UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Corso di Laurea in Fisica

Tesi di Laurea

Frustrazione residua in strutture di proteine topologicamente complesse

Relatore

Prof. Antonio Trovato

Laureando

Giorgio Bonollo

Anno Accademico 2019/2020

Indice

Abstract

La presente tesi ha come obiettivo principale lo studio della relazione tra frustrazione e complessità topologica all'interno delle proteine. Partendo dalla struttura proteica, si parlerà del problema ancora aperto del folding delle proteine, evidenziando il ruolo chiave della frustrazione. Si discuterà poi delle principali complessità topologiche presenti all'interno delle proteine, come nodi e *loop*, e della loro struttura. Si andranno infine ad analizzare i risultati computazionali riguardanti la frustrazione. L'analisi del rapporto tra frustrazione e complessità topologica porterà a risultati interessanti sia nell'andamento globale della frustrazione, sia nell'andamento particolare di ciascuna proteina.

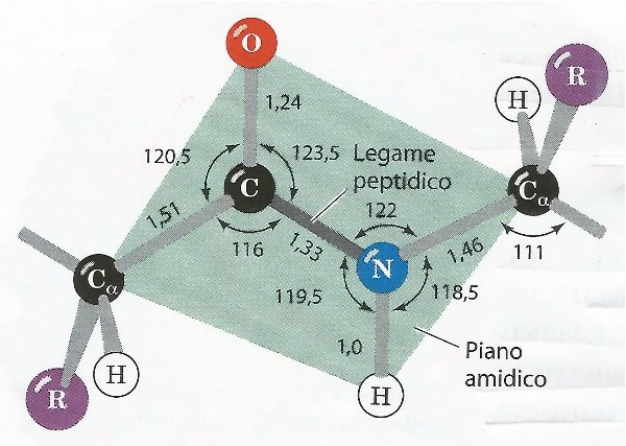
1. Proteine: struttura e folding

1.1 La struttura delle proteine

Le proteine sono polimeri composti da una catena di unità monomeriche più piccole, dette amminoacidi, che si ripiegano dando origine a strutture intricate e di grande importanza biologica[1]. Moltissime funzioni della cellula infatti sono svolte dalle proteine grazie alla loro struttura nativa e la loro complessità strutturale è ancora oggi oggetto di intensi studi scientifici. Lo scopo di questa tesi è cercare e verificare un legame tra complessità topologica nella struttura proteica e frustrazione energetica all'interno delle proteine.

Come detto sopra, le proteine sono composte da sequenze più o meno lunghe di amminoacidi. Gli amminoacidi in natura sono 20, e su questo codice si basa l'intera variabilità delle strutture proteiche. Tutti gli amminoacidi hanno una struttura comune (***Figura 1***): sono formati da un atomo di carbonio, detto C_α al quale sono legati un atomo di idrogeno ($-H$), un gruppo amminico ($-NH_2$), un gruppo carbossilico ($-COOH$) e un residuo (R). Mentre i primi tre sostituenti sono uguali in tutti gli amminoacidi, il quarto, cioè il residuo, è ciò che rende gli amminoacidi uno diverso dall'altro. La classificazione e lo studio della struttura chimico-fisica dei residui sono fondamentali per la comprensione dell'organizzazione e della struttura proteica finale che è oggetto di questa tesi.

Figura 1 (Struttura amminoacidica di base con Carbonio, Ossigeno, Azoto, Idrogeno e Carbonio α; legame peptidico e struttura quasi planare[1])

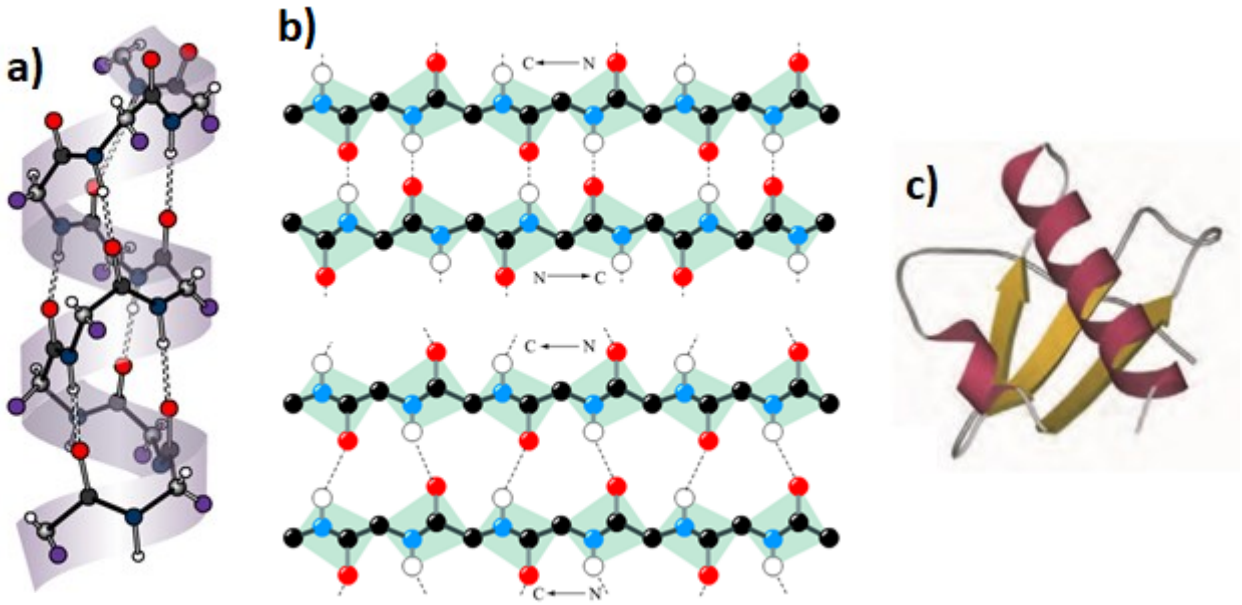


La breve descrizione sulla struttura base amminoacidica non spiega però come è possibile ottenere una sequenza finita di amminoacidi legati tra loro. Questo avviene tramite un complesso meccanismo di sintesi proteica che comincia dal DNA[1], dove il codice genetico viene tradotto in una sequenza di amminoacidi, che vengono legati tra di loro attraverso il legame peptidico (***Figura 1***). Questo legame coinvolge il gruppo amminico ($-NH_2$) e quello carbossilico ($-COOH$); attraverso una reazione di condensazione il C carbossilico si lega con l' N amminico liberando una molecola di H_2O [3]. Studi sulle proprietà di questo legame[4] hanno mostrato la sua "quasi" planarità, che rende di fatto il legame molto rigido e condiziona la flessibilità dell'intera catena. Un altro vincolo sulla struttura proteica finale è dato dalla struttura chimico fisica dei residui, che non possono occupare qualsiasi posizione disponibile a causa di ingombro sterico o interazioni chimiche energeticamente sfavorite dall'ambiente.

La caratteristica principale che emerge da una prima analisi sui residui, è che questi possono essere suddivisi in due tipi: polari e apolari. Ricordando che la maggior parte delle proteine è attiva nella cellula, il cui ambiente è chimicamente molto simile all'acqua, è evidente che in prima approssimazione la struttura base delle proteine sarà costituita da un core idrofobico e da una

superficie idrofila. Infatti, ogni altra soluzione sarebbe energeticamente sfavorita, e l'effetto idrofobico è alla base del ripiegamento delle proteine nella loro struttura nativa[3].
Un'ultima fondamentale peculiarità delle strutture proteiche è la struttura secondaria. Studiando nel dettaglio la disposizione spaziale degli atomi delle proteine si è notata la presenza di alcuni motivi ricorrenti: le α-eliche e i foglietti β[3,5]. Queste strutture secondarie sono caratterizzate dall'avere dei legami idrogeno in posizioni specifiche che rendono molto stabile quella data architettura della sequenza amminoacidica. I legami idrogeno si instaurano tra i gruppi $C = O$ e $H - N$ della catena amminoacidica. Se il legame si forma ripetutamente tra amminoacidi vicini (solitamente tra l'aminoacido *n* e quello *n+4*, ma sono possibili anche altre configurazioni), allora si ha la formazione di una α-elica; se invece i legami idrogeno si sviluppano tra intere porzioni della sequenza amminoacidica allora si assiste alla formazione di un foglietto β[3,5]. In entrambe queste configurazioni i residui laterali delle catene risultano "impacchettati" in maniera estremamente efficiente, rendendo così queste due strutture particolarmente favorite (***Figura 2***).

