UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea Triennale in Fisica

Tesi di Laurea

Modelli semplificati per il ripiegamento cotraslazionale di proteine

Relatore

Prof. Antonio Trovato

Laureando

Antonio Feltrin

Anno Accademico 2021/2022

Indice

Sommario

> Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law
>
> *Douglas Hofstadter, [1]*

Le proteine globulari ripiegano in una struttura tridimensionale nativa, biologicamente attiva, univocamente determinata dalla sequenza di aminoacidi. In anni recenti, è diventato oggetto di studio anche sperimentale il processo di ripiegamento cotraslazionale, che ha luogo "in vivo" mentre la proteina viene sintetizzata nel ribosoma. In questo lavoro di tesi, si propone di adattare al caso cotraslazionale un semplice modello di spin in d=1, già utilizzato in passato ([2]) per studiare la cinetica di ripiegamento di proteine per mezzo di simulazioni Monte Carlo.

Capitolo 1

Struttura delle proteine

Le proteine sono polimeri che presentano una gerarchia di interazioni fra i loro costituenti. Queste interazioni determinano la loro funzione e il loro ripiegamento. I costituenti delle proteine (residui[1]) sono gli aminoacidi, o "peptidi", una famiglia di molecole biologiche dall'architettura comune. Ogni aminoacido è composto da un carbonio centrale, detto C_α, a cui sono legati un gruppo amminico e un gruppo carbossilico. Ciò che differenzia un aminoacido dall'altro è il terzo gruppo legato al C_α, detto catena laterale. La catena laterale può consistere in un solo atomo di idrogeno o in strutture più complicate, come un anello aromatico. Ogni catena polipeptidica[2] presenta un N-terminale, da cui ha inizio la sintesi, e un C-terminale. La catena polipeptidica si forma tramite legami di condensazione fra il C-terminale e l'N-terminale di due aminoacidi adiacenti nella sequenza. Questo particolare tipo di legame, detto legame peptidico, presenta una caratterizzazione planare. Ciò impedisce rotazioni attorno all'asse che unisce gli atomi di carbonio e azoto dei due aminoacidi. Conseguentemente, nella catena polipeptidica i gradi di libertà rotazionali dei singoli aminoacidi sono dovuti ai legami del carbonio centrale: i parametri di rotazione per ogni aminoacido si possono identificare con gli angoli ϕ, ψ di rotazione dei gruppi amminico e carbossilico attorno ai loro legami con il carbonio centrale ([3], pag. 21). Le proteine sono molto varie in lunghezza: le più corte sono composte da qualche decina di residui, mentre le più complesse possono arrivare anche a migliaia di aminoacidi. Ciononostante, esistono tre livelli di organizzazione comuni a tutte le proteine (quattro per le più complesse). La struttura primaria consiste semplicemente nella catena polipeptidica. Quella secondaria si fonda su interazioni locali tra aminoacidi, cioè peptidi vicini nella sequenza (qualche unità). Le principali strutture secondarie sono il foglietto-β e l'α-elica. Nel primo gli aminoacidi assumono una conformazione a linea spezzata planare, mentre nel secondo formano un'elica quasi circolare a passo ridotto (sempre qualche aminoacido). La struttura terziaria coinvolge interazioni non locali fra diverse parti della catena, ed è responsabile della forma finale della proteina (determinando anche la sua funzionalità). Una prima distinzione fra le varie strutture terziarie è quella fra proteine globulari e tubulari, dalla forma compatta o allungata. Infine, la struttura quaternaria riguarda le proteine più grandi, che sono a loro volta composte da proteine più piccole che fungono da

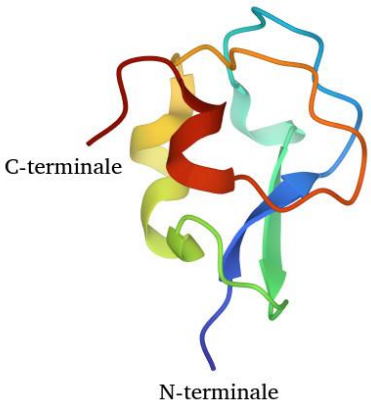

Figura 1.1: Schema della struttura 3D di 1ucs

[1] Si usa il termine "residuo" per indicare il monomero di una macromolecola biologica, come il DNA o una proteina. Nel primo caso i residui sono le basi azotate, nel secondo gli aminoacidi. In questo lavoro si userà "residuo" come sinonimo di aminoacido.

[2] Si usa il termine "polipeptide" o "catena polipeptidica" per concentrarsi sul polimero di aminoacidi, senza tenere conto dello stato (nativo o denaturato) della catena. Parlando di "proteina" invece si presuppone una molecola assemblata, ripiegata e funzionante.

subunità. Nel caso di studio di questa tesi, l'interesse è rivolto verso proteine globulari. La proteina modellizzata in questo lavoro è la "1ucs", secondo la denominazione del Protein Data Bank. Si tratta di una proteina globulare con funzioni antigelo, prodotta da una specie di anguille. Consta di 64 aminoacidi e presenta sia α-eliche che foglietti-β.

1.1 Ripiegamento delle proteine

Il ripiegamento delle proteine è un processo "*all-or-none*", ammette cioè solamente due stati termodinamicamente stabili, quello nativo, detto anche *folded* (F) e quello denaturato, o *unfolded* (U). L'esistenza di due soli stati fa presupporre che esista una barriera di energia libera $F = E - TS$ che li separa (se la barriera non ci fosse si dovrebbe osservare un *continuum* di intermedi tra U ed F). Già solo l'interazione con un ambiente acquoso indirizza gli aminoacidi verso la corretta configurazione. In particolare, le catene laterali giocano un ruolo fondamentale nel caso delle proteine globulari. A seconda che i gruppi laterali siano idrofili o idrofobici, essi inducono il residuo di appartenenza ad assumere una conformazione nella catena che sia tale da portare il gruppo laterale rispettivamente verso l'esterno o l'interno, il tutto nei limiti imposti dai vincoli meccanici negli angoli di legame ϕ, ψ([4]). Un lavoro stima che il 90% dello sforzo per il ripiegamento è compiuto dal lavoro dei gruppi idrofobi e idrofili ([3], pag. 55). Il risultato di questi sforzi è il *molten globule*, una sorta di "bozza" dello stato nativo: i singoli peptidi sono grosso modo nella posizione spaziale corretta, ma i contatti intermolecolari propri della struttura secondaria e terziaria non sono ancora tutti presenti. Al contrario, un grado maggiore di specificità viene ottenuto dalle forze di Van der Waals e dai legami idrogeno, che formando i contatti fra aminoacidi della struttura (secondaria e) terziaria permettono alla proteina di acquisire la sua funzionalità ([3], pag. 55). Il ripiegamento delle proteine viene per questo classificato come un processo cooperativo: la nascita di un contatto ne favorisce altri, permettendo la realizzazione di strutture complesse ([5], pag. 28).

L'intero processo di ripiegamento non dura che pochi millisecondi per proteine corte, e supera il minuto per quelle più lunghe. Le cause di questa rapidità non sono affatto banali, ed anzi costituiscono uno dei primi problemi affrontati dalla ricerca nel settore: come può una catena polipeptidica dotata di innumerevoli gradi di libertà spaziali trovare in così poco tempo la strada per la sua configurazione finale? Levinthal esemplificò queste criticità nel paradosso che porta il suo nome: anche supponendo che ogni aminoacido possa esplorare solo due "stati" nel suo spazio delle configurazioni, una proteina semplice composta da 100 residui si troverebbe di fronte una famiglia di 2^{100} conformazioni possibili. Pur supponendo un tasso di ricerca di miliardi di configurazioni al secondo, appare impossibile che la proteina raggiunga il suo stato nativo esplorando in maniera casuale tale spazio. Per risolvere tale paradosso, sembrò ragionevole imporre che lo stato nativo corrispondesse a un minimo dell'energia libera del sistema. Levinthal teorizzò allora l'esistenza di una "guida cinetica", cioè di un percorso che attraversando lo spazio delle configurazioni fosse capace di condurre a termine il ripiegamento nei limiti temporali osservati negli esperimenti. Come da figura 1.2, si può immaginare un grafico energia-spazio delle fasi come una superficie a imbuto, con lo stato nativo nella configurazione più stabile e quelli denaturati lungo i bordi a energie maggiori ("in alto"). Gli stati intermedi incontrati lungo questa via corrisponderebbero a equilibri metastabili nell'energia libera. Si tornerà sugli intermedi di ripiegamento nelle discussioni dei risultati (e nei confronti con il ripiegamento cotranslazionale).

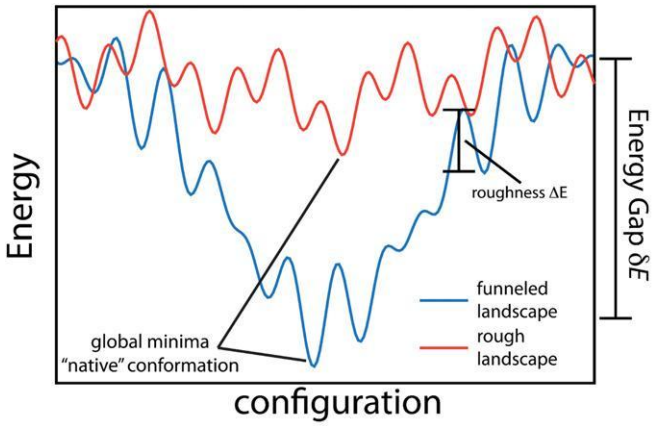

Figura 1.2: Rappresentazione della ruvidità del paesaggio energetico

Per risolvere il paradosso di Levinthal occorre che la "ruvidità (*ruggedness*) dell'imbuto" sia bassa. In questo modo si evitano trappole cinetiche, e il ripiegamento procede come un processo di diffusione nello spazio delle configurazioni ([6], pag. 3). Formalmente, la proteina perde un'energia δE passando dallo stato denaturato a quello nativo. La superficie energia-spazio delle fasi, o "paesaggio energetico" (*energy landscape*) può presentare delle asperità, dei dossi di altezza ΔE. Se la superficie presenta un percorso "liscio" verso lo stato nativo, allora si può definire bassa la ruvidità del sistema, cioè si ha:

$$\Delta E \ll \delta E \tag{1.1}$$

Viceversa, se le asperità costituiscono degli equilibri metastabili in competizione con F, il sistema può bloccarsi in uno stato a metà tra il nativo e il denaturato. Il soddisfacimento della 1.1 coincide con la richiesta che sia verificato il principio di minima frustrazione ([6], pag. 3).La "frustrazione" in un sistema a molti corpi si riferisce all'esistenza di più configurazioni con un confrontabile livello di stabilità dal punto di vista energetico. Un sistema altamente frustrato non possiede quindi uno stato nettamente più stabile degli altri. Il principio di minima frustrazione implica anche che la differenza di energia tra stato nativo e stato denaturato sia grande. Soddisfare il principio sembra quindi perfettamente plausibile con la realtà delle macromolecole biologiche. Riassumendo: durante il ripiegamento la catena perde energia interna grazie alla formazione dei contatti nativi, avvicinandosi verso un minimo di energia. Al tempo stesso però perde anche entropia (esiste un unico stato nativo F, ma ad U corrispondono innumerevoli conformazioni non native), e questo causa l'innalzarsi della barriera di energia libera ([7], pag. 3). La via di ripiegamento ideale allora è quella che porta al minimo di energia con il minor dispendio di entropia possibile.

