deboli. La maggior parte delle proteine assume la forma di un globulo compatto.

Tutti gli amminoacidi sono costituiti da un atomo di carbonio centrale, detto C_α , circondato da quattro gruppi sostituenti: un atomo di idrogeno, un gruppo amminico NH_2 , un gruppo carbossilico $COOH$, e un quarto gruppo detto catena laterale, diverso per ogni amminoacido. Gli atomi non facenti parte della catena laterale vengono definiti anche catena principale, o *backbone*.

La diversità delle catene laterali dà a ciascuno degli amminoacidi, detti anche "residui" quando si trovano all'interno di una proteina, proprietà uniche. Il *backbone*, al contrario, è lo stesso per tutti gli amminoacidi, ed è proprio questa parte della molecola che li connette l'uno all'altro per formare la catena proteica. La connessione avviene tramite un legame peptidico, che si forma tramite una reazione di condensazione tra il gruppo carbossilico e il gruppo amminico di due amminoacidi. Durante la formazione di una nuova catena peptidica, gli amminoacidi sono aggiunti sequenzialmente al gruppo carbossilico finale, chiamato C-terminale, opposto all'N-terminale che è il gruppo amminico iniziale.

Il legame peptidico è planare e rigido, e per questo ha un ruolo importante nella struttura delle proteine. È planare a causa dell'ibridizzazione sp^2 degli atomi N e C', mentre è rigido per la presenza dei loro elettroni non ibridizzati, che assieme agli elettroni di O delocalizzano, formando una nuvola elettronica che ricopre tutti e tre gli atomi. L'angolo di rotazione di questo legame, ω (Figura 1.1), può vibrare solo per agitazione termica e quindi non contribuisce alla flessibilità della catena peptidica; inoltre la repulsione tra due atomi C_α consecutivi rende la conformazione *cis* (Figura 1.2) sfavorevole energeticamente, motivo per cui è raramente trovata. Neanche l'angolo χ , che descrive la rotazione della catena laterale, influisce sulla flessibilità della catena. Le rotazioni attorno agli angoli φ e ψ , invece, sono praticamente libere, e sono proprio queste a dare alla catena peptidica la sua flessibilità. φ e ψ non possono assumere qualsiasi valore, a causa dell'ingombro sterico degli altri atomi della catena principale e soprattutto delle catene laterali. Le conformazioni (φ, ψ) permesse e non sono rappresentate dai grafici di Ramachandran, che mostrano come il 75% dello spazio (φ, ψ) non sia accessibile per il solo ingombro del *backbone* [2].

Alcuni segmenti della catena peptidica tendono a ripiegarsi in strutture semplici, che si estendono lungo l'asse del segmento. Come detto in precedenza, questa è la struttura secondaria della proteina. Vi sono due strutture secondarie regolari, le α -eliche e le strutture- β . Considerate insieme, costituiscono circa la metà della catena di una proteina globulare [1].

Le α -eliche si formano tramite legami a idrogeno: un gruppo carbonile si lega a un gruppo ammidico posizionato più avanti (nella direzione del C-terminale) nella catena principale. In un'elica, i gruppi carbonile e ammidici sui residui estremali non sono coinvolti in legami a idrogeno, ed essendo gruppi polari generano un dipolo su tutta l'elica: questo ha un ruolo importante nell'attività biologica della proteina, trovandosi le α -eliche spesso nei siti di legame [2].

Altri segmenti della catena possono assumere una conformazione meno compatta, detta β , e ogni segmento viene detto filamento- β . In questa conformazione la catena non

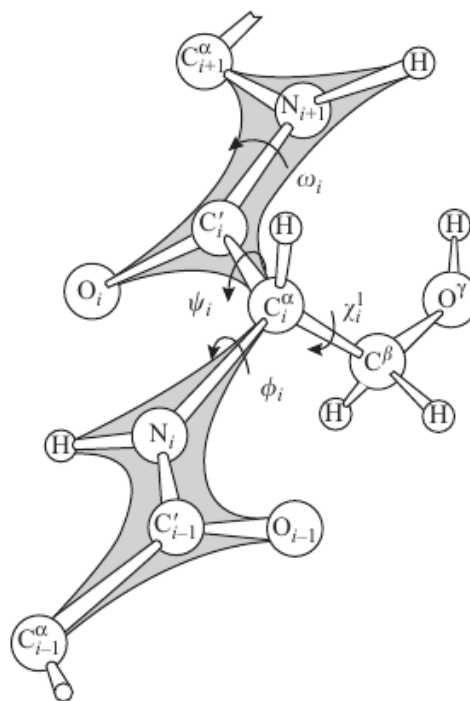


Figura 1.1: Rappresentazione di un legame peptidico, con evidenziati gli angoli di rotazione delle catene principale e laterale. Da [1, p. 17]

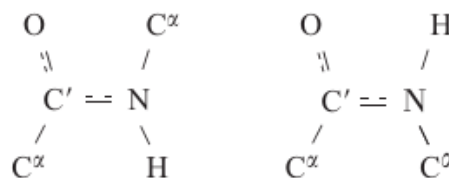


Figura 1.2: Configurazione *trans* (sinistra) e *cis* (destra) del gruppo peptidico. Da [1, p. 21]

è completamente estesa, ma si dispone a zig-zag. Un singolo filamento- β non consente, al suo interno, legami a idrogeno o di van der Waals intramolecolari. Per questo, essi sono raramente isolati nelle proteine, ma tendono a disporsi uno a fianco all'altro per formare i foglietti- β , una struttura piatta. I foglietti- β sono legati tra loro da ponti idrogeno come nelle α -eliche, ma a differenza di queste i gruppi legati sono distanti nella catena, poiché fanno parte di filamenti diversi. A seconda dell'orientamento relativo dei filamenti che lo compongono, un foglietto- β può essere parallelo, antiparallelo, o misto.

Le funzioni complesse delle proteine necessitano di un grado di organizzazione superiore alla struttura secondaria, che si concretizza con il ripiegamento dell'intera catena peptidica, dando origine alla sua struttura terziaria. In base alle strutture terziarie, le proteine si dividono in due famiglie: globulari e fibrose. Le prime assumono una forma sferica, svolgono molteplici funzioni biologiche, e possono essere idrosolubili o liposolubili, a seconda che si trovino nel citoplasma o nelle membrane cellulari; le seconde hanno una forma allungata, non sono solubili in acqua, e hanno funzioni meccaniche o di protezione dei tessuti. Le proteine meglio studiate e su cui sono note maggiori informazioni, come il processo di ripiegamento spontaneo o la struttura 3D, sono quelle globulari, su cui ci concentreremo in questa tesi.

In base alle caratteristiche della catena principale si possono distinguere le proteine globulari in pure α , pure β , e miste. I domini β sono più semplici degli altri: due o più foglietti- β composti da pezzi estesi della catena sono posizionati uno sopra l'altro (Figura 1.3a). La struttura antiparallela è predominante in questi domini. Le proteine α , invece, sono più difficili da categorizzare di quelle β . Il motivo è che mentre il posizionamento dei filamenti β è stabilizzato da legami a idrogeno della catena principale, che sono gli stessi ovunque, quello delle α -eliche dipende dall'incastro delle catene laterali, che hanno molta varietà. Le strutture più semplici sono composte da eliche disposte con gli assi quasi paralleli a formare un cilindro, e sono tipiche di eliche lunghe (Figura 1.3b). Tuttavia, nelle proteine globulari sono più comuni eliche corte, che tendono ad aggregarsi in una forma quasi sferica attorno al *core*, seguendo gli spigoli di un poliedro [1]. Infine, vi sono due tipi di domini misti: α/β e $\alpha+\beta$. Nei primi, le eliche e i filamenti si alternano lungo la catena, creando foglietti- β paralleli sovrapposti a uno strato di eliche. I secondi sono basati sulla struttura β antiparallela, e si possono dividere in due classi, a seconda che le regioni α e β si alternino regolarmente o meno (Figura 1.3c).