Figura 2 (Strutture secondarie: a) struttura atomica dell'α-elica[1], b) struttura atomica dei foglietti β paralleli e antiparalleli[2], c) esempio di proteina ripiegata con eliche e foglietti[6])



1.2 Il problema del protein folding

Da questa breve analisi preliminare emerge chiaramente una cosa: le proteine sono strutture eterogenee e molto complesse. La loro variabilità è molto ampia, dato che sono costruite su un codice di 20 lettere che potrebbe originare un'infinità di strutture valide. La sequenza è dettata dal codice genetico, la struttura nativa (cioè quella che si trova nella cellula) è determinata dalla sequenza, e la funzione di una proteina è a sua volta determinata dalla struttura[7]. Spesso la rappresentazione della proteina non fa chiarezza riguardo alla sua complessità. Le proteine infatti sono simili a cristalli, con un'organizzazione spaziale densa e rigida, perfettamente ordinata nonostante la sua eterogeneità[7]. Questa struttura densa, rigida e compatta è ancora oggi oggetto di studio. Infatti, ciò che è più sorprendente è che le proteine vengano prodotte nella cellula inizialmente come filamenti estesi e poi si organizzino man mano fino a formare una struttura solida che svolge diverse funzioni cellulari. Cyrus Levinthal cercò di stimare quale fosse il tempo necessario ad una catena di *n* amminoacidi per organizzarsi nella sua struttura nativa. Se si ipotizza che ogni amminoacido abbia solamente 2 possibili conformazioni spaziali (in realtà ne ha di più), allora per 100 amminoacidi si hanno $2^{100} \sim 10^{30}$ possibili conformazioni. Se ogni conformazione è sondata in

10^{-12} secondi (tempo tipico di orientazione di un legame), allora il tempo necessario a sondare tutte le possibili conformazioni è di circa 10^{18} s $\sim$ 10^{10} anni[8]. Sperimentalmente il tempo necessario affinché una proteina ripieghi nella sua struttura nativa è di circa un minuto. Allora è chiaro che la proteina non può sondare tutte le possibili configurazioni per trovare quella giusta. Questo argomento, noto come "Paradosso di Levinthal", è alla base del problema del *protein folding*, cioè capire come una sequenza di amminoacidi possa ripiegare nella struttura nativa in tempi relativamente brevi.

Evidentemente quindi, le proteine non esplorano tutte le possibili configurazioni prima di arrivare alla struttura nativa. Levinthal suggerì una teoria, ancora oggi comunemente accettata, per risolvere il paradosso[9,10], secondo cui la proteina segue uno specifico cammino per il ripiegamento (*folding pathway*), riducendo così sia la sua entropia che l'energia a disposizione. Il cammino deve essere abbastanza liscio affinché la proteina possa raggiungere una struttura stabile rapidamente, e anche stretto, in modo tale che il minimo sia energeticamente favorito[9]. In questo modo il *protein folding* diventa un processo altamente cooperativo tra le varie parti della proteina, nel quale il ripiegamento coordinato di ogni singola parte avvicina la proteina alla struttura nativa, limitando le possibilità di errori. Questa *Energy Landscape Theory* è rappresentata tramite un grafico con energia in ordinata ed entropia in ascissa. Appena formata nella cellula la proteina si trova nella parte superiore dell'"imbuto" (*energy-entropy funnel*), possiede cioè un'elevata energia perché le interazioni tra cellula e filamento esteso non sono ottimali (basti pensare alla parte idrofoba), ma possiede anche un'elevata entropia perché le configurazioni possibili sono innumerevoli. La forma dell'*energy funnel* evidenzia quindi alcune proprietà[10]:

- il *folding* prevede un abbassamento sia di energia che di entropia, infatti la proteina raggiunge la struttura nativa quando si trova nel minimo energetico globale;
- il processo prevede il progressivo raggiungimento di minimi locali dopo il superamento di alcune barriere energetiche (il *funnel* non è liscio). Evidentemente la Natura ha selezionato quelle proteine che riescono a superare le barriere energetiche velocemente e che evitano di rimanere intrappolate in minimi locali (vengono privilegiati *funnel* sufficientemente lisci);
- il *folding* può avvenire secondo diversi percorsi, che portano sempre alla struttura nativa[11].

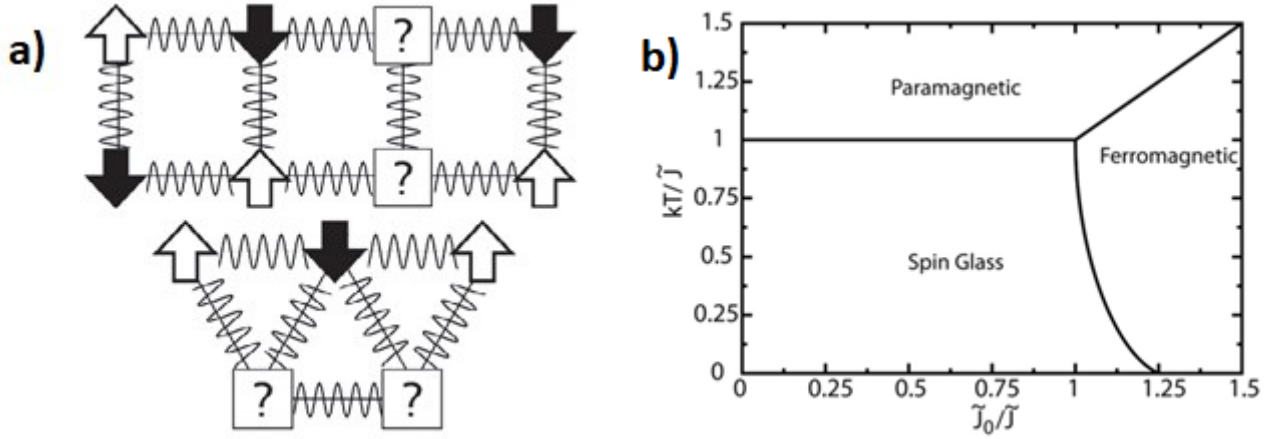
Questa analisi permette una maggiore comprensione della dinamica del *folding*, anche se più qualitativa che quantitativa. Da ormai quasi 60 anni il problema del *protein folding* continua a porre domande. Con l'avvento dei computer e con database sempre più aggiornati sulle strutture proteiche (circa 80 000 strutture di proteine sono conosciute con precisione atomica) si è cercato di simulare la dinamica che porta al *folding* attraverso diversi modelli. Tuttavia, la dinamica di un sistema così eterogeneo è molto complicata, ed è importante costruire teorie che da pochi parametri possano prevedere quanto una struttura sia favorita. Uno dei parametri importanti per la comprensione del *folding* è la cosiddetta frustrazione energetica.

1.3 La frustrazione

In un sistema frustrato, le interazioni tra le diverse componenti non possono essere ottimizzate simultaneamente, causando degenerazione negli stati a bassa energia. Il termine frustrazione contiene diverse sfaccettature a seconda dell'ambito in cui viene utilizzato. In fisica viene usato per descrivere l'approccio all'equilibrio di sistemi tipicamente disordinati che può protrarsi anche molto a lungo nel tempo. Il primo impiego interessante della frustrazione si ha, per esempio, nei vetri di spin. I vetri di spin sono magneti disordinati che, pur osservati in un largo periodo temporale, sembrano non raggiungere mai un punto di equilibrio[12]. Il parallelismo con le proteine è chiaro: si parla infatti di sistemi altamente disordinati che cercano di raggiungere l'equilibrio. È

particolarmente istruttivo vedere cosa accade per un vetro di spin prima di passare alla frustrazione per le proteine.
Si consideri un reticolo rettangolare; ad ogni punto è associato un valore di spin (±1). Il potenziale energetico più semplice per descrivere il sistema è $H = -\sum_{[i,j]} J_{ij} s_i s_j$, dove J_{ij} è il potenziale di interazione tra gli spin e $s_{i,j}$ sono i valori dei singoli spin, solitamente presi come adiacenti[8,12]. Per scelte semplici di J_{ij} il sistema ha un unico stato fondamentale: se, per esempio, $J_{ij} = 1$ l'energia minima si ha per spin che puntano nella stessa direzione, per $J_{ij} = -1$ le direzioni degli spin dovranno essere opposte. Tuttavia, la fisica dei sistemi disordinati non è così semplice; gli accoppiamenti fra spin adiacenti possono essere di entrambi i tipi ($J_{ij} > 0, J_{ij} < 0$) in parti diverse di un vetro di spin. Oppure, cambiando il reticolo da rettangolare a triangolare si vede chiaramente che non esiste una scelta che ottimizzi tutti gli accoppiamenti per $J_{ij} < 0$ (***Figura 3***). Se si prende una distribuzione di probabilità gaussiana per J_{ij}, $p(J_{ij}) = \frac{1}{\sqrt{2\pi}J} e^{-\frac{(J_{ij}-J_0)^2}{2J^2}}$ (con J_0 media e J dispersione) è possibile costruire un diagramma di fase con le quantità normalizzate $\widetilde{J_0} = NJ_0$ e $\tilde{J} = \sqrt{N}J$. Se l'energia termica kT è più grande sia della dispersione del valor medio degli accoppiamenti, allora il sistema è dominato dalle fluttuazioni termiche e diventa paramagnetico, se invece il valor medio degli accoppiamenti è grande rispetto sia alla dispersione, sia all'energia termica, gli spin si allineano facendo diventare il sistema ferromagnetico[8]. Se infine la dispersione è grande rispetto sia al valor medio, sia all'energia termica, allora il sistema diventa frustrato, cioè non è possibile trovare in tempi rapidi una configurazione energetica favorita, le interazioni non sono ottimali, o lo sono solo in piccole regioni di spazio.