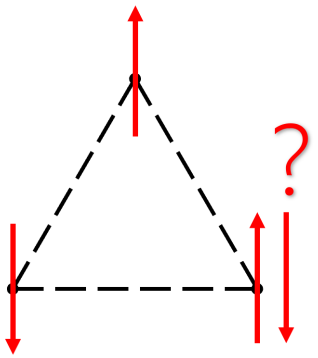

Figura 1.3: Triangolo frustrato di spin: in questo caso l'energia di accoppiamento è $C > 0$, quindi il sistema predilige spin antiparalleli, ma è impossibile trovare una configurazione per cui questa condizione si realizza in tutti e tre i vertici.

Secondo la teoria della cinetica di transizione di Arrhenius, la velocità della trasformazione da uno stato all'altro dipende dall'altezza della barriera di energia libera che li separa (un fenomeno concettualmente simile all'effetto tunnel in meccanica quantistica). Definendo con F_t l'energia libera dello stato di transizione (cui corrisponde l'altezza massima della barriera di energia libera fra gli stati U ed F), si ha che il tempo medio di attraversamento della barriera $\langle\tau\rangle$ nei due casi è:

$$\langle\tau_{U\to F}\rangle = k_0 e^{(F_t - F_U)/k_B T} \qquad \langle\tau_{F\to U}\rangle = k_0 e^{(F_t - F_F)/k_B T} \tag{1.2}$$

Nella sezione 4 discuteremo come questo potrebbe caratterizzare le simulazioni di ripiegamento.

1.2 Denaturazione, *refolding* e *cotranslational folding*

Il processo inverso al ripiegamento, la denaturazione, è anch'esso un processo *all-or-none*, speculare al primo. La denaturazione delle proteine si può schematizzare termodinamicamente come una transizione di fase tra gli stati U ed F[3]. Può avvenire per svariati meccanismi, ad esempio effetti termici o agenti chimici. Poiché il loro funzionamento dipende dalla struttura, e dato che questa si regge a sua volta su legami relativamente deboli (ponti idrogeno e legami di van der Waals), le proteine sono particolarmente sensibili agli sbalzi di temperatura. Ad esempio, nel corpo umano le proteine lavorano a una temperatura di circa $37°C$ $(310K)$, ma denaturano già intorno ai $60°C$. Ci si riferirà alla temperatura oltre la quale la proteina denatura spontaneamente come T_F, o temperatura di *folding*. In corrispondenza di T_F, gli stati U ed F possiedono la stessa energia libera, e l'attraversamento della barriera descritto in 1.2 diventa simmetrico (U ed F sono equiprobabili, $\langle \tau_{U\to F} \rangle = \langle \tau_{F\to U} \rangle$). Se una proteina viene riportata sotto T_F dopo una denaturazione termica può ritornare allo stato nativo. L'interazione con un ambiente interamente acquoso è infatti più importante di quello che si potrebbe pensare: già negli anni '60 Anfinsen scoprì che la denaturazione era un processo reversibile, a patto che la proteina in questione fosse di dimensione ridotta e non avesse subito sostanziali modificazioni chimiche dopo il suo ripiegamento ([3], p.210).

Se il ripiegamento avviene nel suo ambiente naturale, la cellula, viene detto *folding in vivo*. Il processo che avviene nella cellula è condotto dal ribosoma e altre molecole, ma diversi esperimenti osservarono una renaturazione di proteine lasciate libere in acqua, il che fa presupporre che la struttura primaria della proteina contenga le istruzioni necessarie al suo ripiegamento. Questa modalità di ripiegamento prende il nome di *folding in vitro*. I tempi di ripiegamento ([6], pag. 6) sono comparabili con quelli nella cellula, dell'ordine di pochi millisecondi per le proteine di piccola taglia(in realtà il processo cellulare è più veloce, [3], pag. 244). Il ripiegamento cotraslazionale (*cotranslational folding*, da qui abbreviato in CF) consiste nel ripiegamento della catena polipeptidica mentre viene prodotta dal ribosoma. Quest'ultimo è una macromolecola composta da RNA e varie proteine, e ha il compito di assemblare il polipeptide un aminoacido alla volta traducendo le istruzioni contenute nell'RNA messaggero. Il CF per come è biologicamente definito può avvenire esclusivamente nella cellula. Sono le proteine più complesse a beneficiarne maggiormente: questo perché una lunghezza maggiore presenta maggiori rischi di distorsione nel ripiegamento (*misfolding*). La proteina nascente infatti è esposta all'aggregazione con corpi estranei, che possono consistere in altre proteine o generiche molecole presenti nel citoplasma. Per questo sono necessari svariati agenti di supporto durante il processo di prolungamento che avviene nel ribosoma. Il ribosoma svolge parte del lavoro di schermatura ed è accompagnato da vari tipi di "molecole accompagnatrici", o *chaperone molecules*. Il complesso ribosoma-accompagnatrici forma il cosiddetto "*nascent chain welcoming committee*", il cui compito include anche la forgiatura della proteina. Tramite la rotazione del C-terminale e la stabilizzazione dell'N-terminale, il complesso facilita la realizzazione della struttura secondaria della proteina ([3], pag. 244 e [8], pag. 5). Il problema con la riproduzione sperimentale del *folding in vivo* è che studiarlo all'opera in una cellula funzionante è estremamente complicato. Risulta molto più facile porsi in un ambiente più "minimalista" ma controllabile, ad esempio in soluzione acquosa insieme a parte degli enzimi necessari al CF ([3],

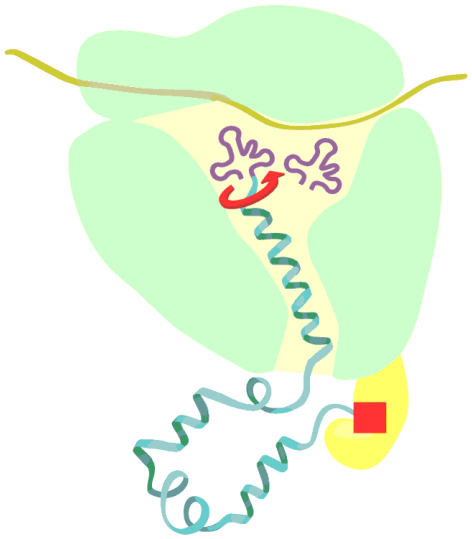

Figura 1.4: Schema del *nascent chain welcoming committee*: quello in verde è il ribosoma, in giallo c'è una *chaperone molecule*, in azzurro la proteina nascente e in viola RNA ribosomiale.

[3]Qui si intende una transizione tra sistemi all'equilibrio: a una certa T_1 il sistema è in equilibrio nello stato F, a $T_2 > T_1$ lo stato di equilibrio diventa U. Invece *in vivo* la proteina viene assemblata e poi evolve verso il suo stato di equilibrio F mentre T resta costante. Nemmeno nel caso del *folding in vitro* si deve intendere il ripiegamento come una transizione di fase: le condizioni ambientali (tra cui T) restano costanti, e il sistema evolve verso l'equilibrio F.

pag. xiv). Questa definizione ridotta di *folding in vivo* viene ad esempio preferita dai fisici, mentre sarebbe ancora classificata come un *folding in vitro* dai biologi. Un approccio simile è adottato per le simulazioni di ripiegamento ("*folding in silico*"). In questo lavoro, le condizioni di simulazione del CF (come la temperatura, v. sezione 4) sono le stesse del *folding in vitro*, fattore che andrà tenuto in considerazione nell'analisi dei risultati.

Capitolo 2

Modelli per il ripiegamento

Negli anni sono stati formulati vari modelli per simulare il ripiegamento di una proteina, e si deve riconoscere a Nobuhiro Gō la paternità di uno dei primi. Sviluppando le idee accennate nella sezione 1.1, Gō le implementò in un modello a reticolo 3D. Va precisato che dalle idee iniziali di Gō furono progettati modelli di diversa specie, ognuno volto a colmare qualche difetto di quello originale: per questo parlando di "modelli Gō" ci si riferisce a una classe di modelli. In misura variabile, vengono definiti "a grana grossa" (*coarse grained*) a seconda del numero di gradi di libertà trascurati, e si basano tutti su due principi:

1. rappresentano transizioni di *folding-unfolding*
2. tengono in considerazione solo i contatti nativi (specificità forte ([7], pag. 4))

Dove con "specificità" si intende il grado di somiglianza con i contatti presenti nello stato nativo. Il punto 1, dall'aspetto più generico, si riferisce al fatto che anche lo stato denaturato deve essere stabile, cioè raggiungibile dalle simulazioni a partire da certe condizioni iniziali. Si può realizzare introducendo un termine di bias entropico nell'energia del sistema. Il punto 2 deriva dalle sperimentazioni di Gō con vari livelli di specificità dei contatti tra gli aminoacidi. Trovò che le sue simulazioni Monte Carlo producevano risultati F solo in casi di specificità forte. Da qui la classificazione dei modelli Gō come *Structure Based Models* (SBM), o "modelli nativocentrici".

In questi lavori Gō discretizzava le coordinate spaziali in un reticolo uniforme (*lattice*), in cui ogni punto era occupabile da un solo aminoacido. Ogni residuo era approssimato da una biglia che si poteva muovere solo da un vertice all'altro del reticolo ([6], pag. 12). Nelle simulazioni si spostava un residuo alla volta in un punto del reticolo che fosse adiacente alla sua posizione originaria, fino al raggiungimento della configurazione nativa. Il problema risiedeva nel fatto che la richiesta di massima autenticità data da una rappresentazione 3D si scontrava proprio con la discretizzazione: pur riducendo i costi computazionali, questa inficiava l'esplorabilità dello spazio delle fasi, mentre una proteina reale è capace di muoversi in uno spazio continuo. In seguito ci si spostò a modelli fuori reticolo (*off-lattice*), con tutti i residui in grado di assumere coordinate in uno spazio tridimensionale continuo. In questo caso, si può distinguere la raffinatezza del modello in base alla selezione degli atomi rappresentati al suo interno. Il più semplice modello fuori reticolo schematizza i residui riducendoli alla posizione dei carboni principali. I modelli più complicati invece contengono tutti gli atomi della proteina, eccezion fatta per gli idrogeni.