Le proteine globulari svolgono funzioni sofisticate, come la catalisi enzimatica, il trasporto di molecole, la trasduzione del segnale, e hanno ruoli di difesa e di regolazione. A questo fine devono essere compatte e in grado di formare un sito attivo e di legame. Queste proprietà sono ottenibili grazie all'abilità della catena peptidica di cambiare direzione, permettendo a residui distanti nella catena di trovarsi vicini nello spazio e interagire. Questo porta la catena peptidica a formare una struttura complessa, la cui stabilità è determinata da una grande varietà di interazioni, la maggior parte delle quali è di natura non covalente. La disposizione dei residui all'interno del globulo dipende principalmente dalla loro polarità. I residui con gruppi polari si trovano solitamente sulla superficie della proteina, a contatto con l'acqua, perché la rottura dei ponti idrogeno con l'acqua e l'immersione dei residui nel *core* della proteina corrisponderebbe a un

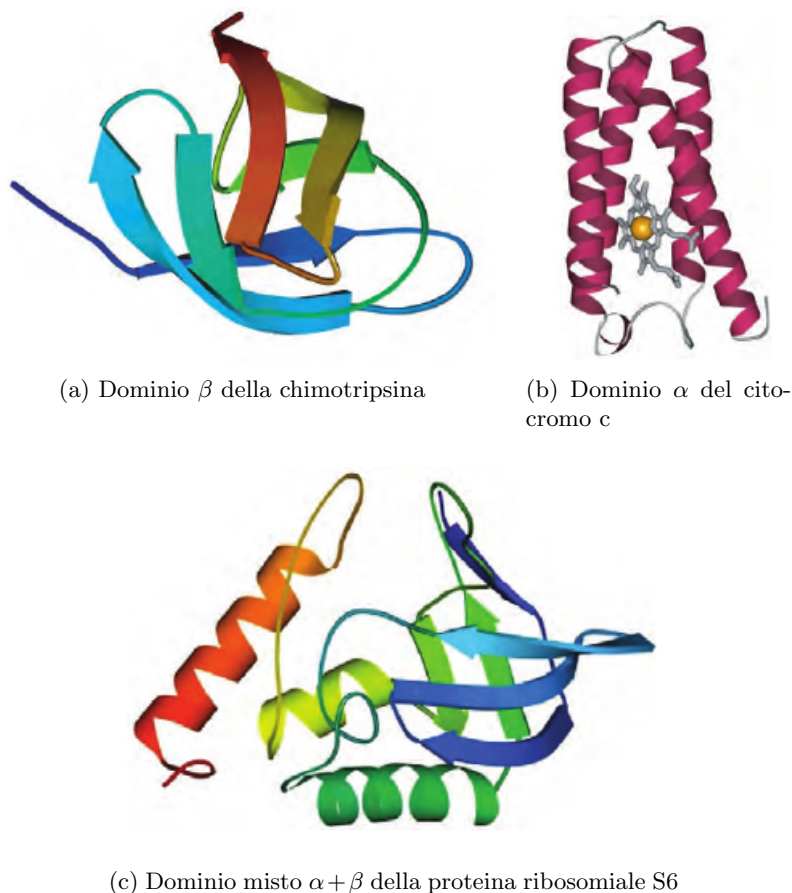


Figura 1.3: Alcuni esempi di strutture terziarie. Da [1]

grande aumento dell'energia libera. Al contrario, i residui con catene laterali idrofobiche tendono a disporsi all'interno della proteina, a causa della repulsione dell'acqua. Questo è l'effetto idrofobico, che è molto importante per la stabilità della struttura della proteina. Esso è infatti responsabile di circa il 90% dello sforzo necessario a formare un globulo compatto [1]. Tuttavia, esso non è sufficiente a stabilizzare la proteina: a questo servono le forze di van der Waals, interazioni deboli e a corto raggio, ma fondamentali in molecole complesse, e i legami a idrogeno, che si formano tra i gruppi polari del *backbone* per stabilizzare le strutture secondarie.

Capitolo 2

Il ripiegamento delle proteine

Inizialmente, si pensava che le proteine fossero oggetti rigidi, di forma complementare al substrato cui si legano. Questa visione cambiò quando David Koshland ipotizzò, nel 1958, che un enzima potesse subire cambiamenti strutturali per adattarsi meglio al substrato. Da allora, vari studi hanno dimostrato che nelle proteine avvengono vari cambiamenti strutturali. La dinamica delle proteine è caratterizzata da un insieme variegato di movimenti di atomi e gruppi, che avvengono su scale temporali tipiche di processi biochimici. Il fatto che queste scale temporali siano confrontabili è un segno che la dinamica delle proteine è importante per il loro funzionamento.

I tipi di movimento che si trovano nelle proteine sono vari, e includono rotazioni su perno, rototraslazioni, e anche il ripiegamento o disfacimento di strutture secondarie e supersecondarie. I perni sono molto comuni nelle proteine, e si è notato che vengono conservati nei processi evolutivisti, confermando la loro importanza funzionale.

La cinetica del ripiegamento non è necessariamente correlata alla struttura proteica; tuttavia, la comprensione di questo argomento può fornire una conoscenza più approfondita del comportamento delle proteine, migliorare gli algoritmi che ne predicono la struttura, e aiutare a comprendere le patologie causate da proteine anomale.

2.1 La denaturazione

Il processo inverso del ripiegamento di una proteina è la denaturazione, che è l'oggetto di numerosi studi, vista la sua relazione con il problema fisico del ripiegamento. La denaturazione di piccole proteine è una transizione con un cambiamento improvviso e simultaneo di molte caratteristiche della molecola, rappresentato dalla forma a S delle curve sperimentali (Figura 2.1). La denaturazione avviene cioè come una reazione “tutto o niente”: solo le fasi iniziale (nativa) e finale (denaturata) sono presenti in quantità osservabili, mentre gli stati intermedi sono pressoché assenti. Questo tipo di transizione è un analogo microscopico delle transizioni di fase di primo ordine nei sistemi macroscopici, ad esempio lo scioglimento dei cristalli.

Questo processo è reversibile: se la proteina non è troppo grande e non ha subito sostanziali modifiche chimiche dopo il ripiegamento *in vivo*, può rinaturarsi. Questo

dimostra che tutte le informazioni necessarie a costruire la struttura della proteina sono contenute nella catena peptidica, e che la struttura stessa è termodinamicamente stabile.

Oltre all'attività proteica, solo due proprietà cambiano rapidamente durante la denaturazione: la disposizione delle catene laterali, e la rigidità della struttura globulare. Al contrario, a seconda delle condizioni della denaturazione, la struttura secondaria, il volume, e la densità del *core* idrofobico possono conservarsi [1, p. 215]. Uno stadio intermedio, quasi universale, della denaturazione (e quindi del ripiegamento) delle proteine, è il cosiddetto *molten globule*. Ha una struttura secondaria pressoché nativa, ma i gruppi laterali non sono ordinati, sebbene alcuni contatti nativi tra di loro rimangano intatti. È meno denso dello stato nativo, tranne che nel *core* che ha pressoché la stessa densità di quello nativo; tuttavia, alcune molecole di solvente penetrano facilmente nel globulo.

Per molte proteine il *molten globule* si forma in seguito a una denaturazione causata da un aumento di temperatura, e può decadere nella catena sciolta sotto l'azione di un denaturante concentrato. Alcune proteine, specialmente quelle piccole, denaturano direttamente in una catena libera senza lo stadio intermedio del globulo fuso.