Figura 3 (Frustrazione nei magneti: a) reticolo rettangolare non frustrato e reticolo triangolare frustrato[12], b) diagramma di fase in funzione di temperatura e distribuzione degli accoppiamenti fra spin[8])

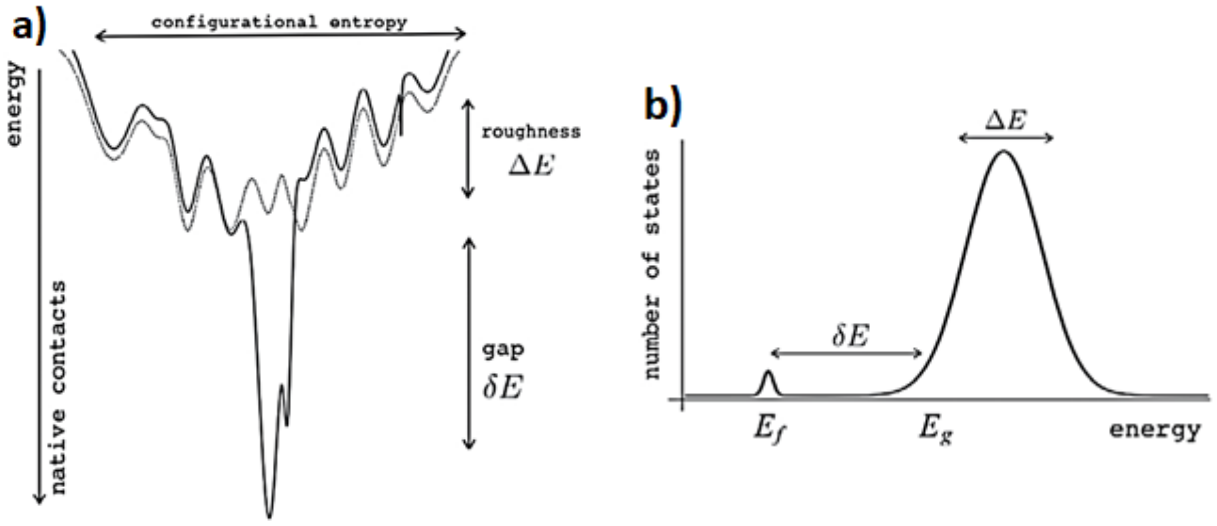


Il parallelismo con le proteine è immediato: l'eterogeneità è data dai 20 diversi tipi di amminoacidi. Si può identificare una transizione tra uno stato *"glassy"*, disordinato, ad uno ordinato e compatto. Inoltre, piccole zone energeticamente frustrate possono modificare di molto la dinamica del sistema perché questo deve rimuovere la degenerazione dovuta alla frustrazione per raggiungere uno stato energeticamente favorito. Tuttavia, la "rimozione" delle zone frustrate deve coinvolgere altre parti del sistema non frustrate; questo costa molto dal punto di vista energetico ed è difficilmente ottenibile solamente con il moto termico casuale[12]. Per questi motivi è importante considerare la frustrazione per i sistemi biologici.

1.4 La frustrazione nelle proteine

Le similitudini tra i vetri di spin e il *protein folding* hanno portato alla formulazione del principio di minima frustrazione. Approcciando il problema in modo simile a quanto fatto precedentemente, si nota subito la difficoltà di descrivere il sistema. Le scelte up/down degli spin sono molto più semplici della miriade di configurazioni spaziali possibili per una proteina. Per quantificare la frustrazione, anche in modo approssimato, è necessario scegliere un modello di rappresentazione spaziale per le proteine. Tra i modelli *"coarse-grained"*, che non considerano i gradi di libertà di ogni singolo atomo, quello più utilizzato consiste nel visualizzare la catena come successione di aminoacidi. In effetti questo modello è abbastanza accurato, se si pensa che l'evoluzione agisce per modifiche di singoli amminoacidi nella catena, più che per cambiamenti dei singoli atomi. Da questo punto di vista, la frustrazione è generata dal contrasto tra i vincoli di legame tra amminoacidi consecutivi lungo la catena, e l'attrazione efficace tra residui idrofobici che tende a portarli verso il nucleo della struttura nativa in contatto tra loro, evitando il contatto con l'ambiente acquoso della cellula[12]. Avendo a disposizione 20 tipi di amminoacidi fra cui scegliere, è stato possibile trovare delle sequenze che, malgrado la dinamica lenta, la frustrazione energetica e la degenerazione, ripiegassero in modo stabile e riproducibile in una data struttura nativa. Un esempio sono le sequenze amminoacidiche che tendono a formare spontaneamente strutture secondarie come foglietti β o α-eliche.

Figura 4 (Frustrazione ed *energy landscape*: a) abbassamento di energia ed entropia in funzione della crescita della struttura nativa[12], b) densità degli stati e gap energetico tra stato *folded* e *unfolded*[12])



Maggiore è il numero di tipi di amminoacidi fra cui scegliere, più velocemente il minimo globale dell'*energy landscape*[12]. Tuttavia, la rugosità del *funnel*, dovuta alla frustrazione, può rallentare il processo intrappolando la proteina in un minimo locale per molto tempo. Dunque, quando il sistema è poco frustrato, il *folding* è un processo altamente cooperativo che si svolge rapidamente e senza grossi rallentamenti. Viceversa, se la frustrazione cresce, il sistema presenta numerosi intermedi con un alto numero di contatti non nativi, e il raggiungimento del ripiegamento corretto non è immediato. Proteine minimamente frustrate mostrano quindi dinamiche simili al congelamento di un cristallo e al suo impacchettamento, mentre proteine altamente frustrate hanno un comportamento caotico e instabile simile a quello dei vetri di spin.

La capacità di una proteina di ripiegare più o meno velocemente è valutata dal principio di minima frustrazione, che può essere enunciato nel modo seguente. Data una proteina, è possibile determinare due energie: E_g, cioè l'energia più bassa della proteina ottenibile per una configurazione compatta non nativa, calcolata con metodi statistici partendo dallo spazio delle configurazioni, ed E_f, cioè l'energia dello stato ripiegato nativo (***Figura 4***). Per un polimero

abbastanza grande, possiamo assumere in prima approssimazione che la distribuzione delle energie non native sia una gaussiana $p(E) = \frac{1}{\sqrt{2\pi\bar{E}}} e^{-\frac{(E-\bar{E})^2}{2\Delta E^2}}$ con ΔE varianza e $\bar{E}$ media della distribuzione. Allora la frustrazione è minima se è soddisfatta la cosiddetta *gap condition*, ovvero se il termine $Z = \frac{(E_f-\bar{E})}{\Delta E}$ è elevato in modulo. Sostanzialmente il principio di minima frustrazione stabilisce che se l'energia dello stato finale, ripiegato e stabile, è molto più bassa della media delle configurazioni possibili (e nello specifico della configurazione E_g), allora il ripiegamento avviene in maniera rapida e ottimale[12]. Questa differenza va normalizzata sulla varianza della distribuzione delle energie non native, ed indica che lo stato ripiegato non solo è energeticamente favorito, ma è accessibile tramite un percorso specifico non casuale.

La connessione tra il principio di minima frustrazione e il *funnel energy landscape* è esplicita. Infatti, le interazioni che stabilizzano la struttura nativa sono energeticamente molto favorite rispetto ad altre interazioni non native, e le possibili *kinetic traps* non ne pregiudicano il corretto ripiegamento.