2.1 Il modello WSME

Lo svantaggio dei modelli fuori reticolo è l'aumento della domanda di risorse computazionali per eseguirli. Con l'intento di muoversi in direzione opposta pur mantenendo le stesse richieste dei modelli Gō, gli autori del modello WSME [1] coniugarono le idee di Gō con il modello di Ising. Nel modello di

[1] questo modello è stato introdotto da Wako e Saito, e poi ripreso indipendentemente da Muñoz e Eaton.

Ising per un reticolo (solitamente 2D) di spin, si definisce un'hamiltoniana

$$H(\{\sigma\}) = -\Sigma_{<i,j>} C_{ij}\sigma_i\sigma_j$$

La sommatoria va eseguita solo sulle coppie $< i, j >$ di particelle che sono prime vicine (adiacenti) nel reticolo. La minima frustrazione si ha quando la costante di accoppiamento energetico $C_{i,j}$ è positiva per tutte le coppie i, j. Così facendo l'energia del sistema ha un minimo nei due soli stati ordinati in cui tutti gli spin valgono $\sigma_i = +1$ o $\sigma_i = -1$. L'idea chiave del modello WSME è quindi abbandonare la rappresentazione 3D degli atomi o dei carboni principali e passare ad una raffigurazione data da una sequenza 1D di N spin $\{m_i\}_{i=1,2,...N}$ con $m_i \in \{0, 1\}$. Il valore $m_i = 1$ indica che l'aminoacido corrispondente è posizionato come nella struttura nativa, cioè gli angoli di legame ϕ, ψ con l'aminoacido precedente e successivo sono gli stessi della proteina funzionante. Viceversa, $m_i = 0$ include tutte le altre possibili configurazioni. L'osservabile che regola l'evoluzione del sistema è una sorta di "hamiltoniana di energia efficace" che tiene conto delle energie di interazione fra gli aminoacidi. Qui sorge la principale differenza con il modello di Ising, dato che il potenziale di interazione tra gli spin è non locale: nella proteina l'ultimo aminoacido può teoricamente essere in contatto con uno nel centro della catena.

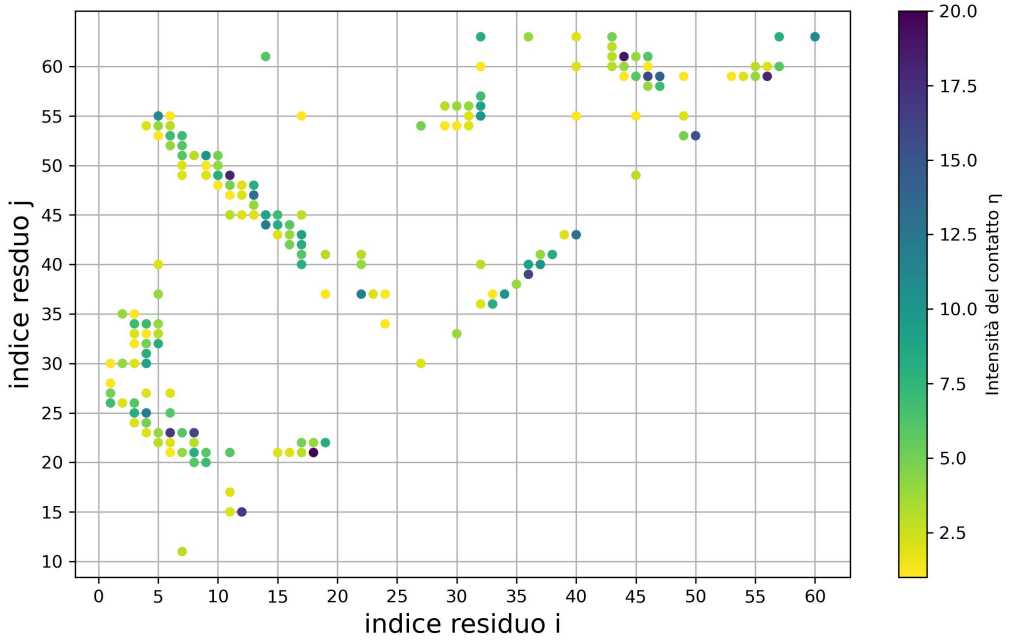

Figura 2.1: Mappa di contatto per 1ucs

Essendo un modello nativocentrico, le interazioni prese in considerazione sono solo quelle tra "contatti" nativi: la struttura nativa 3D è codificata in una "mappa di contatto" che viene definita come segue. Prendendo come riferimento la struttura nativa contenuta nel Protein Data Bank, due aminoacidi vengono definiti "in contatto" se c'è almeno un atomo appartenente all'uno che sia distante meno di R da un atomo dell'altro (atomi di idrogeno esclusi). L'hamiltoniana si presenta con un aspetto del genere:

$$H_{original}(\mu) = -\sum_{i=1}^{N-1}\sum_{j=i}^{N}\epsilon_{ij}\Delta_{ij}\prod_{k=i}^{j} m_k + k_B T\sum_{i=1}^{N} q_i m_i \tag{2.1}$$

Con $\mu = \{m_i\}_{i=1,...,N}$ una qualsiasi configurazione di spin. Il parametro $\epsilon_{ij} > 0$ rappresenta la concentrazione di denaturante, eventualmente pesata con un contributo diverso per ogni contatto. Va ricordato infatti che la proteina ripiega in un ambiente complesso, ricco di acqua e altre sostanze. Il modello però riduce all'essenziale il contesto ambientale in cui avviene il ripiegamento. La matrice Δ_{ij}

è la mappa di contatto; nelle prime versioni vale 1 se fra i peptidi i, j esiste un contatto come definito sopra, 0 altrimenti. La produttoria controlla se il contatto è attivo: questa condizione si realizza solo se tutti gli spin da i a j inclusi sono nello stato "1". Così viene inclusa nel modello la cooperatività della transizione U-F. L'ultimo termine nella H costituisce un contributo di carattere entropico che serve a bilanciare artificialmente lo squilibrio che si genera abbassando la dimensionalità del problema: infatti gli spin possono assumere solo valori binari, ma lo "0" racchiude una pletora di configurazioni 3D non native che vengono sacrificate in favore della semplicità del modello. La perdita di informazione viene compensata penalizzando gli spin nativi, con un peso dato dai parametri $q_i > 0$. In analogia al modello di Ising per il magnetismo, questo termine è chiamato "magnetizzazione" $M = \sum_{i=1}^{N} q_i m_i$. Una grandezza simile alla M ma più utile sperimentalmente è la frazione di contatti nativi formati

$$Q = \frac{\sum_{i<j} \Delta_{ij} \prod_{k=i}^{j} m_k}{\sum_{i<j} \Delta_{ij}}$$

Queste semplificazioni, per quanto brutali, realizzano le condizioni di Gō nei modi seguenti:

1. Per quanto riguarda 1, l'hamiltoniana presenta un chiaro minimo nello stato F, inoltre il secondo termine fornisce un contributo entropico per garantire stabilità anche allo stato denaturato.
2. La 2 è codificata nella mappa di contatto (pesata) dei contatti nativi.

Tralasciando quindi l'evoluzione della cinetica di oggetti 3D, il modello WSME permette di simulare importanti fenomeni e osservabili[2] con un costo computazionale relativamente basso.

Il modello usato è una variante del WSME, dove q_i viene posto uguale ad 1. Il parametro R viene fissato a $R = 4.5$Å, come da [9]. Sono esclusi i contatti tra aminoacidi $i, j : i - j \leq 2$, dato che essendo vicini nella sequenza polipeptidica la probabilità di un'interazione è alta sia in strutture native che denaturate. Inoltre, la mappa di contatto Δ_{ij} viene aggiornata in η_{ij}. La forza del contatto i, j è ora pesata dal numero η_{ij} di coppie di atomi R-vicini: in questo modo si cerca di rispecchiare in modo più accurato il bilancio energetico della proteina reale. Questa è una modifica introdotta in [10] che non è presente nel modello WSME. Inoltre, a differenza di [10], non si prendono come unità di spin i legami, ma gli aminoacidi. Questo genera una minore ridondanza dato che per la proteina 1ucs si passa a 64 elementi anziché 63, ma risulta più intuitivo nella comprensione della cinetica di ripiegamento. L'hamiltoniana risultante è:

$$H(\mu) = -\lambda \sum_{i=1}^{N-1} \sum_{j=i}^{N} \eta_{ij} \prod_{k=i}^{j} m_k + M(\mu) \tag{2.2}$$

Ponendo $\epsilon_{ij} = \epsilon \quad \forall i, j = 1, 2, ..N$ è consentito compiere un'ulteriore importante semplificazione. Dato che in assenza di valori sperimentali l'equazione appena scritta contiene due gradi di libertà nei parametri T ed ϵ, per semplificare le simulazioni e studiare il sistema al variare di uno solo dei due si introduce il parametro $\lambda = \epsilon / k_B T$. Così facendo oltretutto l'hamiltoniana (2.1) viene adimensionalizzata, e il secondo termine coincide con la magnetizzazione della catena di spin.

Il ripiegamento delle proteine avviene al di sotto di T_F, o nel caso dell'equazione 2.2 al di sopra di λ_F. Per $\lambda = \lambda_F$ la proteina oscilla fra gli stati *folded* e *unfolded* con uguale probabilità. Secondo la teoria di Wolynes et al. sulla meccanica statistica del ripiegamento delle proteine, il paesaggio energetico deve possedere bassa ruvidità per evitare fenomeni di "transizione vetrosa". Questa si manifesta come l'incagliamento della cinetica di ripiegamento in uno stato non nativo (*misfolded*). La transizione vetrosa appare al di sotto di una certa temperatura T_g. Si possono allora definire le condizioni di ripiegabilità:

$$T_g < T < T_F \text{ oppure } \lambda_F < \lambda < \lambda_g \tag{2.3}$$

In seguito ci si riferirà anche a λ come a una "temperatura" inversa assumendo di poter tenere ϵ costante.

[2] Come la temperatura di *folding*, la presenza di stati intermedi nel ripiegamento e altro ancora, come verrà discusso in seguito.

Nell'ultima parte dell'analisi, il modello WSME sarà adattato al *folding in vivo*, semplificato come un CF che però avviene nelle stesse condizioni del *folding in vitro*. Si dovrebbe per lo meno cambiare il parametro ϵ - e quindi λ- data la diversa concentrazione di enzimi e altre molecole passando da un *folding* in acqua a uno nella cellula. Del resto ϵ tiene proprio conto della forza dei contatti nativi, che può essere attenuata o favorita dall'ambiente circostante. In conclusione, è sicuramente più comodo studiare il CF con metodi minimali, ma occorre aspettarsi una minore attinenza dei risultati con gli esperimenti.