Per comprendere la denaturazione delle proteine, bisogna spiegare perché esistono due stati ugualmente stabili (stato nativo e catena sciolta) e perché questi sono separati da una barriera di energia libera (necessaria per una transizione “tutto o niente”). Per farlo, serve prendere in considerazione le particolarità delle proteine. I gruppi laterali sono capaci di isomerizzazione conformazionale, cioè di cambiare conformazione senza rompere legami chimici. Ogni cambiamento della conformazione necessita di volume libero attorno alla catena laterale che si sposta, ma la struttura nativa è estremamente compatta, e il *backbone* è particolarmente rigido, perché qui la catena forma le strutture α e β . Affinché il globulo si espanda, queste strutture rigide devono spostarsi interamente, con tutti i loro gruppi laterali, e facendolo liberano degli spazi che o sono insufficienti per l'isomerizzazione di ciascuna catena laterale, o sono sufficienti per l'isomerizzazione di molte contemporaneamente. Perciò la liberazione delle catene laterali avviene solo quando l'espansione del globulo supera una certa soglia. Una piccola espansione del globulo è quindi sempre sfavorevole: aumenta l'energia, visto che le sue parti si allontanano, ma non aumenta l'entropia, perché non libera le catene laterali. Un'espansione maggiore del globulo, invece, porta a una diminuzione dell'energia libera. Di conseguenza, la denaturazione delle proteine non avviene gradualmente, ma con un “salto” oltre la barriera energetica, in accordo con il principio del “tutto o niente” (Figura 2.2).

Prima della scoperta del *molten globule* si credeva che la denaturazione delle proteine fosse un completo decadimento della struttura proteica, cioè una transizione alla catena

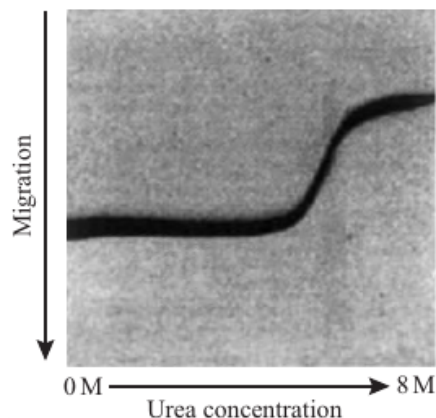


Figura 2.1: Elettroforesi di citocromo c a varie concentrazioni di urea. Da [1, p. 208]

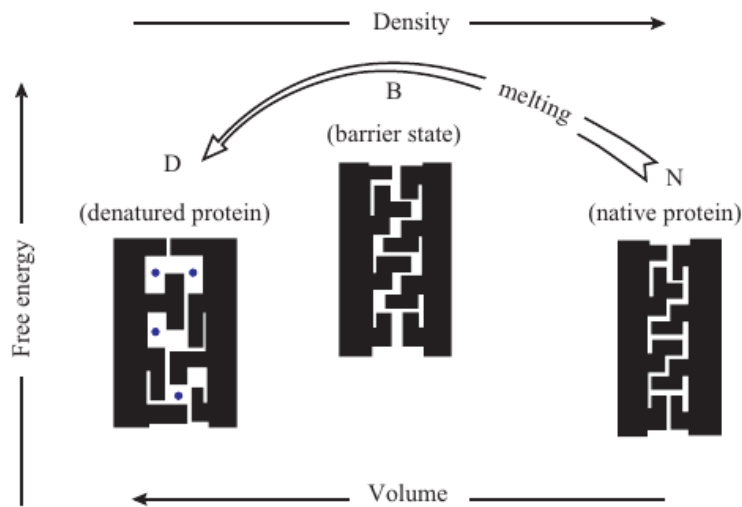


Figura 2.2: Barriera energetica tra stato nativo e stato denaturato. Da [1, p. 229]

sciolta. È divenuto chiaro che una proteina denaturata può essere più o meno densa a seconda della forza del solvente e dell'idrofobicità della catena: più il *core* è idrofilo, più facilmente attrae le molecole d'acqua, che infiltrandosi nella proteina ne causano il rigonfiamento e infine il decadimento.

2.2 Cinetica del ripiegamento

Nelle cellule, le proteine sono sintetizzate dai ribosomi. La biosintesi impiega circa un minuto e la produzione di una proteina ripiegata dura altrettanto: è quindi ragionevole pensare che il ripiegamento di una proteina cominci sul ribosoma, prima che la sintesi della catena sia completa. Infatti numerosi esperimenti su proteine multi-dominio hanno mostrato che i loro domini N-terminali sono in grado di ripiegarsi prima che l'intera catena sia completa. Negli anni '60, fu scoperto che una proteina globulare è capace di renaturazione spontanea (*in vitro*) se la sua catena non è stata modificata chimicamente dopo il ripiegamento iniziale (*in vivo*). Il fenomeno del ripiegamento spontaneo delle strutture native permette di separare, in prima approssimazione, lo studio del *folding* da quello della biosintesi delle proteine.

Il problema fisico fondamentale del ripiegamento delle proteine è noto come paradosso di Levinthal. Da un lato, si ottiene la stessa struttura nativa tramite diversi processi di ripiegamento, e l'esistenza stessa di ripiegamento spontaneo suggerisce che la struttura nativa è la più stabile in condizioni biologiche. Dall'altro, una proteina di 100 residui ha almeno 2^{100} conformazioni a disposizione (visto che ogni residuo ha almeno due conformazioni possibili), e la proteina "percepisce" la struttura stabile corretta solo se ha raggiunto quella precisa conformazione, visto che anche una deviazione di 1 Å può aumentare l'energia della catena. Assumendo che ogni conformazione possa essere campionata in 1 ps, la catena avrebbe bisogno di almeno 2^{100} ps, cioè 10^{10} anni, per testarle tutte. Il paradosso è che il raggiungimento dello stesso stato (nativo) da parte di vari

processi è chiaro indice della sua stabilità; d'altra parte, i calcoli di Levinthal mostrano che una proteina non avrebbe abbastanza tempo per verificare che la struttura nativa è la più stabile tra tutte quelle possibili. Levinthal ipotizzò l'esistenza di un percorso di ripiegamento, e che la struttura nativa sia semplicemente il traguardo di questo percorso, invece che la struttura più stabile. Se questo percorso fosse stretto, solo una piccola porzione dello spazio conformazionale dovrebbe essere esplorata, e il paradosso sarebbe evitato. In altre parole, Levinthal suggerì che la struttura nativa delle proteine è il risultato di un processo cinetico anziché termodinamico, cioè che corrisponde non al minimo globale, ma al minimo locale di energia libera più facilmente accessibile.

L'ipotesi del "controllo cinetico" ha portato allo studio molto intenso degli stati intermedi del ripiegamento. L'idea è che gli intermedi, se isolati, permetterebbero di tracciare il percorso del ripiegamento. Questi intermedi sono stati trovati in molte proteine, ma la domanda di come la catena possa trovare la struttura nativa tra tutte le alternative rimane. Lo studio di proteine piccole, di 50-100 residui, ha portato a dei progressi. Molte di loro ripiegano *in vitro* senza stadi intermedi osservabili, e perciò questi esperimenti sono la migliore opportunità per studiare lo stato di transizione, il collo di bottiglia del ripiegamento. Lo stato di transizione corrisponde a un massimo locale di energia, al contrario del *molten globule* che è un minimo locale, quindi non si accumula e non viene mai osservato direttamente, ma può essere seguito solo tramite la sua influenza sulla velocità di ripiegamento.

La cinetica della transizione tra due stati stabili è descritta dalla teoria dello stato di transizione. L'essenza di questa teoria è che la velocità del processo è limitata dalla bassa occupazione dello stato di transizione, cioè le transizioni tra stato nativo e denaturato sono determinate dall'altezza del massimo di energia libera lungo il percorso di ripiegamento, secondo le relazioni

$$\begin{aligned} k_{A \rightarrow B} &= k_0 \exp(-(F^\# - F_A)/RT) \\ k_{B \rightarrow A} &= k_0 \exp(-(F^\# - F_B)/RT) \end{aligned} \quad (2.1)$$

con i valori delle F rappresentati dalla Figura 2.3, e k_0 il *rate* di un passo elementare. Il rapporto tra i *rate* delle reazioni diretta e inversa è la costante di equilibrio della reazione.