Ora che è stato compreso come le proteine riescono a ripiegare così facilmente, minimizzando la frustrazione, è necessario darne una definizione quantitativa. Si tratta cioè di capire quanto un legame, un contatto o una configurazione di amminoacidi sia favorita rispetto ad altre possibili. Per fare questo si può analizzare localmente una certa zona della proteina facendo delle mutazioni casuali degli amminoacidi coinvolti. Dopo aver campionato un numero statisticamente rilevante di mutazioni si procede al calcolo della frustrazione.

In questa tesi in particolare, il modello adottato prevede di cambiare la natura degli amminoacidi coinvolti in un certo contatto nella proteina, calcolando così la frustrazione nello spazio delle sequenze a struttura fissata. Le mutazioni sono casuali, ma ogni amminoacido viene campionato in base alla frequenza per esso osservata nella proteina in esame. Prendendo una coppia di amminoacidi in contatto, detti i, j la frustrazione sarà:

$$F_{i,j} = \frac{E^N_{i,j} - \langle E^M_{i',j'} \rangle}{\sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(E^{l,M}_{i',j'} - \langle E^M_{i',j'} \rangle\right)^2}} \qquad (\mathbf{1.1})$$

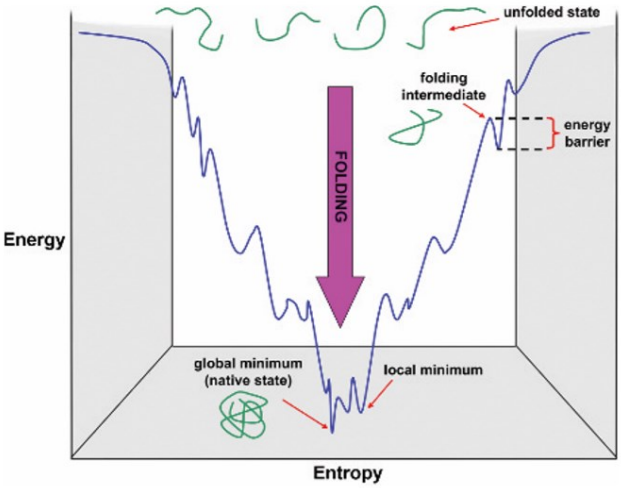
Con $E^N_{i,j}$ energia dei residui i, j in contatto nella struttura nativa, $\langle E^M_{i',j'} \rangle$ energia media ottenuta con i residui mutati, $E^{l,M}_{i',j'}$ energia della l-esima mutazione, N numero di mutazioni.

Da questa definizione[12] emerge come la frustrazione dica quanto effettivamente un contatto è energeticamente favorito rispetto a tutte le possibili combinazioni in una data sequenza. Questa definizione coinvolge anche altri residui vicini a quelli mutati in contatto tra loro, perché la mutazione modifica anche le altre interazioni formate dagli amminoacidi mutati. Un'altra peculiarità di questa definizione è che dipende solamente dalla struttura della catena e dalla natura degli amminoacidi; è quindi relativamente semplice e verrà utilizzata in questa tesi. Sono state introdotte altre definizioni di frustrazione basate, per esempio, su cambiamenti nella posizione spaziale degli amminoacidi.

Ci sono due aspetti fondamentali nel principio di minima frustrazione:

- proteine poco frustrate ripiegano in maniera veloce e riproducibile in una struttura nativa ben definita, mentre proteine molto frustrate non hanno una struttura nettamente favorita rispetto alle altre[12];
- la poca frustrazione richiede la presenza di una sorta di "gerarchia strutturale" nelle proteine; i diversi elementi di struttura nativa si devono formare nel giusto ordine, per non compromettere il *folding* corretto[8].

Figura 5 (*Energy landscape* e *folding*[10])



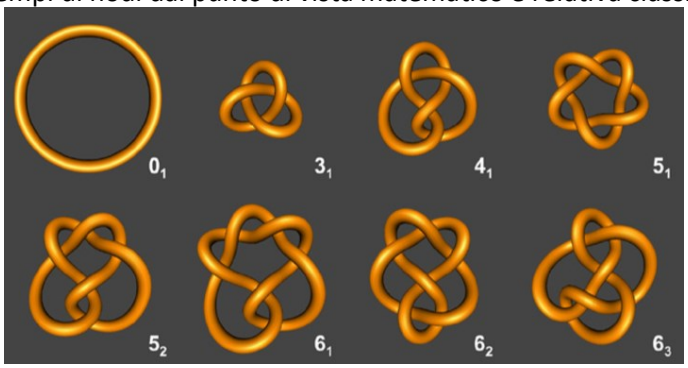
L'assenza di frustrazione permette quindi di guidare il ripiegamento verso una struttura energeticamente favorita. Il ruolo della frustrazione è ancora più importante nello studio di alcune strutture proteiche complesse: i nodi.

2. Nodi, *loop* e il ruolo dell'entanglement nella struttura proteica

2.1 I nodi nelle proteine

I nodi sono strutture intricate e complesse che si possono ritrovare molto spesso nel mondo macroscopico. Negli ultimi anni, con l'avvento di tecniche sperimentali più precise, sono state rivelate molte conformazioni topologicamente complesse anche all'interno di strutture biologiche microscopiche come DNA e proteine. Il ruolo di questi complessi avvolgimenti deve ancora essere ben chiarito e compreso, ma quello che risulta chiaro è che essi giocano un ruolo fondamentale sotto diversi punti di vista[13] (*folding* o struttura biologica per esempio). Matematicamente un nodo è definito come una curva chiusa che non può essere slegata senza essere tagliata o aperta; nel caso di una curva aperta il nodo è una struttura che non può essere slegata "tirando" gli estremi della curva[13].

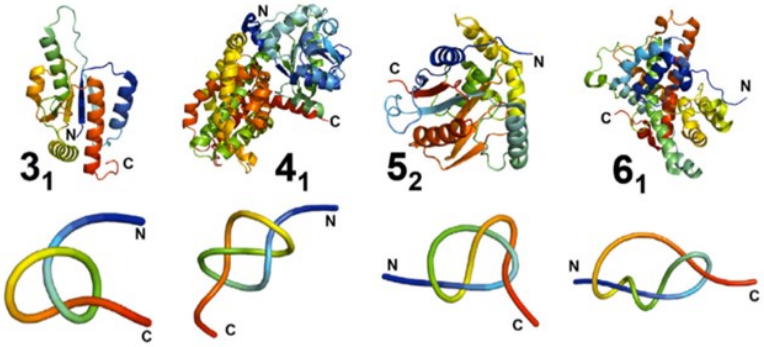
Figura 6 (Esempi di nodi dal punto di vista matematico e relativa classificazione[13])



Dal momento in cui la percentuale di proteine con la presenza di nodi (ad oggi si parla di circa il 6% delle strutture presenti nel PDB) è aumentata, si è cercato di trovare indizi sul loro ruolo biologico. Ricordando quanto detto sopra, le proteine sono strutturalmente paragonabili a cristalli; la presenza di struttura come nodi o "lacci" come quelle in figura diventa quindi difficile da spiegare. Molte domande sono tuttora senza risposta, e gli indizi su ruolo, stabilità, pressione evolutiva di queste

architetture biologiche sono contrastanti. Si sono fatte ipotesi sul miglioramento dell'attività catalitica dovuta a questi nodi, o alla loro importanza in struttura allosteriche, ma le prove sperimentali sono ancora poche.