2.2 Algoritmo di Metropolis

Per le simulazioni di ripiegamento si utilizza l'algoritmo di Metropolis. L'algoritmo serve a calcolare il valor medio di alcune osservabili X per un sistema all'equilibrio termodinamico [3], e risulta utile nel caso queste grandezze siano impossibili da ottenere in modo analitico. Per calcolare una certa X, la meccanica statistica potrebbe ricorrere all'ensemble canonico:

$$\langle X\rangle = \frac{\int X e^{-H/k_BT} d^N p d^N q}{\int e^{-H/k_BT} d^N p d^N q} \tag{2.4}$$

Con $H = H(q_1, p_1, ..., q_N, p_N)$ l'energia del sistema di N particelle. Questa è semplicemente una media di X valutata su tutti i microstati $\{(q,p)_i\}_{i=1,...,N}$, e pesata per la loro probabilità $e^{-H/k_BT} = e^{-\beta H}$. Il fattore di normalizzazione nell'espressione sopra è la funzione di partizione Z[4], ma per ottenerla è necessario risolvere un'integrale della stessa complessità di quello per ricavare $\langle X\rangle$([11]). L'idea è allora sfruttare un'integrazione Monte Carlo modificata ottenendo il processo che segue. In un'integrazione Monte Carlo classica si sceglierebbero equiprobabilmente punti dallo spazio delle fasi, per poi pesarli nell'integrale con $e^{-\beta H}$ ([12]). Questo è sconveniente nel caso il valore dell'esponenziale sia piccolo. Per cui si procede "al contrario", selezionando le configurazioni con una probabilità del tipo $e^{-\beta H}$ e pesandole uniformemente.

Il procedimento di Metropolis applicato al modello WSME è il seguente:

1. Partendo da una configurazione $\{m_i\}^{ini}_{i=1,...,N}$ si genera un numero casuale $j \in [1, 64]$ e si inverte lo spin m_j dell'aminoacido corrispondente. Si ottiene così una nuova configurazione $\{m_i\}^{fin}_{i=1,...,N}$
2. Si valuta la differenza di energia $\Delta H = H_{fin} - H_{ini}$
3. La nuova configurazione viene scelta con probabilità $P = min\{1, e^{-\beta\Delta H}\}$. In altre parole, se $\Delta H < 0$ si conferma la nuova configurazione; se $\Delta H > 0$ si resta nel nuovo stato (meno stabile) con probabilità $e^{-\beta\Delta H}$, altrimenti si torna allo stato iniziale.

La probabilità di passare dallo stato a allo stato b è definita in modo da soddisfare il "bilancio dettagliato": questa è una condizione necessaria per ottenere il campionamento della 2.4 all'equilibrio. Il bilancio dettagliato chiede che all'equilibrio $p(a)T(a \to b) = p(b)T(b \to a)$, con $p(a)$ la probabilità di essere nello stato a e $T(a \to b)$ la probabilità di passare dallo stato a a b. In Metropolis

$$\frac{p(b)}{p(a)} = \frac{T(a \to b)}{T(b \to a)} = e^{-\beta(H_b - H_a)} \tag{2.5}$$

Se $H_b > H_a$, l'esponenziale è minore di 1 e si sceglie $T(b \to a) = 1$, restando con $T(a \to b) = e^{-\beta(H_b - H_a)}$. Se $H_b < H_a$, l'esponenziale è maggiore di 1 e si fissa $T(a \to b) = 1$. In altri lavori[5] si sfrutta l'algoritmo di Metropolis per stimare anche il tempo impiegato dal sistema a raggiungere l'equilibrio. Questa è un'assunzione importante per l'analisi dati che seguirà, e si basa sull'idea che il sistema evolva verso lo stato di equilibrio percorrendo un cammino nello spazio delle fasi tanto più rapido quanto più stabile sia lo stato finale (plausibile per la dinamica Metropolis nel limite ideale

[3] Quelle che sono di interesse per questo lavoro sono l'energia efficace H, la magnetizzazione M e la frequenza di contatti nativi Q, oltre alla raffigurazione $\{m_i\}$ della catena di spin.

[4] A meno della costante $\frac{1}{N!h^N}$

[5] In [10], pag. 3

$\delta E \gg \Delta E$). Nel caso della proteina, se lo stato di partenza è U, ci si aspetta che a $T < T_F$ il "tempo di *folding*" (numero di passi Metropolis richiesti) sia in qualche modo paragonabile al numero di passi τ_{ciclo} nella simulazione. Volendo essere rigorosi fino in fondo, purtroppo non esiste alcuna dimostrazione che l'utilizzo di Metropolis sia rappresentativo della cinetica di *folding* in questo modello. La principale virtù delle tecniche Monte Carlo è proprio quella di portare al campionamento di integrali del tipo descritto in 2.4. Ogni semplificazione, ogni livello di *coarse graining* distorce la dinamica del sistema. L'approssimazione del modello WSME di rinunciare alla descrizione della proteina in 3D per usare una catena di spin in 1D è di suo già così grande che la scelta arbitraria della cinetica Monte Carlo può anche passare in secondo piano. Come ultima nota di carattere tecnico, l'implementazione del WSME in Metropolis non è rapida come per un modello di Ising, perché l'interazione non locale fra gli aminoacidi rende unica la rete di contatti di ogni residuo. Questo obbliga a richiamare la funzione energia ad ogni iterazione senza poter ricorrere ad approssimazioni.

Capitolo 3

Simulazioni di *Folding in vitro*

3.1 Stima temperatura di *folding*

Stimare λ_F è importante per avere una simulazione di CF il più fedele possibile alla realtà: per una proteina umana ad esempio, la temperatura operativa è circa $T_{nat} = 37°C$ e la denaturazione avviene a $T_F = 60°C$. Questo si traduce in una $\lambda_{nat} = \lambda_F/0.9$ per il modello impiegato, supponendo che i rapporti del caso umano valgano anche per 1ucs. Il primo tentativo di stimare λ_F parte dallo studio di Q e σ^2_M al variare di λ per simulazioni di *refolding* a τ_{ciclo} fissata. Con $\sigma^2_M = \langle(M - \langle M\rangle)^2\rangle = \langle M^2\rangle - \langle M\rangle^2$ si intende la varianza di M calcolata su r simulazioni, che deve essere massima in prossimità della temperatura di *folding*. Questo è intuitivo se si pensa che lontano da T_F la proteina è sempre nello stato F o nello stato U, mentre a T_F si il sistema dovrebbe presentarsi negli stati U ed F con uguale probabilità, da cui la varianza maggiore. Similmente, nello stato F $Q \approx 1$ e in U $Q \approx 0$, quindi a T_F si ha $Q \approx 0.5$. Lo studio di Q consiste in un fit lineare semplice, scegliendo $\langle Q\rangle = 0.5$ come valore di soglia per individuare λ_F. Vengono eseguite $r = 100$ simulazioni di $\tau_{ciclo} = 4 \cdot 10^5$ passi Metropolis per ogni valore di λ, e i valori di M e Q sono calcolati come medie nell'ultimo 5% dei passi di ogni simulazione (si suppone quindi che nell'ultimo 5% della simulazione si sia raggiunto l'equilibrio). Viene poi eseguita un'ulteriore media su tutte le r simulazioni, il cui risultato è usato per i fit.

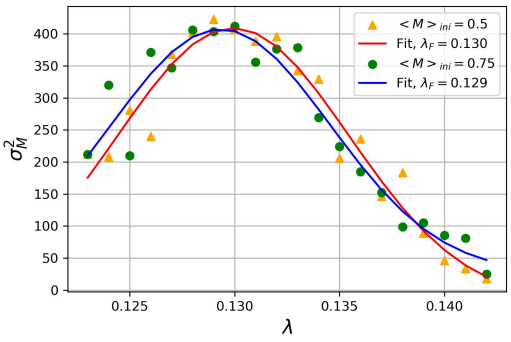

Figura 3.1: Stima di λ_F tramite σ^2_M

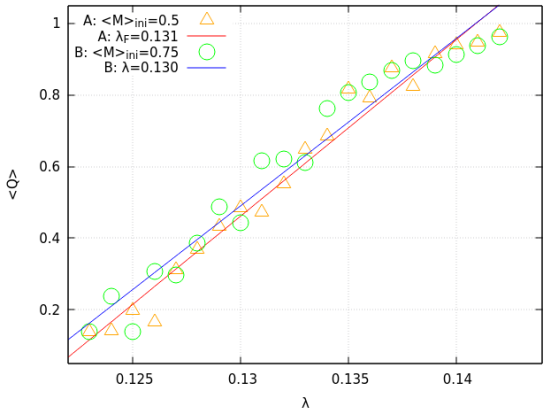

Figura 3.2: Stima di λ_F tramite Q

La distribuzione di σ^2_M dovrebbe presentare un andamento a campana con il picco a λ_F, e un fondo costante altrove. Tale occorrenza è verificata dalle simulazioni graficate sopra, distinte in due famiglie a seconda delle condizioni iniziali: una parte da una M iniziale media per la catena di $\langle M\rangle_{ini} = 0.5$, l'altra da $\langle M\rangle_{ini} = 0.75$. Le interpolazioni concordano nel fornire una prima stima $\lambda_{F,1} \approx 0.13$, ma c'è un problema. Una simulazione prolungata dovrebbe mostrare un'oscillazione del sistema tra i due stati, che però non si verifica. Una prima spiegazione è che, come accennato dalla 1.2, il tempo medio di attraversamento della barriera di energia libera sia troppo elevato, anche a λ_F. Una ragione di questo comportamento potrebbe essere il tempo di simulazione troppo breve: se il sistema non riesce

a raggiungere lo stato di equilibrio, allora λ_F è sovrastimato. In altre parole, i valori di λ per cui $\tau_{ciclo} = 4 \cdot 10^5$ non produce *folding* dovrebbero in realtà mostrarlo, per cui $\lambda_F < \lambda_{F,1}$.

Si esegue allora uno *sweep* anche rispetto a τ_{ciclo}. Vengono eseguite due classi di simulazioni, di *folding* e *unfolding*, a seconda che le condizioni iniziali corrispondano a uno stato rispettivamente U o F. Per le simulazioni di *folding* è stata imposta una magnetizzazione media iniziale di 0.5, che si traduce in una $\langle Q \rangle_{ini} < 0.1$.