Il tasso di avvicinamento all'equilibrio k_{app} è la somma dei *rate* delle equazioni (2.1) [1, pp. 255-256]. La più rapida delle due reazioni domina il valore di k_{app} , i cui valori sono solitamente presentati in funzione della concentrazione di denaturante, assumendo una forma a V (Figura 2.4). I *rate* delle due reazioni dipendono in maniera opposta dalla concentrazione di denaturante. Al punto di metà transizione, la proteina ripiega e denatura nelle stesse quantità. A basse concentrazioni, invece, il tasso di ripiegamento ha il suo massimo, ed è pressoché indipendente dalla concentrazione: il tempo di ripiegamento in acqua è quindi preso come tempo caratteristico di ripiegamento di una proteina.

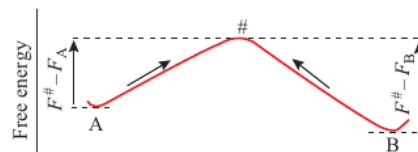


Figura 2.3: Barriera di energia per transizione tra due stati A e B. Da [1, p. 252]

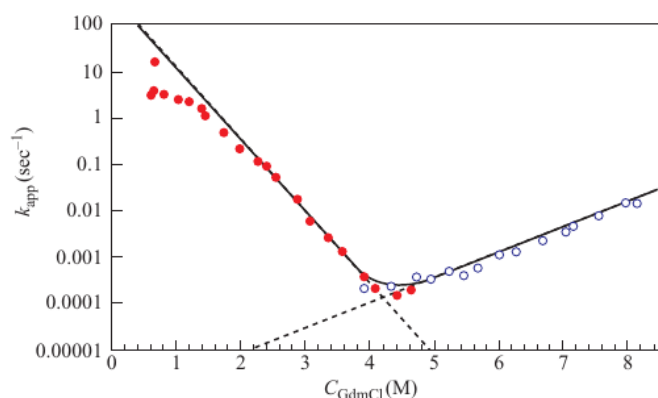


Figura 2.4: Frequenza di avvicinamento all'equilibrio in funzione della concentrazione di denaturante. Da [1, p. 257]

Lo stato di transizione è meno uniforme del globulo fuso. Per la transizione da catena sciolta a struttura nativa, lo stato di transizione si può immaginare come un pezzo di proteina nativa, con il resto della catena non ancora ripiegata. La stabilità della struttura nativa della proteina rispetto alla catena non ripiegata è determinata dalla differenza di energia libera tra i due stati, il cui valore dipende dalle condizioni ambientali.

Le mutazioni della catena peptidica influenzano sia il tasso

di ripiegamento che la stabilità della proteina, quindi possono essere utilizzate per determinare sperimentalmente il nucleo di ripiegamento, cioè la parte globulare dello stato di transizione. Il meccanismo di nucleazione è tipico delle transizioni di fase di primo ordine della fisica convenzionale, e quindi è plausibile che avvenga anche nel ripiegamento delle proteine. Se una mutazione modifica le differenze di energia tra stato denaturato e stato di transizione (da cui dipende il tasso di ripiegamento), e tra stato nativo e denaturato (indice della stabilità della proteina) di quantità simili, allora il residuo mutato fa parte del nucleo. Se il tasso non cambia, il residuo non fa parte del nucleo. Solo una porzione trascurabile di residui si comporta in maniera che non può essere spiegata da tale meccanismo: questo permette di considerare il meccanismo di nucleazione corretto.

Per dimostrare che la struttura più stabile è in grado di ripiegare rapidamente, è sufficiente dimostrare che almeno un percorso di ripiegamento rapido porta a questa struttura. Ulteriori percorsi possono solo accelerare il ripiegamento visto che reazioni parallele sono additive. Per essere rapido, un percorso deve consistere di pochi passi, e soprattutto non deve scavalcare una barriera di energia troppo alta. Se i contatti che stabilizzano la struttura iniziano a formarsi solo quando la catena è vicina al suo stato nativo, perdendo molta entropia prima di perdere energia, l'aumento iniziale di energia libera costituirebbe una barriera molto alta. Il paradosso di Levinthal ha origine proprio da questa situazione, ma può essere evitato se esiste un percorso in cui la perdita di entropia è rapidamente compensata da una perdita di energia.

Un percorso di ripiegamento molto semplice è quello sequenziale, in cui a ogni passo un residuo assume la sua posizione finale nella struttura nativa. Sembra artificiale, ma è ciò che si osserverebbe guardando un processo di denaturazione al contrario, e la proteina non può usare percorsi diversi per le reazioni opposte per il principio del bilancio dettagliato, secondo cui le reazioni diretta e inversa devono seguire lo stesso percorso a parità di condizioni.

Quando un pezzo del globulo cresce sequenzialmente, anche le interazioni stabilizzanti

vengono ripristinate sequenzialmente. Se questa parte della proteina rimane compatta, il numero di interazioni native cresce circa linearmente con il numero di residui nelle loro posizioni native. All'inizio del ripiegamento la diminuzione di energia è più lenta, dato che il contatto di un residuo con la superficie di un globulo piccolo è mediamente minore che con un globulo più grande. Questo risulta in un termine di superficie (proporzionale a $N^{2/3}$) nell'energia del globulo. L'entropia ha un comportamento simile, a causa di loop chiusi ma disordinati che protrudono dal globulo. Quando il globulo è in equilibrio con la catena libera, i termini lineari si annullano a vicenda e rimangono solo quelli di superficie. La barriera di energia libera, quindi, dipende solo dagli effetti di superficie al confine tra globulo e catena, ed è proporzionale non a N , come implicherebbe la stima di Levinthal, ma a $N^{2/3}$. Il tempo di ripiegamento, quindi, non cresce con il numero di residui come 2^N , ma solo come $\exp(\lambda N^{2/3})$. λ è stimato valere 1 ± 0.5 , con il valore preciso che dipende dalla distribuzione di legami forti e deboli e dalla topologia della struttura di energia minima [1, pp. 267-269].

Questa stima mostra che una catena di 100 residui trova la sua struttura più stabile in alcuni minuti, e spiega anche perché proteine più grandi devono essere formate da domini separati: catene di più di 300 residui ripiegherebbero troppo lentamente.

Trovando l'energia libera dello stato di transizione si possono stimare le dimensioni della sua parte globulare. Da questa stima si vede che il nucleo dev'essere grande almeno la metà della proteina: un nucleo di queste dimensioni non lascia spazio a molti percorsi di ripiegamento, per cui considerare un percorso soltanto può dare una stima accurata del tempo di ripiegamento [1, p. 270].

Capitolo 3

L'entanglement topologico

Nonostante già nel 1977 fosse conosciuto un nodo nell'anidra carbonica B, per molto tempo fu considerato impossibile che una catena peptidica potesse annodarsi e formare una proteina funzionante, e fu solo nel 2000 che un'analisi approfondita del Protein Data Bank (PDB) dimostrò inequivocabilmente l'esistenza di proteine con nodi, anche profondi, che ammontano all'1-3% delle proteine registrate [3, 4]. Esse appartengono a una grande varietà di famiglie funzionali, perciò si ipotizza che i nodi possano contribuire a migliorare l'attività enzimatica o ad aumentare la stabilità delle proteine.

Da allora, numerosi studi hanno stabilito che alcune proprietà del ripiegamento si possono ricavare a partire dalle caratteristiche dello stato nativo: ad esempio, conoscendo quali coppie di residui sono in contatto tra loro, si può ricostruire il nucleo di ripiegamento della proteina [5]. Tuttavia per le proteine dalla struttura più complessa, ad esempio quelle che contengono nodi (normali e scorsoi), lacci, e anelli (presenti complessivamente nel 6% delle proteine nel PDB), lo studio delle proprietà locali non è sufficiente a spiegare il loro ripiegamento. In questo caso è più utile studiare aspetti più generali dello stato nativo, come ad esempio il suo grado di *entanglement* topologico. Questa grandezza descrive l'avvolgimento reciproco che sussiste tra due curve o tra due segmenti della stessa curva.