Figura 7 (Esempi di nodi in strutture native[13])



Diversi studi riportano che proteine che contengono nodi sono più stabili dal punto di vista termodinamico ed è più difficile che raggiungano uno stato *unfolded* se confrontate con le rispettive strutture senza nodi[13]. I nodi aumentano anche la stabilità meccanica della proteina e sembrano dunque molto importanti per proteine particolarmente grandi e complesse.
Se quindi i nodi riescono a dare queste maggiori garanzie dal punto di vista meccanico e termodinamico, come riesce la proteina a ottenerle durante il percorso di *folding*? Come fa la proteina ad evitare un percorso di ripiegamento che la porterebbe verso possibili trappole cinetiche? Diversi studi hanno mostrato che:

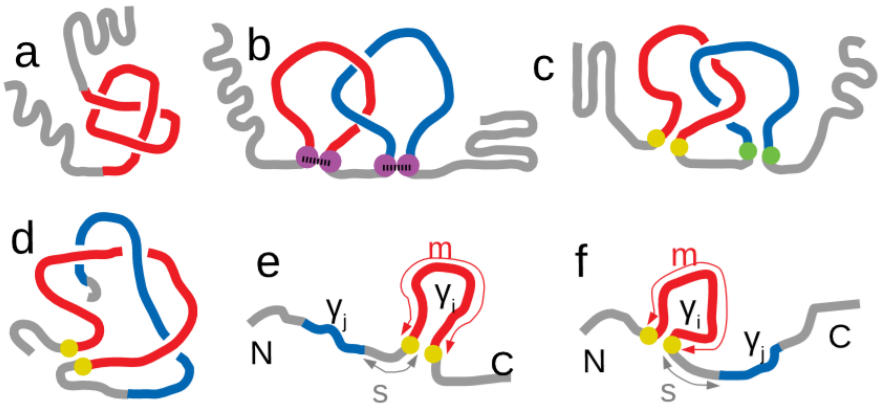
- i nodi hanno un importante ruolo sia nella funzione biologica della proteina che nella sua stabilità energetica[13];
- l'*energy landscape* associato a proteine con strutture *entangled* dovrebbe essere ricco di trappole cinetiche e questo è in contrasto con il cosiddetto *funnel* relativamente "liscio" previsto dal principio di minima frustrazione[13];
- il C terminale della catena amminoacidica risulta più spesso presente nella struttura *entangled* rispetto all'N terminale[14].

2.2 Loop e folding co-traslazionale

La complessità topologica delle proteine, come detto sopra, può risultare molto elevata. Comprendere la struttura proteica sino in fondo è essenziale per capire la funzione di ogni singola struttura. Per conoscere l'organizzazione spaziale delle proteine e le interazioni fondamentali tra gli amminoacidi della catena un metodo di analisi interessante è quello dei contatti[15,16]. Per contatti si intendono quelle coppie di amminoacidi che sono distanti tra loro lungo la catena amminoacidica, ma vicini nella struttura nativa ripiegata. Tuttavia, proprio a causa delle molteplici conformazioni possibili, la mappa dei contatti risulta talvolta poco utile. Inoltre, come visto sopra, la presenza di strutture *entangled* restringe di molto i possibili cammini per il *folding*. Lo studio di queste ultime potrebbe fornire interessanti particolari su come le proteine riescano a raggiungere, dopo la sintesi, la forma corretta nella cellula che permette loro di svolgere la loro funzione biologica.
Per questo motivo sono state fatte analisi più approfondite sui cosiddetti *loop* nelle proteine[15], cioè la porzione di catena fra due amminoacidi in contatto. Analizzando le caratteristiche dei contatti sono state rivelate diverse strutture come quelle in ***Figura 8***, quantificandone l'*entanglement* e la loro posizione lungo la catena. L'indice di *entanglement* gaussiano G_c' è una generalizzazione del *linking number* per una curva chiusa; più questo indice è alto in modulo, più il *loop* analizzato risulta topologicamente complesso in relazione ad altre porzioni di catena. Studiando la natura e la disposizione spaziale degli amminoacidi coinvolti nel *loop*, è possibile trarre alcune conclusioni

riguardo la sua frustrazione. Infine, rispetto all'altra porzione di catena con cui si attorciglia si può verificare se il *loop* sia più vicino al terminale N o C della catena. Da quest'analisi si verifica o meno la possibilità di un *folding* co-traslazionale (se il ripiegamento avviene già durante la sintesi proteica ci si aspetta che strutture complesse non siano troppo vicine all'N-terminale che viene sintetizzato per primo). Questi risultati verranno messi a confronto con misure sulla frustrazione dei *loop* e analizzati in dettaglio per studiare l'eventuale correlazione tra i parametri G_c' e F.

Figura 8 (Esempi di nodi e loop (a, b, c) e di strutture *loop-thread* (d), con N-thread (e) e C-thread (f)[15])



2.3 L'analisi dell'indice di entanglement

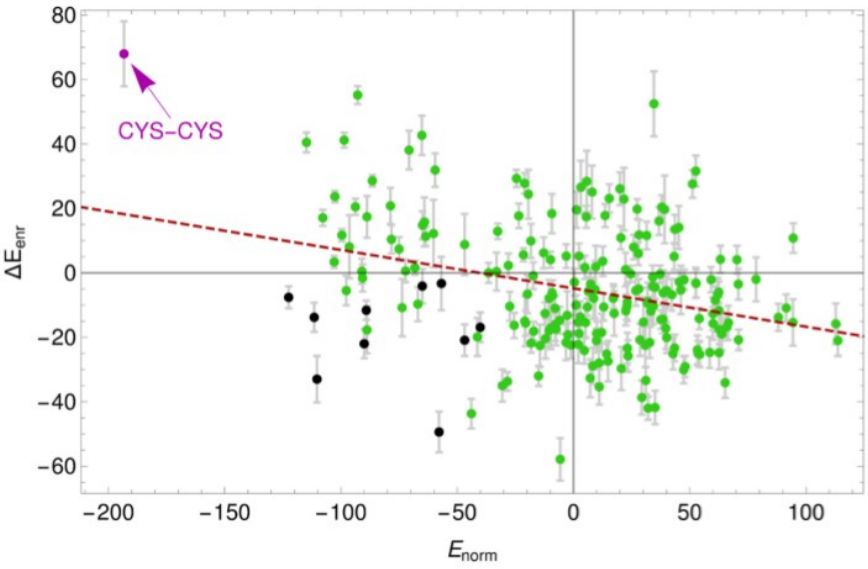
I risultati[15] che sono riportati ora sono il punto di partenza del presente lavoro di tesi. L'analisi approfondita sull'indice di *entanglement* è stata eseguita su circa 17 000 proteine presenti nel PDB. Attraverso uno studio strutturale è possibile trovare all'interno delle proteine diverse conformazioni *entangled* (***Figura 8***), che vanno dai nodi a *loop* concatenati. In particolare, si è focalizzata l'attenzione su strutture del tipo *loop-thread*, cioè sequenze in cui fosse possibile trovare una porzione di catena con un *loop* γ_i e un "filo" γ_j intrecciato con esso. Dopo aver trovato il contatto, quindi il *loop*, si massimizza l'indice $|G_c'|$ sui possibili *thread* e si determina l'indice di *entanglement* per il contatto; massimizzando $|G_c'|$ sui contatti si trova il valore di *entanglement* per la proteina (nel seguito $|G_{c-max}'|$).

Se si ha $|G_c'| > 1$, allora il *loop* viene detto *entangled*, e si è trovato che circa il 32% delle proteine analizzate possiede almeno un *loop* di questo tipo. Studiando accuratamente la posizione relativa tra *loop* e *thread* si è visto, tramite analisi statistiche, che le sequenze favorite sono gli *N-thread*, cioè sequenze dove il *thread* è tra il terminale N della catena e il *loop*. Questa asimmetria si trova scegliendo gli indici che massimizzano $|G_c'|$; ciò suggerirebbe che strutture di questo tipo sono favorite per la proteina. Nel caso di *folding* co-traslazionale infatti, è ragionevole pensare che per la proteina risulti più "semplice" sintetizzare un *thread* e avvolgerci intorno un *loop*, piuttosto che il contrario.

Sono state svolte anche analisi sulla natura energetica dei contatti tra gli amminoacidi alle estremità del *loop*. Usando un potenziale statistico che tiene conto delle frequenze dei contatti tra i diversi amminoacidi (E_{norm}), si è calcolata l'energia di ogni contatto (si calcola per esempio la frazione con cui contatti del tipo CYS-CYS avvengono nell'intero set di proteine e da questo si calcola E_{norm}). Se le frequenze analizzate sono invece relative ai soli *loop entangled*, si ottiene E_{GE}. Lo studio della correlazione tra E_{norm} e ΔE_{enr} $(= E_{GE} - E_{norm})$ calcolati per coppie (a, b) di amminoacidi in contatto, ha dato un esito interessante. Si è visto che (r di Pearson pari a -0,31) ΔE_{enr} ed E_{norm} sono anticorrelati in maniera statisticamente significativa (***Grafico 1***). Questo si traduce nel fatto che le interazioni fra gli amminoacidi agli estremi di *loop entangled* sono mediamente meno favorevoli

rispetto a quelle fra amminoacidi agli estremi di *loop* generici, suggerendo quindi che le interazioni fra i contatti *entangled* risultino essere energeticamente frustrati.
Questa tesi si pone proprio l'obiettivo di stabilire un legame tra questi risultati e il principio di minima frustrazione, analizzando nel dettaglio la relazione tra gli indicatori F e G_c'.