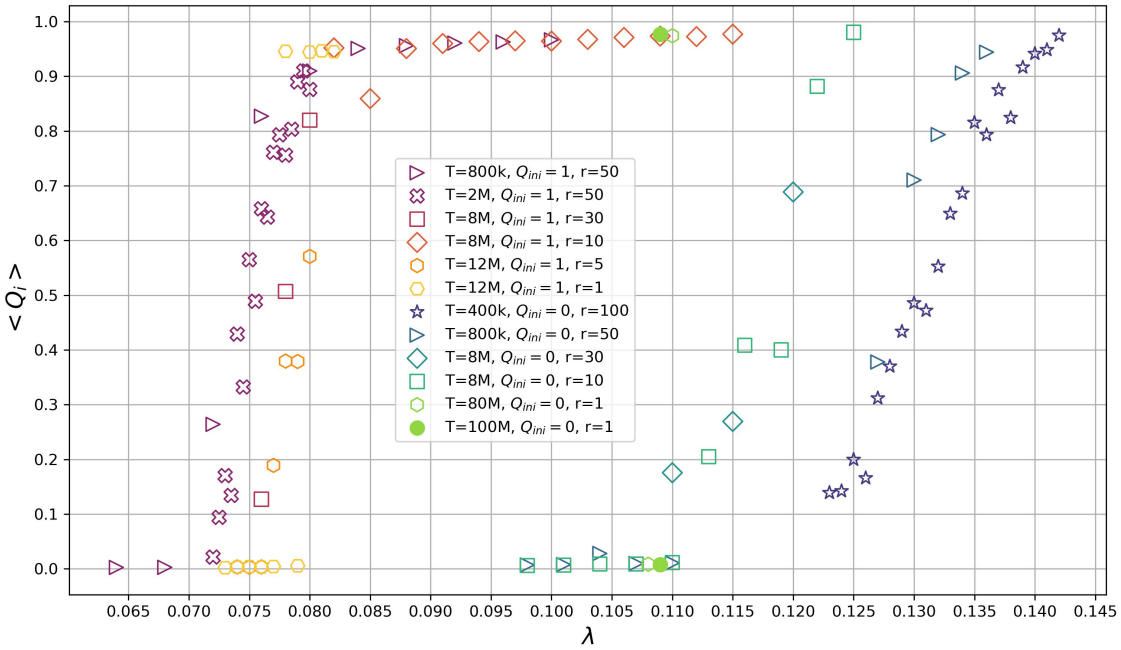

Figura 3.3: Ricerca di λ_F. Le prime 6 serie di dati provengono dal set di *unfolding*, le altre da quello di *folding*.

Vengono prodotte due famiglie di set, che producono delle stime $\lambda_F(\mathrm{U}, \tau_{ciclo})$ e $\lambda_F(\mathrm{F}, \tau_{ciclo})$ (con λ_F selezionabile come quella a cui $\langle Q \rangle = 0.5$). In figura 3.3 si possono osservare le varie serie di $\langle Q \rangle(\lambda, \tau_{ciclo})$ mediate su r simulazioni, corrispondenti alle curve di transizione di fase di *unfolding* e *folding*. Il numero r per ogni valore di λ è stato scelto in base alle disponibilità computazionali. Se il tempo di simulazione fosse abbastanza lungo, $\tau_{ciclo} = \tau^*$, si dovrebbe osservare una convergenza delle curve di *folding* e *unfolding*, con $\lambda_F(\mathrm{U}, \tau^*) = \lambda_F(\mathrm{F}, \tau^*)$. Invece, al variare di λ si osserva come non ci sia aderenza tra le curve di *folding* e *unfolding*, che anzi restano molto distinte anche per una grande varietà di tempi Metropolis (da $800k$ fino a $8M$ di passi). Il tempo macchina richiesto per eseguire i calcoli ha reso sconveniente procedere oltre con le simulazioni. In particolare, nemmeno per dei $\tau_{ciclo} > 80M$ passi si verificano oscillazioni tra gli stati U ed F. Si può di nuovo ipotizzare che ciò sia dovuto a un elevato tempo caratteristico di attraversamento della barriera di energia libera tra i due stati. Questa condizione è soddisfatta anche con un ΔF relativamente basso, come spiegato in 1.2. Purtroppo è difficile stimare $\tau_{U \to F}$ o ΔF non conoscendo k_0, inoltre va ricordato che non è garantita una corrispondenza tra i passi Metropolis e il tempo reale.

Per stimare la "temperatura" λ_{nat} si sceglie allora un valore che sia più compatibile con i test di *folding*, che sono quelli in cui applicare il CF. La ricerca di λ_F è fallita perché in quel range di λ il tempo richiesto per raggiungere l'equilibrio era troppo grande. Ci si pone allora a $\lambda_{nat} = 0.135$, un valore sufficientemente conservativo da garantire tale richiesta. Con simulazioni da 1 milione di passi e una $\langle M \rangle_{ini} = 0.5$, $\langle Q \rangle_{ini} \approx 0$, il polipeptide ripiega infatti nel 97% circa dei casi, in un tempo medio di $266k$ passi.

3.2 Stati intermedi di ripiegamento

Un elemento interessante nella cinetica del ripiegamento consiste nella presenza di stati intermedi, ossia porzioni (non necessariamente contigue) della catena polipeptidica che assumono la conformazione nativa prima delle altre. Formalmente, si fissano dei valori di riferimento per gli stati nativo e denaturato: si caratterizza U con $Q < 0.1$, e F con $Q > 0.95$. Anche la scelta di Q come variabile di riferimento è arbitraria, ma rappresenta meglio la complessità del problema: infatti a parità di M, due stati possono possedere energie molto variabili, perché residui diversi prendono parte in un diverso numero di contatti nativi. Come esempio limite, "spegnere" l'aminoacido 20 in una catena completamente nativa ha come risultato la disattivazione di ben 99 contatti. Inoltre nel computo dell'energia lo studio di Q ha una rilevanza maggiore, dato che lo stato fondamentale ($\mu = F$, $Q = 1$)ha un'energia $H_{fond} = H(F) - M(F) = -\lambda \cdot \Sigma_{i<j}\eta_{ij} = -90.1$ con $\lambda = 0.100$, che è un valore medio nel range studiato nella sezione 3.1 ma relativamente conservativo (basso) per le simulazioni di *folding in vitro*. D'altro canto, $M = 64$ porta un contributo minore all'energia totale, confermando che la vicinanza di Q a 1 è un fattore di maggior rilievo nel raggiungimento dello stato nativo. Infine, esistono dei residui estremali (m_i, $i = 63, 64$) che non prendono parte ad alcun contatto nativo, quindi sono irrilevanti per la "vera" definizione di stato nativo. Come conseguenza, la configurazione con questi due residui a $m_i = 0$ costituisce il vero minimo rispetto ad H.

Si definisce allora "stato intermedio" una famiglia di configurazioni di spin I che siano abbastanza simili fra loro, cioè tali da avere un $\langle Q \rangle \in [Q(I) - \Delta Q, Q(I) + \Delta Q]$ [1], $\Delta Q = 0.06$. Il valore di riferimento va mantenuto per almeno $\tau_{inter} = 5000$ passi Metropolis. La famiglia di configurazioni soprattutto deve presentare una sequenza di spin consistentemente nativi (con un $\langle m_i \rangle > 0.9$). Inoltre la ricerca viene eseguita ogni $\tau_{refresh} = 1000$ passi, per evitare troppa correlazione tra gli stati inseriti nel calcolo degli $\langle m_i \rangle$.

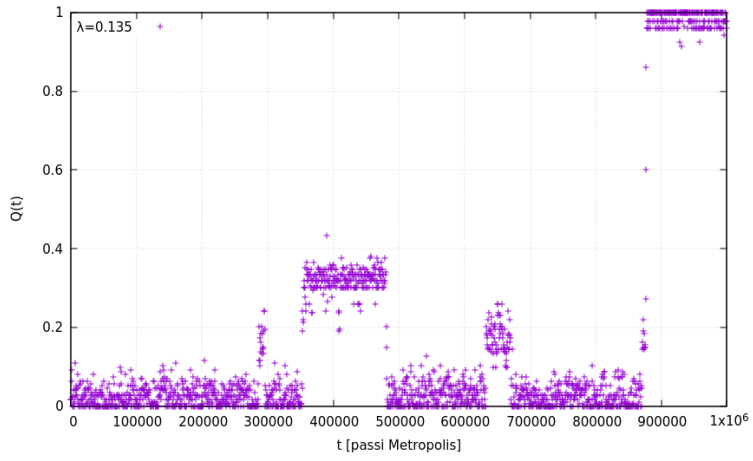

Figura 3.4: Simulazione con apparizione degli intermedi N e C

I valori di τ_{inter}, $\tau_{refresh}$ e ΔQ sono scelti arbitrariamente, cercando un equilibrio tra la sensibilità nell'individuazione di intermedi e il margine d'errore dato dalle naturali oscillazioni casuali della dinamica Monte Carlo. Siccome la ricerca di intermedi parte dall'individuazione di intervalli $Q \pm \Delta Q$ che vengano mantenuti per almeno un τ_{inter}, non è garantito che a questi intervalli in Q corrisponda un unico intermedio. Per confermare che non esistano altre configurazioni I' degeneri, si costruisce un semplice algoritmo di identificazione. Si campiona un piccolo (di dimensione $L = 100$) insieme $S_{standard}$ di simulazioni in cui un certo intermedio I si presenta nella "stessa"[2] sequenza di spin nativi. Si calcola quindi la mappa di spin m_i medi $\langle \nu \rangle = (\langle m_1 \rangle, \langle m_2 \rangle, ..., \langle m_N \rangle)$. La $\langle \nu \rangle$ è ottenuta da *snapshots* della catena di spin in tutti gli istanti di tutte le L simulazioni in $S_{standard}$ in cui si presenta il candidato-I. Da $\langle \nu \rangle$ si ricava anche la deviazione standard campionaria per tutti i vettori $\nu \in S_{standard}$ (ricordando che $dim(\nu) = N = 64$):

$$\sigma_m = \sqrt{\frac{\sum_{l=1}^{L}\sum_{i=1}^{N}(\langle \nu \rangle_i - \nu_{l,i})^2}{L-1}} \tag{3.1}$$

Con questa si opera una misura di compatibilità su un altro campione S_{test} (≈ 100 simulazioni), computando per ogni candidato-I $\nu_{test} \in S_{test}$ il parametro di identificazione $\Delta m = \langle \nu \rangle - \nu_{test} =$

[1] Si descrive l'intermedio da un punto di vista "a basso livello" fondato sugli M, ma lo si valuta a un "livello più alto" nelle Q.

[2] Oscillazioni nella mappa di spin sono plausibili sia nel contesto del ripiegamento che del Metropolis.