Per calcolare questa grandezza si usa l'*entanglement* gaussiano, una generalizzazione dell'integrale di Gauss usato per calcolare il *linking number* [5]. Il *linking number* tra due curve chiuse $\gamma_i = \{\mathbf{r}^{(i)}\}$ e $\gamma_j = \{\mathbf{r}^{(j)}\}$ in $\mathbb{R}^3$ vale

$$G = \frac{1}{4\pi} \oint_{\gamma_i} \oint_{\gamma_j} \frac{\mathbf{r}^{(i)} - \mathbf{r}^{(j)}}{|\mathbf{r}^{(i)} - \mathbf{r}^{(j)}|^3} \cdot (d\mathbf{r}^{(i)} \times d\mathbf{r}^{(j)}), \quad (3.1)$$

è un numero intero e un invariante topologico, e descrive quante volte una curva è concatenata all'altra. Se calcolato su due curve aperte, si ottiene un numero reale G' , chiamato appunto *entanglement* gaussiano, che quantifica il mutuo avvolgimento tra le due curve.

Nelle proteine, le curve in questione sono linee spezzate che uniscono gli atomi di C_α consecutivi. In particolare, γ_i unisce i C_α che vanno dall'indice i_1 all'indice i_2 , γ_j da j_1 a j_2 , e le due porzioni di catena non devono avere sovrapposizioni. Visto che i legami

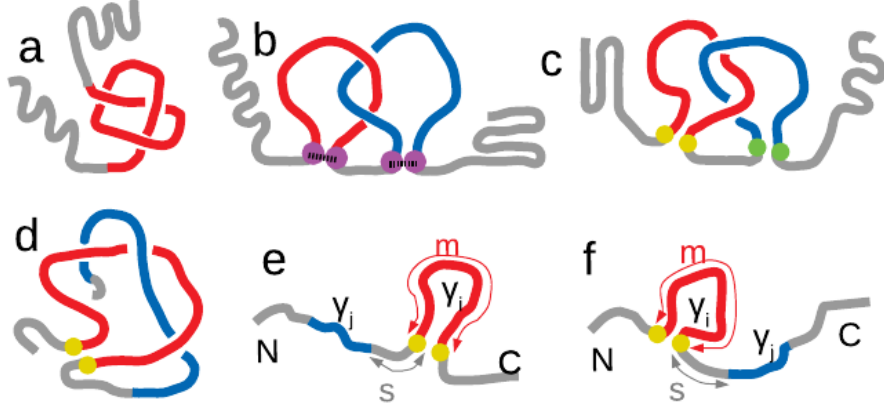


Figura 3.1: Schemi di varie strutture *entangled*: un nodo (a), anelli concatenati chiusi da un legame covalente (b) e non covalente (c), un anello attraversato da una porzione di catena (d), che può trovarsi a monte (e) o a valle (f) dell'anello. Da [5].

che uniscono i C_α sono abbastanza distanti tra loro, si può approssimare l'integrale a una somma discreta. Poste $\mathbf{r}_i$ le coordinate dei C_α , allora $\mathbf{R}_i = \frac{1}{2}(\mathbf{r}_i + \mathbf{r}_{i+1})$ sono le posizioni medie dei legami, $\Delta\mathbf{R}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$ sono i vettori dei legami, e l'equazione (3.1) diventa

$$G = \frac{1}{4\pi} \sum_{i=i_1}^{i_2-1} \sum_{j=j_1}^{j_2-1} \frac{\mathbf{R}_i - \mathbf{R}_j}{|\mathbf{R}_i - \mathbf{R}_j|^3} \cdot (\Delta\mathbf{R}_i \times \Delta\mathbf{R}_j). \quad (3.2)$$

Per limitare i calcoli alle configurazioni illustrate in Figura 3.1, si impone che i residui di indici i_1 e i_2 siano in contatto tra loro. L'*entanglement* gaussiano “di contatto” di queste configurazioni è detto $G'_c(i, j)$. A ogni anello γ_i è associato $G'_c(i)$, che è il valore estremaale tra i $G'_c(i, j)$, con γ_j lunga almeno 10 residui. Infine, l'*entanglement* di contatto dell'intera proteina è il valore estremaale tra i $G'_c(i)$.

L'analisi di un ampio campione di domini proteici effettuata in [5] ha evidenziato che le strutture mostrate nelle Figure 3.1c-d sono abbastanza comuni, essendo presenti circa un terzo delle 17000 proteine considerate. È stato anche rilevato che gli anelli *entangled* si trovano in posizione successiva lungo la sequenza rispetto al segmento che li attraversa (Figura 3.1e) più frequentemente che il contrario (55% dei casi): infatti, sembra più efficiente dal punto di vista cinetico avvolgere l'anello attorno a una porzione di catena già correttamente ripiegata nella struttura nativa piuttosto che il contrario, in particolare nel contesto del ripiegamento cotraslazionale. Infine, i residui che formano i contatti che chiudono gli anelli sono in larga parte più idrofili, rispetto ai contatti intramolecolari generici che hanno luogo generalmente tra amminoacidi idrofobici. Questo rende l'interazione fra le estremità di un anello *entangled* più debole della media. Un possibile motivo per cui questi contatti sono stati selezionati dall'evoluzione è che un'interazione più debole è necessaria a evitare trappole cinetiche, ritardando la formazione degli anelli *entangled* agli ultimi passi del ripiegamento [5].

Capitolo 4

Modelli di ripiegamento

Le transizioni di fase si osservano in molti ambiti sperimentali, quindi è logico pensare di poter trovare analogie tra le proteine e altri sistemi fisici, ad esempio i vetri di spin. Questi sono descritti dal modello di Ising, che rappresenta un materiale ferromagnetico come un reticolo discreto, in cui ogni sito ha un valore di spin di $+1$ o -1 . L'energia del sistema è data da $H = - \sum_{\{ij\}} J_{ij} s_i s_j$, dove (s_i, s_j) sono coppie di spin adiacenti, e J_{ij} la loro interazione magnetica. Se $J_{ij} > 0$, il sistema è definito “non frustrato”, poiché tutte le interazioni possono essere ottimizzate prendendo tutti gli spin paralleli tra loro. In condizioni fisiologiche le proteine ripiegano spontaneamente e rapidamente e, in maniera simile, nei sistemi di Ising le transizioni da paramagnetico a ferromagnetico avvengono senza trappole cinetiche se la dispersione delle interazioni J_{ij} è bassa. Coerentemente con ciò, lo stato (nativo o non) delle proteine è determinato dal rapporto tra temperatura e dispersione delle interazioni [6]. Queste somiglianze hanno portato alla formulazione del principio di minima frustrazione. Nelle proteine, la frustrazione si manifesta in interazioni stabilizzanti non native.

Negli ultimi decenni, molti studi teorici e sperimentali hanno studiato il problema del ripiegamento delle proteine. I primi modelli di simulazione rappresentavano le catene polimeriche come un reticolo (come i vetri di spin) per esplorare la relazione tra cinetica, frustrazione energetica, e meccanismi strutturali. In questi modelli, il polimero è trattato come una catena, in cui ogni amminoacido è una biglia che si sposta attraverso un insieme discreto di posizioni sul reticolo. Il loro basso costo computazionale ha permesso uno studio esteso della relazione tra stabilità energetica, contributi entropici, e frustrazione energetica nel ripiegamento, e l'elaborazione di nuove strategie per il disegno di proteine [6].