Grafico 1 (Anticorrelazione tra ΔE_{enr} e E_{norm}[15])



3. Analisi dati

3.1 Scopi e definizioni

Lo scopo di questa analisi è lo studio della relazione tra i due indicatori di frustrazione, F, e di *entanglement* G_c'. L'indice di *entanglement* G_c' è definito a partire dal doppio integrale di Gauss

$$G \equiv \frac{1}{4\pi} \oint_{\gamma_i} \oint_{\gamma_j} \frac{r^{(i)} - r^{(j)}}{|r^{(i)} - r^{(j)}|^3} \cdot \left(dr^{(i)} \times dr^{(j)}\right) \quad (\mathbf{3.1})$$

generalizzato per due curve aperte. Tuttavia, dovendo analizzare strutture discrete e non continue, questo indice va modificato. Servono quindi delle definizioni[15] per quantificare i parametri in gioco e per chiarire la natura dell'analisi. Due amminoacidi (a, b) sono quindi in contatto se almeno una coppia di atomi pesanti (viene escluso dunque H) appartenenti uno all'amminoacido a e l'altro all'amminoacido b hanno una distanza relativa $d \leq 4{,}5$ Å (le distanze sono misurate conoscendo le posizioni atomiche dai file PDB). Per poter estendere G al caso discreto si definisce la posizione dell'atomo C_α come r_i, la posizione media di legame $R_i \equiv \frac{1}{2}(r_i + r_{i+1})$ e il vettore di legame $\Delta R_i = r_{i+1} - r_i$. L'indice di *entanglement* sarà allora la somma discreta:

$$G_c'(i,j) = \frac{1}{4\pi} \sum_{i=i_1}^{i_2-1} \sum_{j=j_1}^{j_2-1} \frac{R_i - R_j}{\left|R_i - R_j\right|^3} \cdot (\Delta R_i \times \Delta R_j) \quad (\mathbf{3.2})$$

Dove $i_{1,2}$ sono gli indici del *loop*, con la condizione $i_2 - i_1 \geq 10$, e $j_{1,2}$ sono gli indici del *thread*, sempre con la condizione $j_2 - j_1 \geq 10$ e il *loop* e il *thread* non sono sovrapposti in sequenza. Per la frustrazione si è invece utilizzata la formula:

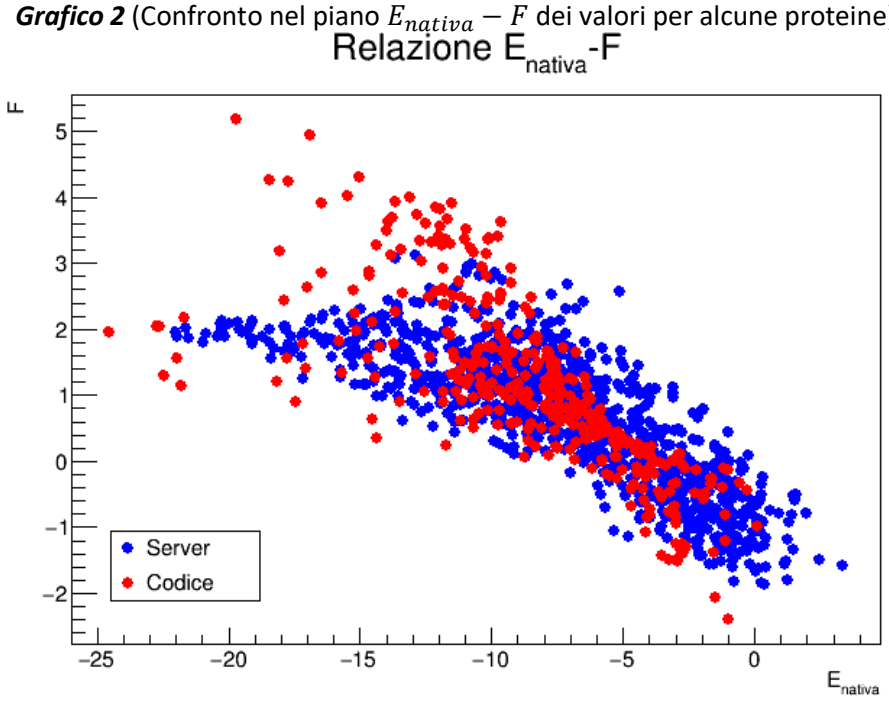
$$F_{i_1,i_2} = \frac{E^N_{i_1,i_2} - \langle E^M_{i_1',i_2'} \rangle}{\sqrt{\frac{1}{N}\sum_{l=1}^{N} \left(E^{l,M}_{i_1',i_2'} - \langle E^M_{i_1',i_2'} \rangle\right)^2}} \quad (\mathbf{3.3})$$

Dove i_1, i_2 sono gli amminoacidi agli estremi del *loop*, $E^N_{i_1,i_2}$ è l'energia del contatto nella struttura nativa, $\langle E^M_{i'_1,i'_2} \rangle$ è la media delle energie dei contatti mutati, $E^{l,M}_{i'_1,i'_2}$ è l'energia della singola mutazione, N è il numero di mutazioni totali. Per il calcolo delle energie, si è utilizzato il potenziale statistico[15]:

$$E_{norm}(a,b) = -log\frac{f_c(a,b)}{f(a,b)} \qquad (\mathbf{3.4})$$

Dove (a,b) è una data coppia di amminoacidi, $f_c(a,b) = N_c(a,b)/N_c$ è la frazione di contatti (a,b) su tutti gli N_c contatti, $f(a,b) = N(a,b)/N$ è la frazione di possibili coppie (a,b) sull'insieme totale delle generiche coppie N. A questo punto si è dovuta usare una convenzione per F. Si è allora preso come riferimento il server "Frustratometer"[17], dove viene implementata la $(\mathbf{3.3})$. Nello specifico si sono operate $N = 1000$ mutazioni per ogni coppia amminoacidica. Una mutazione è una generazione casuale di una coppia, che deve essere però coerente con la proteina analizzata. Dunque, per ciascuna proteina si campionano gli amminoacidi, dando maggior peso a quelli più presenti nella sequenza. In questo modo la mutazione generata sarà più facilmente compatibile con la struttura proteica (per esempio se una proteina contiene una minima frazione di cisteine, è difficile che questa venga estratta spesso nelle mutazioni). Per ogni mutazione poi, si calcola $E^{l,M}_{i'_1,i'_2}$, che è l'energia della struttura nativa con tale contatto cambiato. Prendendo poi la media sulle mutazioni effettuate e l'energia nella sequenza originaria è possibile determinare F. Il server è stato utilizzato per calcolare la frustrazione in alcune proteine in modo da poter verificare la validità del codice e del semplice potenziale statistico E_{norm} utilizzato in questa analisi. La funzione energia utilizzata in Frustratometer è infatti molto più sofisticata. Il confronto server/codice è analizzato nel ***Grafico 2***, dove sono presenti i valori di F vs E_{nativa} calcolati dal server (blu) e dal codice (rosso) per i contatti in 10 proteine scelte casualmente tra quelle analizzate.

Grafico 2 (Confronto nel piano $E_{nativa} - F$ dei valori per alcune proteine)



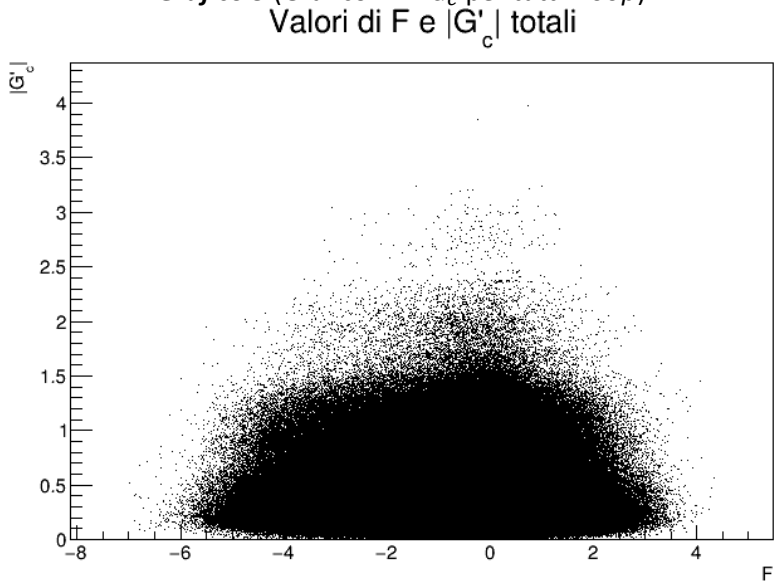
Si nota subito il diverso numero di punti; questo è dovuto al fatto che il server analizza la frustrazione anche per coppie amminoacidiche distanti anche meno di 10 residui tra loro. Inoltre, il server rinumera gli amminoacidi in modo automatico, e questo rende difficile stabilire a quale coppia appartenga quel dato indice. La buona sovrapposizione conferma un ottimo accordo tra le misure del codice e quelle del server; tale sovrapposizione è stata ottenuta tramite uno shift di 5 unità per l'energia (ciò non deve preoccupare perché i potenziali non sono uguali, e possono essere definiti a