$\sum_{i=1}^{N}\langle m_i(\nu)\rangle - m_i(\nu_{test})$. Se $|\Delta m| < \sigma_m$, allora si accetta lo stato ν_{test} come intermedio I. Per gli intermedi trovati, si conclude che non esiste degenerazione, ossia ad ogni valore di $\langle Q\rangle$ corrisponde un solo intermedio. L'ultimo passo è verificare che gli intermedi restino gli stessi al variare della temperatura. Oltre a simulazioni eseguite a $\lambda_{nat} = 0.135$, si cercano intermedi a $\lambda_2 = 0.146$, confrontandoli poi nei due casi. Il valore di λ_2 è molto prudente (si discosta molto dalla "vera" λ_F), e c'è il rischio di finire sotto la temperatura di transizione vetrosa. Basta però fare un test di ripiegamento per verificare che a λ_2 la catena si piega con altissima percentuale (quasi 100% con $1M$ di passi).

Si trovano due stati intermedi, uno con una $\langle Q\rangle$ di circa 0.17 e uno con $\langle Q\rangle = 0.32$. Quest'ultimo è di maggior interesse perché ha una durata maggiore, ma soprattutto consiste nella seconda metà del polipeptide (quella verso il C-terminale) interamente nativa; per tale motivo, viene denominato "stato C". L'altro stato presenta una sequenza di spin nativi nella prima metà della catena, per cui viene battezzato "stato N". Lo stato N presenta una "porzione nativa" più debole rispetto a quella dell'intermedio C, assumendo valori nel range $\langle m_i\rangle \in [0.85, 0.95]$. Per confronto, gli $\langle m_i\rangle$ nativi (per $i > 32$) di C sono nel range $[0.95, 1]$. Il test di identificazione viene superato nel 96% dei casi per l'intermedio C, e nel 73% per l'intermedio N. Il risultato per N è plausibile considerando che N è più vicino allo stato denaturato rispetto a C, e presenta una varianza maggiore nella mappa di spin. Lo stato C si presenta con una frequenza del 54% nelle simulazioni alla $\lambda = \lambda_{nat}$, mentre l'intermedio N appare nel 62% dei casi a fronte di un tempo di permanenza medio 10 volte più piccolo. Bisogna osservare ora come gli intermedi citati nella sezione 1.1 come equilibri metastabili nell'energia libera non sono del tipo di quelli descritti nell'analisi, che invece hanno natura transiente (e per questo si potrebbero definire "intermedi cinetici").

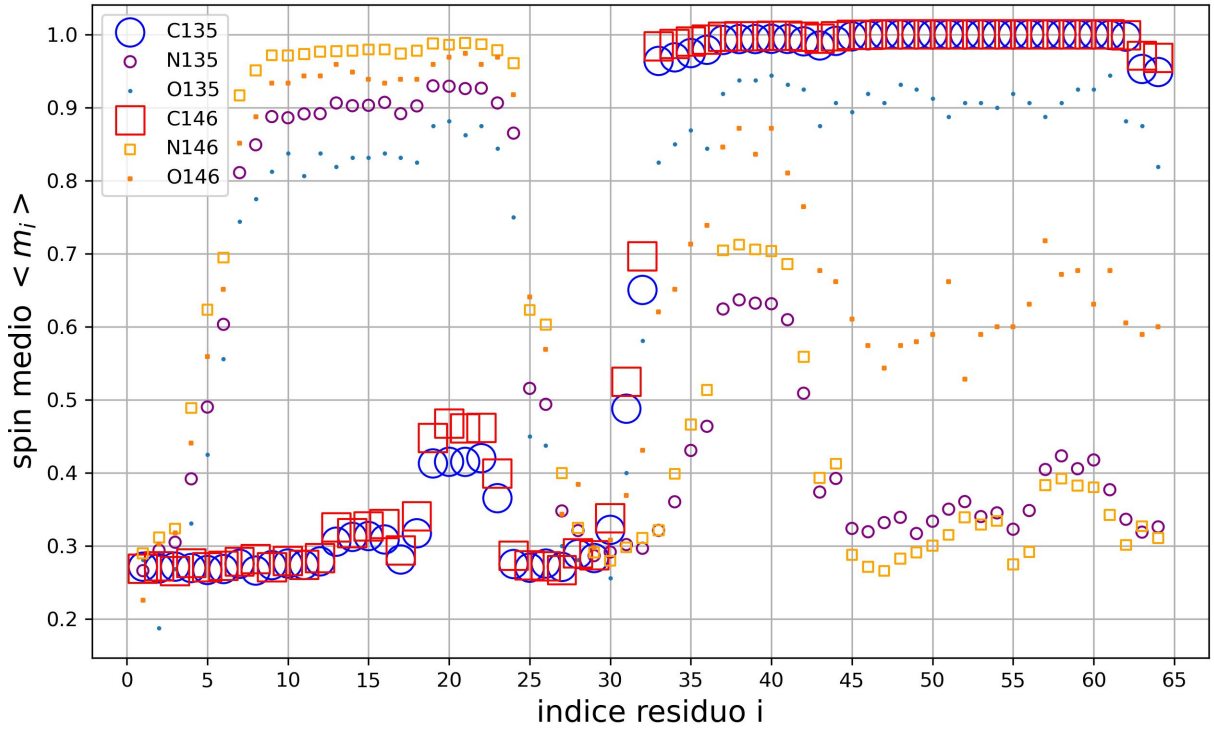

Figura 3.5: Confronto fra i due stadi intermedi nei due valori di λ. Il formato della legenda è "lettera identificativa+valore di λ" ("C135" indica la mappa di spin di C a $\lambda = 0.135$).

Il grafico 3.5 mostra la mappa di spin medi per i due intermedi e la classe "altro"/O. La dimensione dei punti delle simulazioni alla stessa λ è proporzionale alla frequenza di apparizione dell'intermedio

corrispondente all'interno delle 500 simulazioni. Si può osservare come esistano due porzioni di catena particolarmente stabili, intorno al residuo 18-21 e a quelli 37-41. Questi segmenti corrispondono a due α-eliche, in particolare quella di cui fanno parte gli aminoacidi 37-41 è la più lunga di tutta la proteina, e presenta dunque vari contatti fra aminoacidi vicini nella sequenza polipeptidica. Invece fra i residui 18 e 21 esiste un solo contatto, ma molto forte ($\eta_{18,21} = 20$). Questi fattori determinano la stabilità intrinseca dei due segmenti, e spiegano perché appaiano nello stato nativo (con frequenza relativamente alta) nell'intermedio di cui non farebbero parte.

L'analisi degli stati intermedi non restituisce altre configurazioni degne di nota oltre alle due suddette. Nella ricerca di N e C sono comparsi in realtà anche degli altri intermedi, inseriti nella classe "O". Osservando le mappe di spin, gli intermedi O si possono ricondurre ad anomalie di derivazione statistica, per i seguenti motivi:

1. Gli stati O si possono spiegare come deviazioni o combinazioni dei due intermedi già descritti. Alcuni sono una sorta di "N/C incompleto", e si possono attribuire a code (meno frequenti) nella distribuzione delle mappe di spin. Questi corrispondono a valori di $Q(O) < Q(N)$ o $Q(O) < Q(C)$. Gli stati O a $Q(O) > Q(C)$ invece corrispondono all'unione di stati N e C.
2. La frequenza di apparizione di O è minore di quella di N e C ed è di durata nettamente inferiore. Inoltre la correlazione nella comparsa di O e N/C è alta (O non compare quasi mai senza alcuno degli altri intermedi, v. tabella 1).

Nelle due famiglie di simulazioni a $\lambda_{nat} = 0.135$ e $\lambda_2 = 0.146$ si osserva che gli stati intermedi coincidono, anche se con una stabilità (una $\langle m_i \rangle$) leggermente minore a λ_2. Quindi si può studiare il CF con la sicurezza che gli intermedi restino gli stessi anche a temperature più prossime alla λ_F.

Ci interessa studiare gli stati intermedi per verificare se costituiscono delle scorciatoie nel ripiegamento. La presenza di stati intermedi che che inducano ripiegamento in tempi minori (agendo da "guida cinetica") potrebbe indicare che il CF sia più rapido del *folding in vitro*. In realtà, dalle simulazioni riportate nella tabella 1 e in figura 3.6 emerge l'opposto: i casi in cui compaiono gli intermedi hanno un tempo medio di ripiegamento maggiore rispetto a quelli che accedono direttamente allo stato nativo. Questo è compatibile con l'idea che gli intermedi C ed N non siano parte di una guida verso lo stato nativo, ma costituiscano piuttosto delle "trappole" cinetiche. Nello specifico, il tempo medio di ripiegamento τ_{avg} è minimo in assenza di intermedi, e cresce in base alla comparsa di N, C o entrambi, in quest'ordine (si veda la tabella 1).

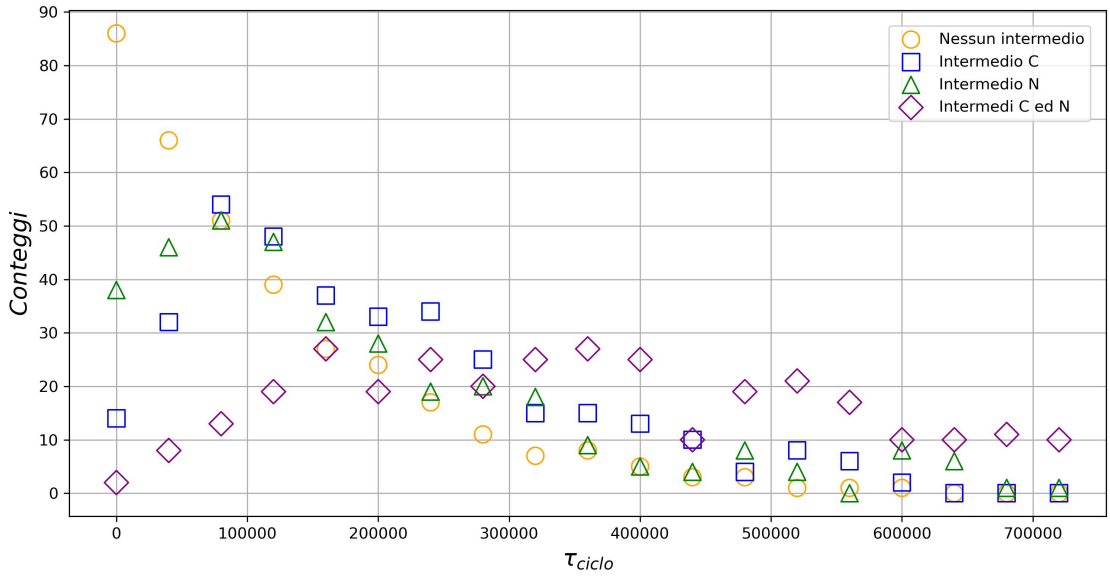

Figura 3.6: Distribuzione dei tempi di ripiegamento a seconda della presenza di intermedi. Bin da 40k passi.