Le proteine, però, non risiedono su un reticolo, ma possono assumere coordinate continue in tre dimensioni. Per permettere una descrizione continua dello spazio delle fasi, sono stati elaborati modelli fuori reticolo, che permettono di studiare proteine specifiche e non solo dei sistemi modello. I primi modelli fuori reticolo rappresentavano ogni residuo come una sfera, centrata sul suo C_α (e sono quindi chiamati modelli C_α). Il principio di minima frustrazione dice che la differenza di energia tra stato nativo e non nativo deve essere grande: basandosi su ciò, questi modelli usano delle funzioni energia

costruite a partire dalla struttura nativa, in cui hanno il loro minimo globale, e sono per questo detti *structure-based models* (SBM). Gli SBM hanno dimostrato che la connettività della catena e il volume escluso restringono significativamente i possibili percorsi di ripiegamento dei biopolimeri [6]. Questa restrizione dello spazio delle fasi può portare a una situazione chiamata “frustrazione topologica”, che può essere caratterizzata come la necessità che alcune subunità si ripieghino per ultime, indipendentemente da quanta stabilità aggiungono, perché la formazione precoce di interazioni specifiche può far insorgere barriere energetiche che rallentano o impediscono il ripiegamento.

Un’evoluzione naturale degli SBM a grana grossa è la loro estensione a una rappresentazione più fine, in cui ogni atomo pesante (e in alcuni casi anche l’idrogeno) è considerato esplicitamente. Come nei modelli C_α , ogni interazione presente nella struttura nativa è stabilizzante. Questo modello ha due vantaggi principali: il volume escluso della struttura atomica è considerato esplicitamente, e si possono includere interazioni non specifiche (cioè non formate nella configurazione nativa), usando l’SBM nativo come base.

Il modello utilizzato in questa tesi è il Wako-Saitô-Muñoz-Eaton (WSME). Questo è un modello molto semplificato su un reticolo unidimensionale, in pratica un modello di spin con interazioni a lungo raggio, che però sono calcolate utilizzando le coordinate tridimensionali degli atomi pesanti nella struttura nativa. Una versione omogenea di questo modello fu introdotta nel 1978 da Wako e Saitô, che ne risolsero la termodinamica dell’equilibrio. Il caso eterogeneo fu considerato successivamente da Muñoz e colleghi [7], che considerarono solo configurazioni con un numero limitato di segmenti nativi durante il ripiegamento, sia per l’equilibrio che per la cinetica. In [7], il modello fu usato per misurare l’effetto di una mutazione puntiforme sul valore di $\varphi = \Delta \log k_f / \Delta \log K$, con k_f il tasso di ripiegamento (Eq. 2.1) e K la costante di equilibrio. Un valore di φ pari a 1.0 indica che l’ambiente strutturale attorno al residuo è identico a quello dello stato nativo, uno pari a 0.0 che è identico a quello dello stato denaturato. Nonostante la semplicità della funzione utilizzata come energia libera, fu trovato un buon accordo tra i valori teorici e sperimentali di φ .

Nel 2002, Pelizzola [8] ne formulò una versione non approssimata, calcolando una soluzione esatta per l’energia libera e qualunque altra quantità termodinamica rilevante, per qualsiasi proteina. Poiché il fine ultimo di questa classe di modelli è lo studio della cinetica, assumendo che questa sia determinata principalmente dalla struttura nativa, ha senso chiedersi se la soluzione esatta all’equilibrio può essere estesa alla cinetica. Nel caso generale, non si può determinare una soluzione esatta per la cinetica; tuttavia, si può escogitare un approccio di equilibrio locale che produce lo stato di equilibrio esatto nel limite di tempo infinito [9].

Il modello WSME descrive una proteina di $N + 1$ amminoacidi come una catena di N legami peptidici. La variabile $m_i \in \{0, 1\}$ è associata all’ i -esimo legame peptidico, e può essere pensata come una discretizzazione molto grossolana degli angoli diedri: $m_i = 1$ rappresenta la configurazione in cui gli angoli dell’ i -esimo legame peptidico sono nella conformazione nativa, $m_i = 0$ tutte le altre configurazioni. Due residui possono interagire solo se sono in contatto nella struttura nativa e se tutti i legami peptidici

compresi tra loro nella catena sono nella conformazione nativa. L'energia libera efficace (impropriamente anche detta Hamiltoniana) di una proteina con N legami può essere scritta come

$$H(m) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \varepsilon_{ij} \Delta_{ij} \prod_{k=i}^j m_k + k_B T \sum_{i=1}^N q_i m_i, \quad (4.1)$$

con k_B la costante di Boltzmann e T la temperatura assoluta. Il secondo termine della formula rappresenta il costo entropico di ordinare l'unità peptidica i . In [10] e in questa tesi, si pone $q_i = 1$. Il primo termine di (4.1) assegna un'energia ai contatti tra i legami peptidici i e j se questi sono in contatto nella struttura nativa ($\Delta_{ij} = 1$ in questo caso, 0 altrimenti), cioè se almeno una coppia di atomi pesanti, uno dal residuo i e uno dal residuo $j + 1$, sono più vicini di $4.5 \text{ \AA}^1$. I contatti con $j = i + 1$ (cioè tra i residui i -esimo e $(i + 2)$ -esimo) sono esclusi, perché sono presenti con probabilità alta anche nello stato denaturato. L'energia ε_{ij} è pari a $-\varepsilon n_{ij}$, proporzionale al numero n_{ij} di atomi pesanti a contatto tra loro. Infine, si pone $\lambda = \varepsilon/k_B T$. Questo parametro tiene conto delle condizioni ambientali di temperatura e concentrazione di denaturante (tramite ε).

All'inizio di una simulazione di ripiegamento, la proteina è completamente denaturata, cioè $m_i = 0$ per ogni i . Il sistema viene poi fatto evolvere tramite una dinamica di Metropolis, con cambiamento di una singola variabile per ogni passo. Ad ogni passo della simulazione viene selezionato un legame i campionando i diversi legami in maniera uniforme, e viene calcolata l'energia che il sistema avrebbe se l' m_i corrispondente cambiasse valore. La probabilità che il sistema evolva dallo stato iniziale x allo stato di prova x' è

$$W(x \rightarrow x') = \min \left\{ 1, \exp \left(\frac{H(x) - H(x')}{k_B T} \right) \right\}. \quad (4.2)$$

Se lo stato di prova non viene selezionato, il sistema rimane nello stato x anche al passo successivo.

¹In [8–10] si usano $4 \text{ \AA}$.

Capitolo 5

Simulazione del ripiegamento di tre proteine

Si è interessati a verificare l'ipotesi, formulata in [5], che i contatti che chiudono gli anelli *entangled* si formino negli ultimi stadi del ripiegamento delle proteine. A questo fine si è simulato il ripiegamento di tre proteine con diverso grado di entanglement:

- 1r6j, formata da 82 residui, con $|G'_c| = 0.395$, non presenta *entanglement*;
- 2hqx, formata da 81 residui, con $|G'_c| = 1.241$, ha *entanglement* che consiste in un avvolgimento;
- 3cry, formata da 169 residui, con $|G'_c| = 3.068$, ha *entanglement* che consiste in tre avvolgimenti.

Le matrici di energia ε_{ij} e di contatto Δ_{ij} sono state calcolate usando le coordinate spaziali degli atomi pesanti delle proteine, disponibili nel Protein Data Bank. Come detto in precedenza, due atomi vengono considerati in contatto se distano meno di 4.5 Å. Prima di passare alla simulazione vera e propria, è necessario stabilire le condizioni ambientali in cui il ripiegamento avviene, cioè il valore del parametro λ . Affinché il ripiegamento sia possibile, è necessario che $\lambda > \lambda_c$, dove quest'ultimo è il valore di λ per cui l'energia libera ha due minimi locali equivalenti, corrispondenti allo stato nativo e a quello denaturato. Al di sopra di questo valore, il minimo globale dell'energia libera corrisponde allo stato nativo. Per trovare un valore nell'intervallo corretto, si eseguono delle simulazioni relativamente brevi e si seleziona un λ per cui una buona parte dei residui è nella configurazione nativa ($m_i = 1$). Qui si è scelto di eseguire 100 simulazioni da 10000 *step* Monte Carlo ciascuna, ripetendo il processo aumentando λ ogni volta, finché, in media, alla fine delle simulazioni l'80% dei residui non fosse nella configurazione nativa. I valori di λ così ottenuti sono 0.137842 per 1r6j, 0.115928 per 2hqx, e 0.17813 per 3cry.