meno di costanti additive) e uno scambio sul segno di F. La seconda operazione di scambio di segno potrebbe apparire erronea, tuttavia, analizzando la formula, se $F < 0 \implies E^N_{i_1,i_2} < \langle E^M_{i'_1,i'_2} \rangle$, cioè l'energia nativa è minore della media delle mutazioni possibili. F negative corrispondono quindi a strutture favorite e poco frustrate. Nel server la definizione è contraria, F positive corrispondono a contatti poco frustrati. Data l'ambiguità presente tra definizione e utilizzo del parametro nel server, si è scelto di cambiare il segno di F solo in questo grafico, per mostrare che il codice utilizzato trova valori di F ed E_{nativa} che si distribuiscono in modo simile. Nel resto dell'analisi il segno di F sarà quello calcolato dal codice, con la convenzione che F negativa corrisponde a frustrazione minore.

3.2 Struttura dell'analisi e risultati globali

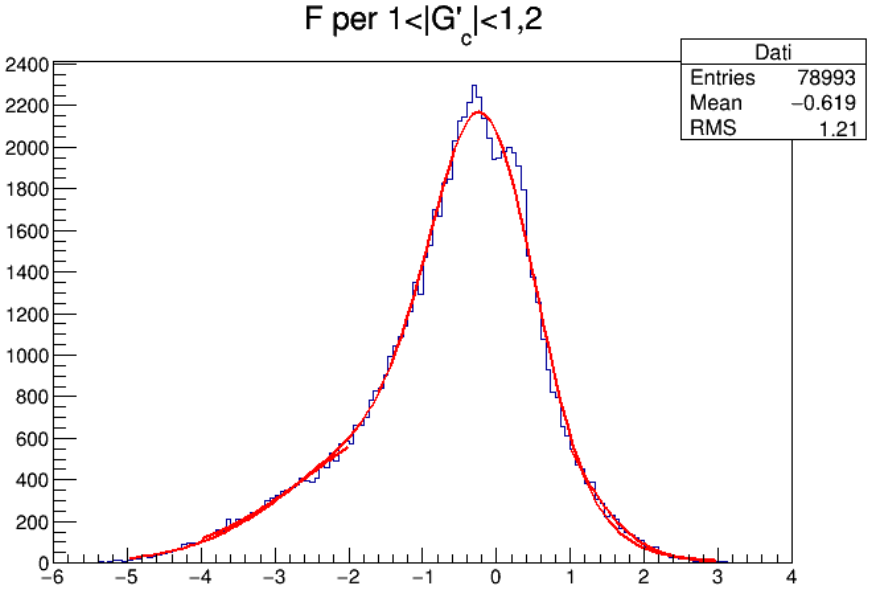
Per analizzare la frustrazione è stato scritto un codice che potesse calcolare F per coppie amminoacidiche agli estremi di *loop* nella catena. Il codice quindi deve innanzitutto ritrovare le stesse coppie di amminoacidi già analizzate[15] (caratterizzate dal numero i_1, i_2 che determina la posizione del *loop* nella catena, e dalla natura degli amminoacidi a, b coinvolti nel contatto). Sono state prese in considerazione circa 21 000 proteine e di queste circa il 70% aveva all'interno della sua struttura uno o più *loop*. Dei 3,6 milioni di contatti analizzati[15], il codice ne ritrova circa 3,3, dunque oltre il 90%. Dopo aver trovato i contatti il programma calcola la frustrazione di ciascuno.
Nel ***Grafico 3*** si vedono i risultati totali dell'analisi, cioè i valori di F e $|G'_c|$ per ogni contatto.

Grafico 3 (Grafico $F - G'_c$ per tutti i *loop*)



Ad un primo sguardo non emerge alcuna struttura comune. Visti i risultati[15], ci si aspetterebbe una relazione esplicita tra i due parametri. Infatti, all'aumentare dell'*entanglement* e dell'intricatezza della struttura, sarebbe atteso un aumento della frustrazione.
Si è deciso allora di analizzare le distribuzioni, degli indicatori. Dopo aver suddiviso $|G'_c|$ in intervalli di 0,2 unità (il parametro varia, con contributi significativi, tra 0 e 1,6), sono stati graficati gli istogrammi dei valori di F per contatti con $0{,}2 \cdot n < |G'_c| < 0{,}2 \cdot (n+1)$. Il pattern tipico della distribuzione è visibile al ***Grafico 4***. La distribuzione dei valori di F ricorda abbastanza la somma di gaussiane e per questo motivo sono stati eseguiti diversi fit. Tra i diversi fit provati (con 1, 2, 3 e 4 gaussiane) quello risultato migliore è stato quello con 2 gaussiane (un esempio al ***Grafico 4***). Tutti i fit presenti in questa tesi sono stati eseguiti con ROOT[18].

Grafico 4 (Esempio di distribuzione di F in un intervallo di $|G_c'|$ e relativo fit a due gaussiane)



Nonostante il fit non descriva bene l'andamento a due picchi, la funzione somma di due gaussiane è risultata la migliore sia perché non affetta da problemi di *overfitting* e dispersione dei parametri, ma anche perché descrive bene l'andamento diverso tra $F < 0$ e $F > 0$, cosa che non riesce con una sola gaussiana. Il fit ha determinato i 6 parametri della somma delle gaussiane, ma i due più importanti sono senza dubbio le medie μ_1 e μ_2. Le due medie infatti restituiscono la posizione dei due picchi principali di F; andando poi a monitorare questi valori al crescere di $|G_c'|$ sarà possibile stabilire un legame tra i due parametri. Per avere una stima migliore su $|G_c'|$ si è preferito usare non il centroide dell'intervallo, ma, dato l'andamento non uniforme di $|G_c'|$ (***Grafico 5***), è stato scelto il valore più probabile, definito come $\langle|G_c'|\rangle = \frac{\sum_i G_i \Delta p_i}{\sum_i \Delta p_i}$. Dividendo l'intervallo in piccoli segmenti centrati in G_i, ciascuno con una probabilità relativa Δp_i, si è calcolato empiricamente $\langle|G_c'|\rangle$. I risultati dei fit sono riportati in ***Tabella 1***, mentre il fit lineare fra μ_1, μ_2 e $\langle|G_c'|\rangle$ e i relativi risultati sono al ***Grafico 6*** e ***Tabella 2***.

Grafico 5 (Distribuzione dei valori di $|G_c'|$)

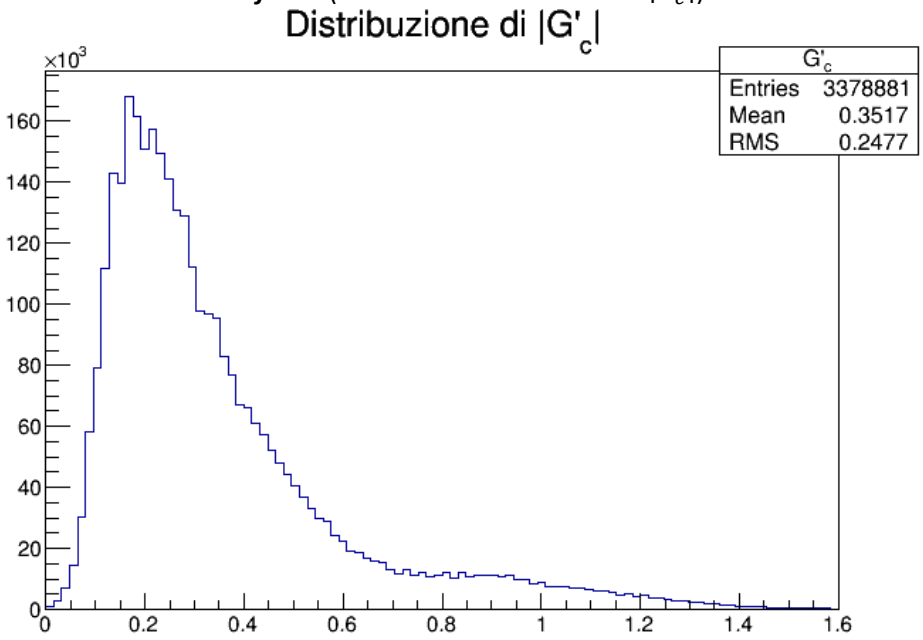


Tabella 1 (Risultati del fit a due gaussiane)