Capitolo 4

Ripiegamento cotraslazionale

Le simulazioni per il *folding* cotraslazionale sono modellate come segue. Si definisce la variabile τ_{append}, pari al numero di passi attesi per l'inserimento di un nuovo aminoacido nella catena. Questa è un'approssimazione, dato che in natura il ribosoma sosta per un tempo maggiore in corrispondenza dell'inserimento di aminoacidi rari nella sequenza in formazione ([3], p.242). Il peptide di partenza consta già di 10 aminoacidi, perché il primo contatto che può formarsi è quello fra gli aminoacidi 7 e 11: partire da 10 aminoacidi o 1 avrebbe lo stesso profilo di energia efficace (senza componente sui contatti), quindi la scelta serve a risparmiare tempo computazionale. Dopo 54 iterazioni, corrispondenti a una fase di prolungamento, il sistema si ritrova sostanzialmente nelle stesse condizioni del *folding in vitro*. Si eseguono 500 simulazioni da 10^6 passi Metropolis per ogni valore di τ_{append}, con condizioni iniziali $\langle M\rangle_{ini} = 0.5$ ($\langle Q\rangle_{ini} \approx 0$ dato che si parte con pochi aminoacidi). Dalle simulazioni della sezione 3.1 si può dedurre che il τ_{ciclo} è sufficientemente grande da permettere il raggiungimento dell'equilibrio, infatti la catena ripiega nel 97% dei casi circa, tenendo conto del "tempo di assemblaggio".

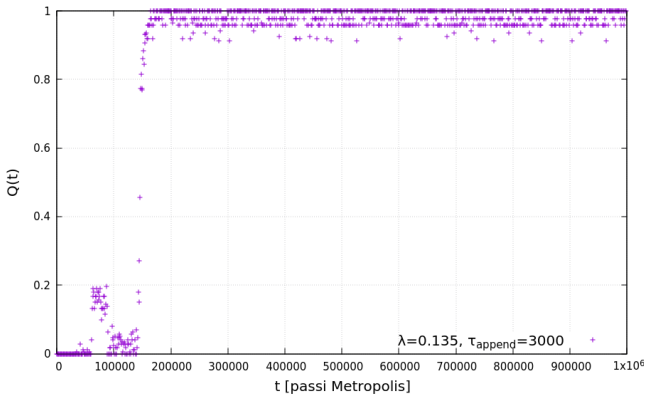

Figura 4.1: CF con intermedio N

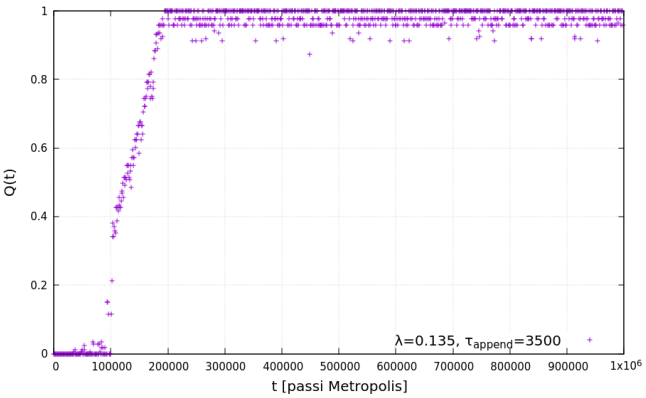

Figura 4.2: CF senza intermedio

Anche in questo caso si esegue la ricerca di stati intermedi. Qui si osserva che in alcuni casi compare una nuova specie di intermedio rispetto al caso del *folding in vitro*. I plot delle simulazioni mostrano come nel caso compaia un "intermedio di CF", questo consiste nel ripiegamento totale della catena parziale: non si potrebbe dunque paragonare agli intermedi del *folding in vitro*. Una volta che si genera, un "nucleo di condensazione" (catena parziale) di spin nativi non si scioglie più, e gli spin aggiunti successivamente assumono valore nativo ($m_i = 1$) molto rapidamente. Il tempo di ripiegamento τ_{avg} coincide con il tempo di prolungamento totale[1], $\tau_{avg} \cong 52\tau_{append}$. Si ottiene un τ_{avg} ridotto rispetto al *folding in vitro*, (v. tabella 2) con una bassissima deviazione standard: nel caso $\tau_{append} = 3000$, $\tau_{avg} = (156.10 \pm 0.09) \cdot 10^3$ passi, cioè una fluttuazione relativa dello 0.06%. In pratica il CF impone un'unica via privilegiata di ripiegamento, che procede in maniera ordinata dall'N-terminale al C-terminale.

[1] Lo stato F viene dichiarato raggiunto a $Q > 0.95$, quindi leggermente prima dell'aggiunta dell'ultimo aminoacido (54 iterazioni di τ_{append}).

Confrontando i tempi medi di ripiegamento nelle simulazioni in cui compare oppure no, si può notare come il CF sia più conveniente (il tempo di ripiegamento è minore) del *folding in vitro* per $\tau_{append} < 5000$. Nei casi in cui si manifesta il CF, il tempo di *folding* resta lo stesso indipendentemente dalla comparsa dell'intermedio N; l'intermedio C non appare dato che che presuppone l'estremità N non nativa. Quindi si potrebbe dedurre che gli intermedi per lo meno non accelerino il processo di ripiegamento, in parziale accordo con quanto accennato alla fine della sezione 3.2. Un altro risultato degno di nota è che nonostante convenga a τ_{append} più basse, la frequenza di apparizione del CF cresce con τ_{append}. Come caso aggiuntivo, il CF viene implementato nella modalità "inversa", ossia partendo dal C-terminale. Qui c'è un'asimmetria data dal fatto che gli ultimi 10 aminoacidi presentano svariati contatti, infatti il tempo di *folding* nei casi non-CF è significativamente maggiore nel CF "diretto". Questa asimmetria però sarebbe ineliminabile anche partendo da 1 aminoacido: è un dato di fatto che il C-terminale è più ricco di contatti nativi. Come per il CF diretto, anche nel metodo inverso si osserva un ridotto tempo di *folding* a bassa varianza che cresce linearmente con τ_{append}, ma con frequenze significativamente minori rispetto al primo caso. Questo comporta un tempo medio di *folding* maggiore nel caso inverso. Il tutto sembra plausibile dato che si verifica che il metodo osservato in natura è anche quello più economico in termini di tempo. Sembra ragionevole supporre che la probabilità P_{CF} di innescare il CF sia proporzionale al tempo speso nella fase di prolungamento. Come dalla sezione 2.2,un tempo maggiore equivale a un maggiore numero di passi spesi a esplorare lo spazio delle fasi della catena ridotta, e i risultati ottenuti fanno credere che questa possegga un grado di stabilità sufficiente a conferirle proprietà di *primer* (cioè di fungere da innesco per il ripiegamento) e velocizzare poi il ripiegamento della proteina intera.

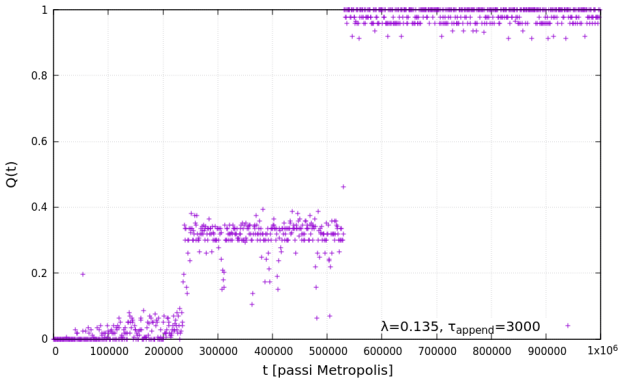

Figura 4.3: Simulazione senza innesco di CF. Si notino le maggiori fluttuazioni in Q nello stato U dovute alla presenza di una catena incompleta.