Con questi valori per ogni proteina sono state eseguite 25000 simulazioni da 50000 *step* Monte Carlo, tenendo traccia dei tempi di formazione e disfacimento dei contatti presenti nella matrice ε_{ij} . Questi tempi sono stati utilizzati per costruire un istogramma

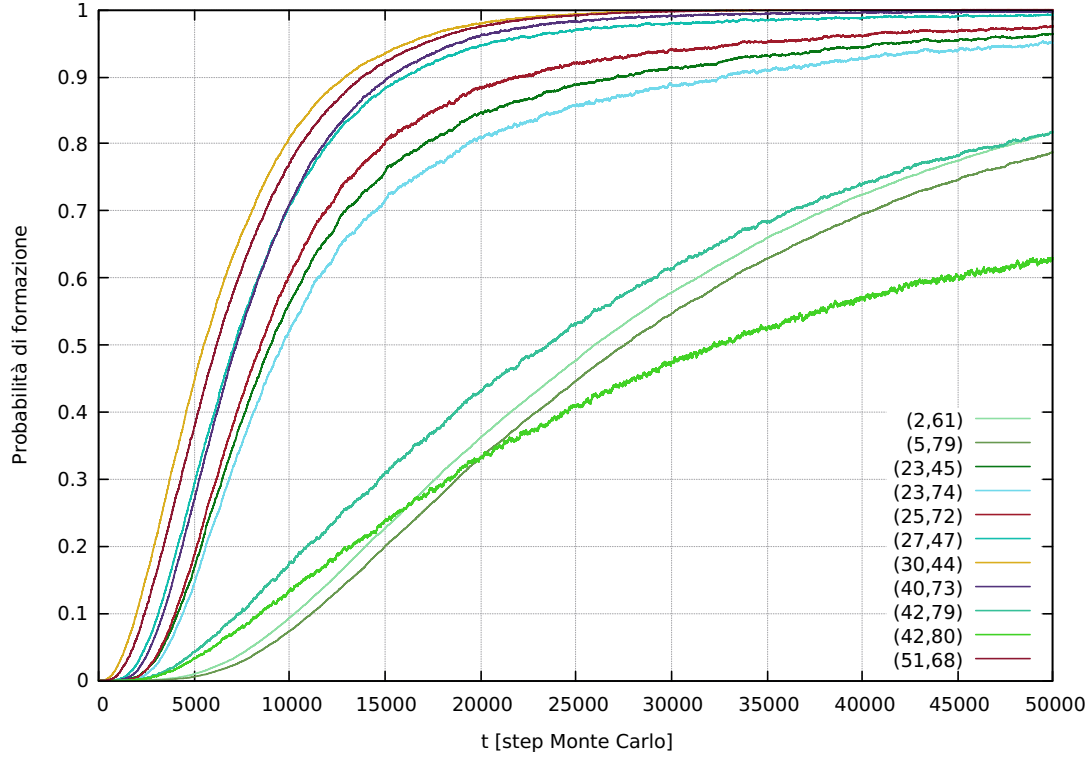


Figura 5.1: Probabilità di formazione di alcuni contatti (i_1, i_2) della proteina 2hqx.

del numero di contatti formati al tempo t , poi normalizzati per ottenere una probabilità (un esempio in Figura 5.1 per la proteina 2hqx). Sono poi state selezionate le probabilità a un quarto, a metà e a fine simulazione ($t = 12500$, $t = 25000$ e $t = 50000$) di tutti i contatti per cui è definito l'*entanglement* gaussiano per verificare come queste sono distribuite in funzione di $|G'_c(i)|$. Di particolare interesse sono le distribuzioni a un quarto e a metà simulazione. Queste mostrano chiaramente che i contatti con $|G'| \geq 0.7$ si formano in media più tardi degli altri, e che le righe su cui si dispongono i dati sembrano corrispondere al numero di avvolgimenti quantificato da $|G'|$. Per 1r6j, che non presenta *entanglement*, a metà simulazione le probabilità si dispongono su una sola riga (Figura 5.2). Per 2hqx, che ha un avvolgimento, i contatti entangled sono disposti su una riga ben separata dal resto dei contatti per tutta la durata della simulazione (Figura 5.3). Per 3cry, si vedono tre righe distinte separate tra loro e dal resto dei contatti: la riga che ha i contatti con $|G'| \geq 2.5$ si forma in media per ultima, e le altre due righe coinvolgono contatti con $|G'| \geq 2$ perché corrispondono a due anelli doppiamente concatenati tra loro nella struttura nativa (Figura 5.4).