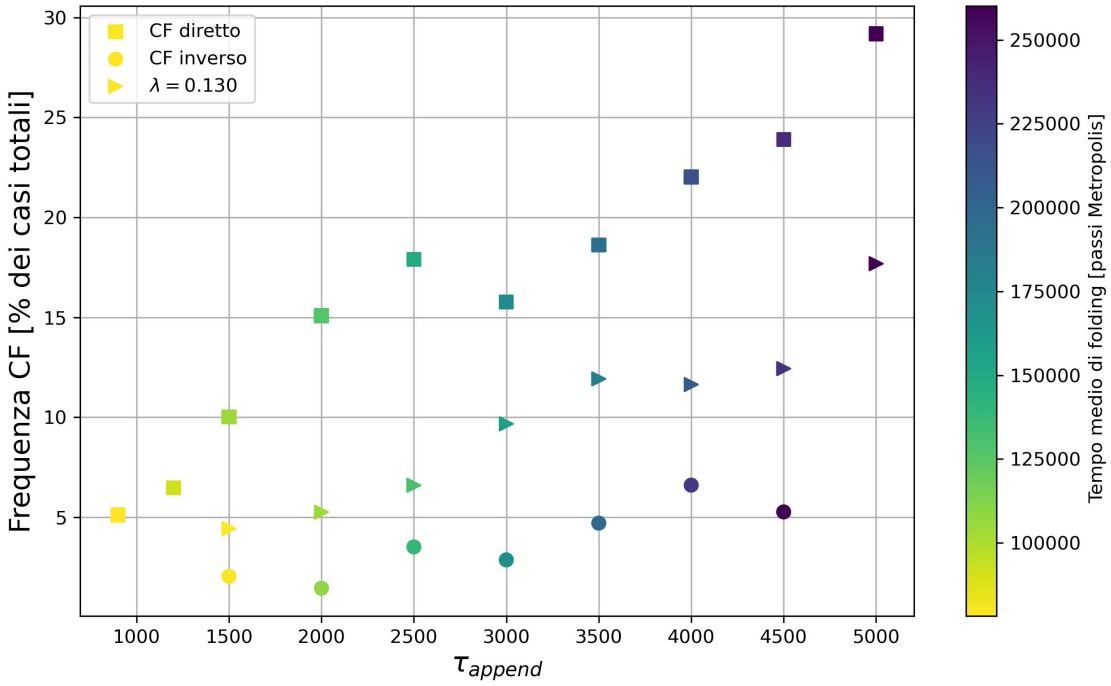

Figura 4.4: Confronto CF diretto e inverso

Nonostante tutto, calcolando il tempo medio di ripiegamento su tutte le simulazioni si nota che la modalità favorita è sempre il *folding* in vitro. La modalità *in vivo* è chiaramente sfavorita dal tempo speso in prolungamento, visto che nella maggior parte dei casi non conduce all'innesco del CF. Secondo [8], nel CF i meccanismi biologici della cellula abbassano l'entropia oltre i livelli normalmente accessibili *in vitro*, creando una sorta di ascensore nell'energia libera che permette di accedere in minor tempo agli stati nativi più ordinati (ad esempio compiendo movimenti di torsione sulla catena che sarebbero impossibili se isolata; passando per questi angoli "innaturali" sarebbe quindi possibile raggiungere più velocemente lo stato nativo). Quindi per ottenere una produzione *in silico* di proteine che sia biologicamente accurata, occorrerebbe prendere in considerazione le interazioni con l'ambiente cellulare. L'idea di Mushegian prende origine dall'osservazione che il campione di proteine selezionate per le simulazioni di *folding in vitro* possiede un *bias* per le sole - semplici - proteine che sono in grado di eseguire *refolding in vitro* spontaneamente. Le simulazioni eseguite in questo lavoro hanno mostrato che anche per proteine semplici - per cui non dovrebbe costituire un vantaggio così apparente - il CF può costituire una valida via di ripiegamento. Il tutto con un modello molto semplificato. Infatti, bisogna ricordare che il modello implementato non considera nessuna differenza fra i casi del ripiegamento *in vitro* e *in vivo*: λ resta la stessa. Di fatto la temperatura e ϵ non vengono toccati, il *folding in vivo* differisce da quello *in vitro* solo per l'aggiunta della fase di prolungamento. Se già *in vitro* si tiene conto solo in modo "efficace" dell'interazione con l'acqua e il denaturante, nel caso del *folding in vivo* gli agenti biomolecolari da tenere in considerazione sono ancora maggiori, e risulta ancora meno plausibile che siano ben rappresentate nel modello (ad esempio tramite ϵ).

Non sappiamo se a temperature diverse i risultati siano attendibili. Bisogna scegliere il giusto equilibrio tra una temperatura più alta possibile (per avvicinarsi a T_F) e un τ_{ciclo} più basso possibile (per non allungare eccessivamente le simulazioni). La percentuale di catene che ripiegano cala drasticamente con il decrescere di λ, come si vede in figura 4.4. Questo è compatibile con i risultati della sezione 3.1 che prevedono un aumento del tempo medio di ripiegamento e della frequenza di stati intermedi (si veda anche la 1).

Capitolo 5

Conclusioni

Questa tesi ha presentato un modello 1D per il ripiegamento di una proteina. Partendo da simulazioni di *folding in vitro*, è stata individuata una λ_{nat} adatta alle simulazioni di CF, trovando che la presenza di intermedi rallentava il processo di ripiegamento. Questi ostacoli sono stati superati passando a una rappresentazione semplificata di CF, che sembra offrire l'esistenza di una via privilegiata per il ripiegamento. Tuttavia a causa delle approssimazioni insite nel modello non è possibile determinare le condizioni ambientali per massimizzarne la resa. Pur trascurando molti degli agenti biochimici che dovrebbero rendere ancora più efficiente il CF, si osservano comunque dei benefici in una categoria di proteine per cui tali benefici dovrebbero essere minimi. Futuri sviluppi potrebbero ripartire da modelli più accurati, o nella complessità del sistema studiato, come uno SBM fuori reticolo in 3D, o nella fedeltà all'ambiente con cui interagisce, ad esempio cercando di implementare l'interazione della proteina con il *nascent chain welcoming committee*. In aggiunta si potrebbe testare il modello con una verifica sperimentale, anche per capire come le semplificazioni adottate influenzino le previsioni delle simulazioni. Per esempio, in [13] si riporta come il tasso medio di prolungamento della proteina nel ribosoma sia di circa 5 aminoacidi/secondo nelle cellule eucarioti, corrispondente a un τ_{append} molto maggiore di quelli studiati: il modello WSME ha previsto un rapporto $\tau_{append}/\tau_{fv} \ll 1$ come condizione favorevole all'insorgere del CF. Resta allora aperto il quesito sull'identità del meccanismo che favorisce il CF nell'ambiente cellulare.

Bibliografia

[1] Douglas R Hofstadter. *Gödel, Escher, Bach*. Basic books New York, 1979.

[2] Andrea Camerra. «Tesi di laurea triennale: Cinetica di ripiegamento su stati nativi topologicamente complessi». In: *Università degli studi di Padova* (2019).

[3] Alexei V Finkelstein e Oleg Ptitsyn. *Protein Physics*. Elsevier, 2002.

[4] Ken A Dill e Justin L MacCallum. «The protein-folding problem, 50 years on». In: *science* 338.6110 (2012), pp. 1042–1046.

[5] Amit Kessel e Nir Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. Chapman e Hall/CRC, 2018.

[6] Paul C Whitford, Karissa Y Sanbonmatsu e José N Onuchic. «Biomolecular dynamics: order–disorder transitions and energy landscapes». In: *Reports on Progress in Physics* 75.7 (2012), p. 076601.

[7] Shoji Takada. «Gō model revisited». In: *Biophysics and physicobiology* 16 (2019), pp. 248–255.

[8] Irina Sorokina e Arcady Mushegian. «Modeling protein folding in vivo». In: *Biology Direct* 13.1 (2018), pp. 1–14.

[9] Marco Baiesi et al. «Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding». In: *Scientific Reports* 9.1 (2019), pp. 1–12.

[10] Marco Zamparo e Alessandro Pelizzola. «Nearly symmetrical proteins: Folding pathways and transition states». In: *The Journal of chemical physics* 131.3 (2009), 07B609.

[11] Benjamin A Stickler e Ewald Schachinger. *Basic concepts in computational physics*. Springer, 2016.

[12] Nicholas Metropolis et al. «Equation of state calculations by fast computing machines». In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[13] Marija Liutkute, Ekaterina Samatova e Marina V Rodnina. «Cotranslational folding of proteins on the ribosome». In: *Biomolecules* 10.1 (2020), p. 97.

[14] *Type III Antifreeze Protein RD1 from an Antarctic Eel Pout*. URL: `https://www.rcsb.org/structure/1ucs`. (accessed: 15.11.2022).

Risultati delle simulazioni

$r = 100$	$\lambda = 135$	$r = 500$	$\lambda = 135$	$\lambda = 146$
n_{tot}	97	n_{UF}	68	117
$n_C(Q)$	59	$\tau_{UF}[10^3$ passi]	124	87
$n_C(m)$	80	n_N	155	190
$C_{count}(Q)$	4237	$\tau_N[10^3$ passi]	202	102
$C_{corr}(Q,m)$	4065	N_{durata}	11	8.2
$C_{corr}(Q,m)/C_{count}$	0.96	n_C	115	165
$n_N(Q)$	60	$\tau_C[10^3$ passi]	255	98
$n_N(m)$	84	C_{durata}	99	56
$N_{count}(Q)$	678	n_{NC}	149	27
$N_{corr}(Q,m)$	497	$\tau_{NC}[10^3$ passi]	406	102
$N_{corr}(Q,m)/N_{count}$	0.73	n_{tot}	487	499
		$\tau_{fv}[10^3$ passi]	266	97
		n_O	134	97
		O_{durata}	2.4	2.0
		n_{CO}	84	34
		n_{NO}	115	70
		n_{NCO}	67	10

Tabella 1: Analisi intermedi

La tabella riporta i risultati dell'analisi degli stati intermedi, con il test di identificazione (confronto dei metodi basati su Q e sugli spin m_i) nella metà di sinistra e a destra i dati di apparizione degli intermedi a diverse λ. La legenda, nella prima: n_{tot} è il numero di simulazioni in cui si raggiunge lo stato nativo; $n_I(Q,m)$ è il conteggio delle simulazioni in cui appare l'intermedio I ($I \in \{N, C\}$); $I_{count}(Q)$ è il numero di istanze in cui la funzione ricerca intermedi (metodo-Q) ha trovato l'intermedio I all'interno di tutte le simulazioni; $I_{durata} = I_{count}(Q)/n_I$; $I_{corr}(Q,m)$ conta le volte in cui il metodo-Q e il metodo-m danno entrambi esito positivo nella ricerca di Q. Nella seconda: τ_{UF} è il tempo di ripiegamento di simulazioni senza intermedi, τ_{fv} è la media del tempo di ripiegamento su tutte le simulazioni di *folding in vitro*, indipendentemente dalla comparsa di intermedi; n_{OI} è il numero di volte in cui O appare insieme a I.

Sorgenti delle figure
La figura 1.1 è presa dalla pagina del Protein Data Bank su 1ucs ([14]), la figura 1.2 viene da [6], la figura 1.3 viene da `https://arxiv.org/pdf/1810.10481.pdf` e la 1.4 da [8].

τ_{append} [passi]	τ_{fv} [10^3 passi]	n_{fv}	τ_{CF}	n_{CF}	n_{CF}/n_{tot} [%]	τ_{avg}	n_{tot}
CF diretto							
300	259	489	0	0	0	259	489
600	275	490	0	0	0	275	490
900	307	464	47	25	5.11	294	489
1200	316	449	63	31	6.46	299	480
1500	332	440	78	49	10.02	306	489
2000	377	411	104	73	15.08	336	484
2500	374	399	130	87	17.90	330	486
3000	393	406	156	76	15.77	356	482
3500	416	389	182	89	18.62	373	478
4000	442	379	208	107	22.02	390	486
4500	459	363	234	114	23.90	405	477
5000	498	342	260	141	29.19	429	483
CF inverso							
1500	312	480	77	10	2.04	307	490
2000	323	479	102	7	1.44	320	486
2500	346	468	128	17	3.51	338	485
3000	352	475	153	14	2.86	346	489
3500	379	466	179	23	4.70	370	489
4000	414	453	204	32	6.60	400	485
4500	440	450	230	25	5.26	429	475
$\lambda = 0.130$, CF diretto							
1500	410	389	78	18	4.42	396	407
2000	458	361	104	20	5.25	439	381
2500	487	340	130	24	6.59	464	364
3000	497	327	156	35	9.67	464	362
3500	497	340	182	46	11.92	460	386
4000	542	319	208	42	11.63	503	361
4500	572	317	234	45	12.43	530	362
5000	566	284	260	61	17.68	512	345

Tabella 2: Conteggi n e tempi di ripiegamento (in migliaia di passi Metropolis) per il CF diretto e inverso a $\lambda = 0.135$, diretto a $\lambda = 0.130$. Il tempo medio di ripiegamento per il *folding in vitro* a $\lambda = 0.135$ è di $\tau_{fv} = 266k$ passi.