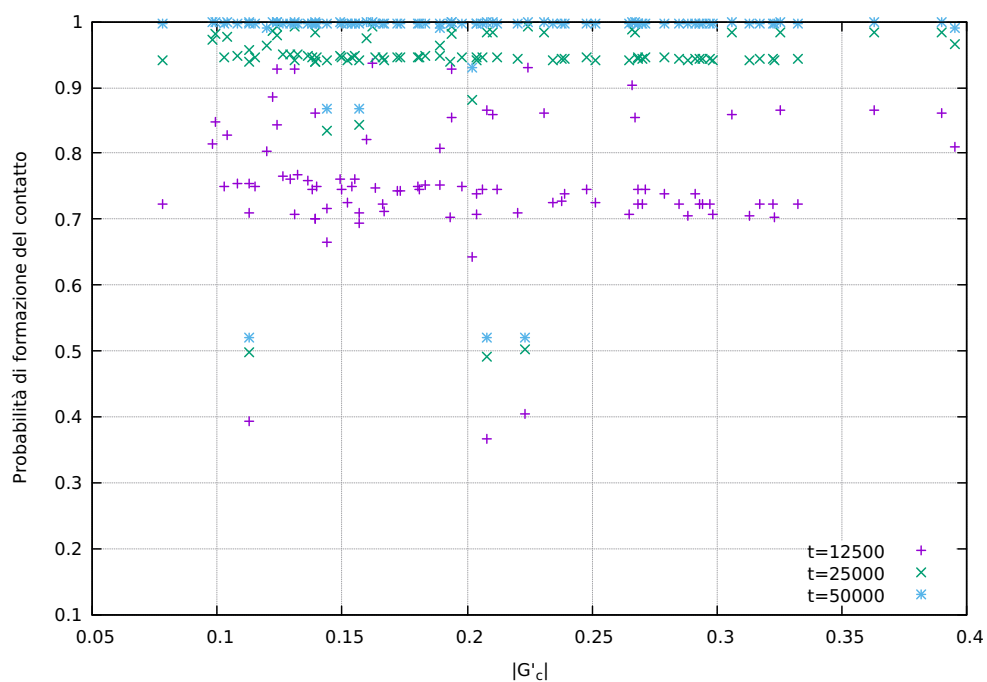


Figura 5.2: Probabilità di formazione dei contatti a vari tempi per 1r6j

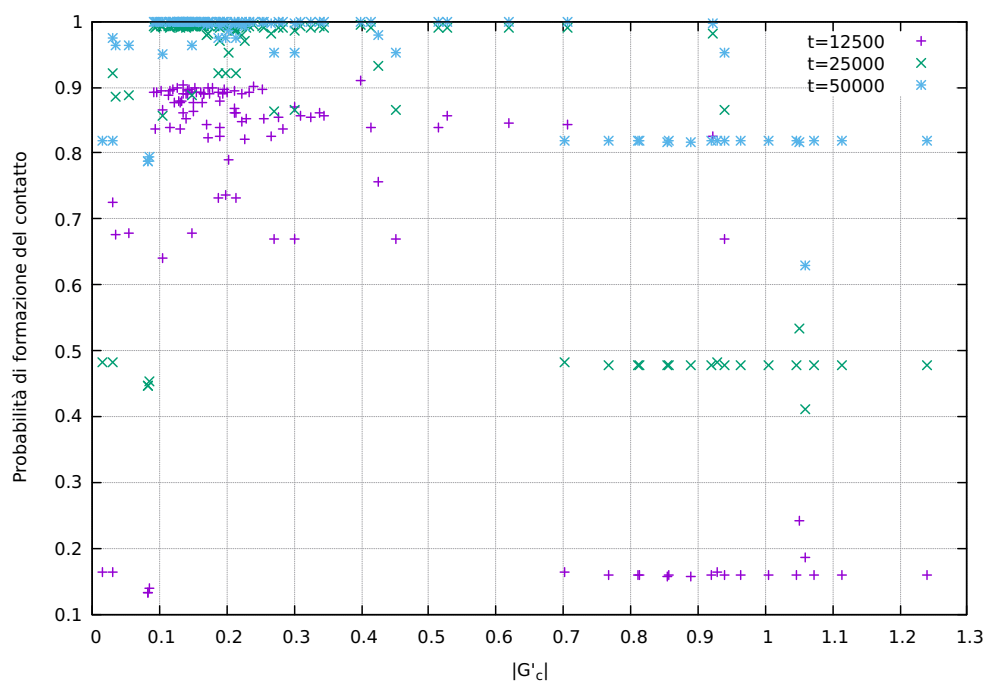


Figura 5.3: Probabilità di formazione dei contatti a vari tempi per 2hqx

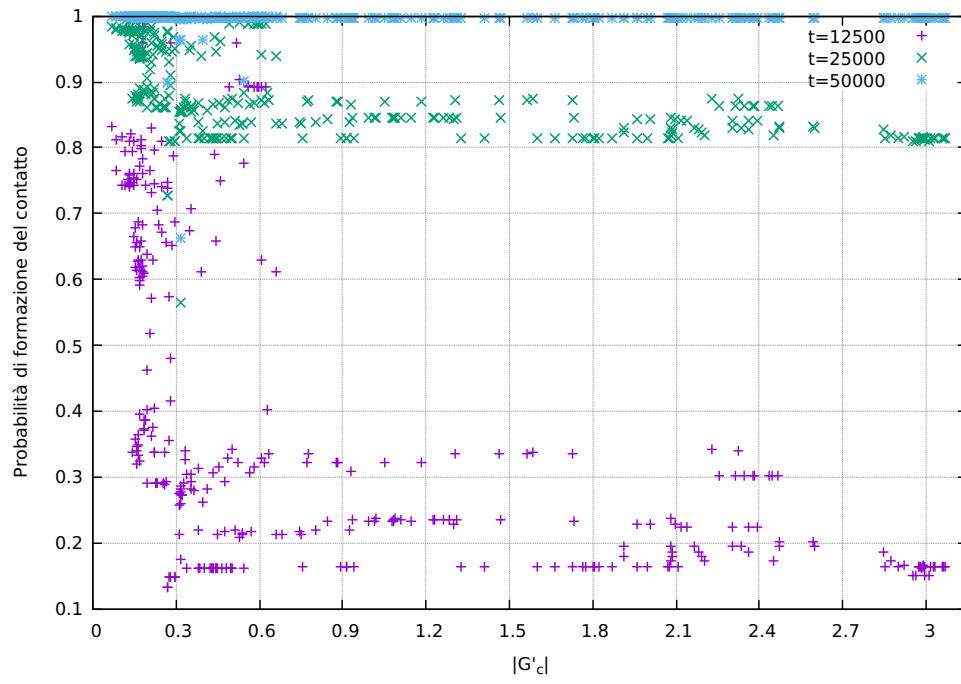


Figura 5.4: Probabilità di formazione dei contatti a vari tempi per 3cry

Capitolo 6

Conclusioni

Lo scopo di questo lavoro di tesi è quello di verificare l'ipotesi, formulata in [5], che i contatti che chiudono gli anelli *entangled* si formino più tardi degli altri. Per farlo, è stato simulato il ripiegamento di tre proteine, scelte in base al loro diverso grado di *entanglement*, usando un modello unidimensionale con interazioni a lungo raggio, calcolate a partire dalla struttura tridimensionale della proteina. Sebbene il modello sia appunto unidimensionale, esso è riuscito a dare un segnale correlato con la presenza di *entanglement* topologico tridimensionale che è in accordo con l'ipotesi di partenza.

Per confermare questa conclusione saranno necessari ulteriori studi. Ad esempio si potrebbe estendere la stessa analisi a un numero maggiore di proteine *entangled*; in particolare sarebbe interessante verificare se la corrispondenza tra il numero di righe nei grafici dei contatti e il numero di avvolgimenti presenti in una proteina non è solo una coincidenza. Si potrebbero analizzare le stesse proteine con un modello più sofisticato, ad esempio un SBM, anche a grana grossa, per verificare se si ottengono gli stessi risultati. Infine, dati la sua semplicità, il suo basso costo computazionale, e il fatto che esistono buone approssimazioni analitiche per studiare la cinetica [9], lo stesso modello WSME può essere usato per simulare un meccanismo evolutivo che selezioni i percorsi di ripiegamento più veloci (in maniera simile a quanto fatto in [11]).

Bibliografia

- ¹A. Finkelstein e O. Ptitsyn, *Protein Physics* (Academic Press, 2002).
- ²A. Kassel e N. Ben-Tal, *Introduction to Proteins: Structure, Function, and Motion* (Chapman e Hall/CRC, 2018).
- ³N. Lim e S. Jackson, «Molecular knots in biology and chemistry», *Journal of Physics: Condensed Matter* **27** (2015).
- ⁴S. Jackson, A. Suma e C. Micheletti, «How to fold intricately: using theory and experiments to unravel the properties of knotted proteins», *Current Opinion in Structural Biology* **42**, 6–14 (2017).
- ⁵M. Baiesi, E. Orlandini, F. Seno e A. Trovato, «Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding», *Scientific Reports* **9**, 8426 (2019).
- ⁶P. Whitford, K. Sanbonmatsu e J. Onuchic, «Biomolecular dynamics: order-disorder transitions and energy landscapes», *Reports on Progress in Physics* **75**, 076601 (2012).
- ⁷V. Muñoz e W. Eaton, «A simple model for calculating the kinetics of protein folding from three-dimensional structures», *Proceedings of the National Academy of Sciences* **96**, 11311–11316 (1999).
- ⁸P. Bruscolini e A. Pelizzola, «Exact Solution of the Muñoz-Eaton Model for Protein Folding», *Physical Review Letters* **88**, 258101 (2002).
- ⁹M. Zamparo e A. Pelizzola, «Kinetics of the Wako-Saitô-Muñoz-Eaton Model of Protein Folding», *Physical Review Letters* **97**, 068106 (2006).
- ¹⁰M. Zamparo e A. Pelizzola, «Nearly symmetrical proteins: Folding pathways and transition states», *The Journal of Chemical Physics* **131**, 035101 (2009).
- ¹¹F. Norbiato, F. Seno, A. Trovato e M. Baiesi, «Folding Rate Optimization Promotes Frustrated Interactions in Entangled Protein Structures», *International Journal of Molecular Sciences* **21**, 213 (2020).
- ¹²K. Dill e J. MacCallum, «The Protein-Folding Problem, 50 Years On», *Science* **338**, 1042 (2012).