$\langle\lvert G'_c\rvert\rangle$	μ_1	σ_1	μ_2	σ_2	χ^2/dof
0,14	-1,235±0,006	1,452±0,002	-0,229±0,002	0,799±0,003	$3{,}7\cdot10^3$/154
0,28	-1,291±0,005	1,479±0,002	-0,250±0,002	0,793±0,002	$5{,}4\cdot10^3$/154
0,48	-1,237±0,007	1,519±0,003	-0,172±0,003	0,754±0,003	$2{,}4\cdot10^3$/154
0,69	-1,15±0,01	1,509±0,005	-0,113±0,004	0,697±0,005	770/154
0,90	-1,18±0,01	1,489±0,006	-0,171±0,005	0,696±0,006	570/113
1,09	-1,18±0,02	1,457±0,007	-0,163±0,006	0,693±0,007	408/113
1,28	-1,21±0,03	1,47±0,01	-0,117±0,009	0,70±0,01	197/73
1,48	-0,98±0,04	1,42±0,02	-0,09±0,02	0,61±0,02	95/58

Grafico 6 (Picchi gaussiani di F e relativi fit)

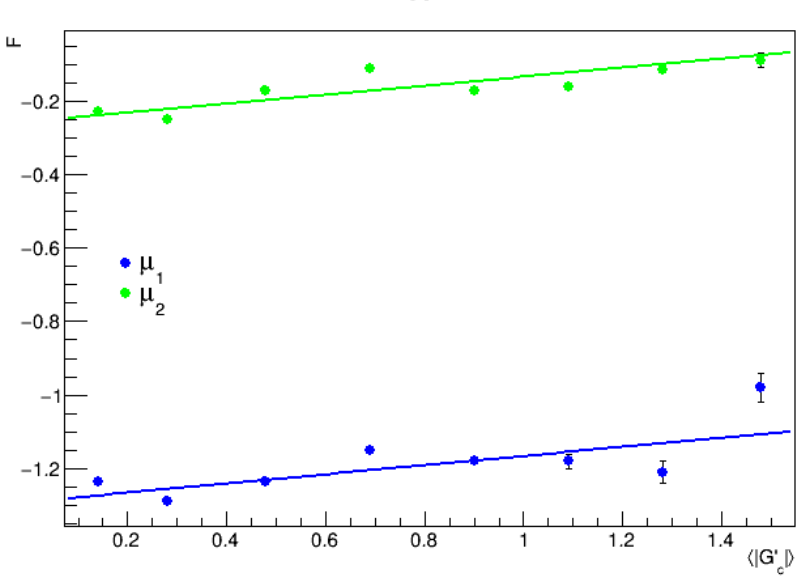


Tabella 2 (Risultati del fit)

Fit	μ_1	μ_2
a	-1,291±0,005	-0,257±0,002
b	0,12±0,01	0,122±0,004
χ^2	139/6	584/6
r	0,73	0,84

Dalla ***Tabella 1*** si nota che il valore del χ^2 risulta molto elevato rispetto ai gradi di libertà, soprattutto per piccoli valori di $\langle|G'_c|\rangle$. Questo si può spiegare osservando che i conteggi dei *loop* sono correlati fra loro e che una stima dei conteggi statisticamente indipendenti riduce di un fattore circa 8 l'istogramma totale[15]. Scalando quindi il χ^2 di tale fattore, i risultati dei fit tornano ad essere comunque buoni. Il ***Grafico 6*** e la ***Tabella 2*** evidenziano ottimamente la relazione attesa: la frustrazione aumenta all'aumentare dell'indice di *entanglement*. Inoltre, il fattore di crescita è lo stesso per entrambi i picchi analizzati, questo conferma sia che l'analisi con 2 picchi è una buona approssimazione, sia che l'*entanglement* agisce allo stesso modo su valori diversi della frustrazione. Infine, il fattore r di Pearson mostra che la relazione tra i due indici è ben approssimata da una relazione lineare. Il valore alto del χ^2 è principalmente da attribuirsi ad una sottostima dell'errore sulla posizione dei picchi.

Dopo aver verificato che globalmente l'andamento è quello atteso, si è fatta un'analisi anche sulle proteine singole, scegliendo due parametri per la frustrazione della proteina: F_{max} (trovata per il contatto con frustrazione massima) e $\langle F\rangle$ (media dei valori di F sui contatti di una proteina).

3.3 Analisi per proteine singole

Mentre l'*entanglement* di una proteina è ben descritto nel complesso da $|G'_{c-max}|$, per la frustrazione sono stati usati $\langle F\rangle$ e F_{max}. Per quanto riguarda il primo parametro, la distribuzione dei suoi valori al variare di $|G'_{c-max}|$ si comporta in modo piuttosto semplice. In prima approssimazione il pattern ricorda una gaussiana principale con altre due ai suoi lati (***Grafico 7***). Aumentando i valori di $|G'_{c-max}|$ però, la struttura a tre gaussiane non è più adatta e il grafico viene meglio descritto da

una gaussiana singola. Sono riportati al ***Grafico 8*** e alla ***Tabella 3*** i risultati del fit su gaussiana singola che, anche se impreciso per piccoli valori di $|G'_{c-max}|$, mostra comunque molto bene l'andamento atteso. Il coefficiente r di Pearson conferma la relazione lineare tra $|G'_{c-max}|$ e $\langle F \rangle$; dunque all'aumentare dell'*entanglement* nella proteina, anche la frustrazione media aumenta. Le proteine allora sviluppano strutture altamente *entangled* anche se localmente molto sfavorite energeticamente. Lo studio di questi complessi si rivela quindi interessante per la comprensione del *folding*.

Grafico 7 (Fit a gaussiana singola per $\langle F \rangle$ a valori estremi di $|G'_{c-max}|$)

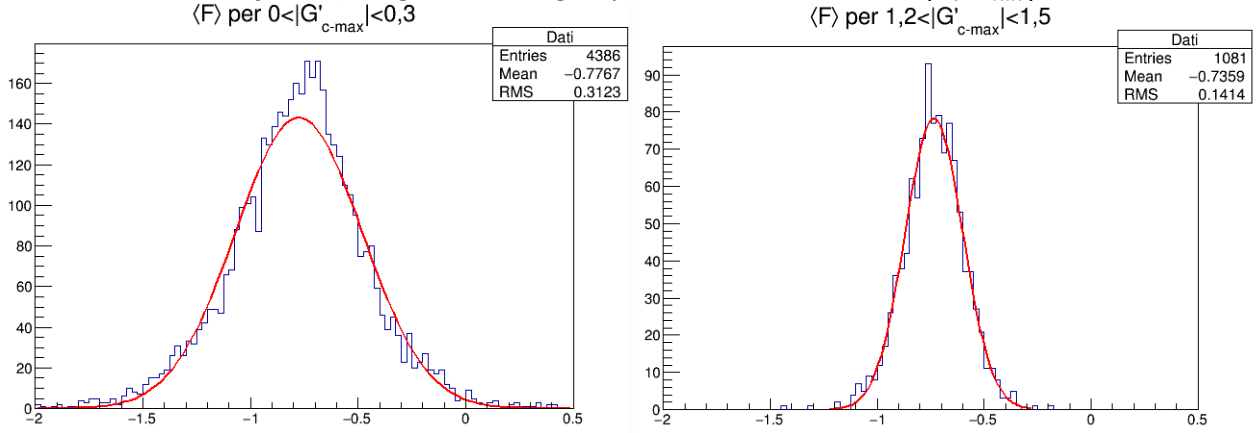


Tabella 3 (Risultati del fit a gaussiana singola per $\langle F \rangle$)

$\langle \lvert G'_{c-max} \rvert \rangle$	μ_g	σ_g	χ^2/dof
0,22	-0,776±0,005	0,294±0,004	144/94
0,42	-0,791±0,003	0,235±0,003	100/64
0,76	-0,751±0,004	0,184±0,003	71/54
1,03	-0,766±0,003	0,156±0,002	40/34
1,31	-0,735±0,004	0,134±0,004	18/27

Grafico 8 (Picco gaussiano di $\langle F \rangle$ e relativo fit)

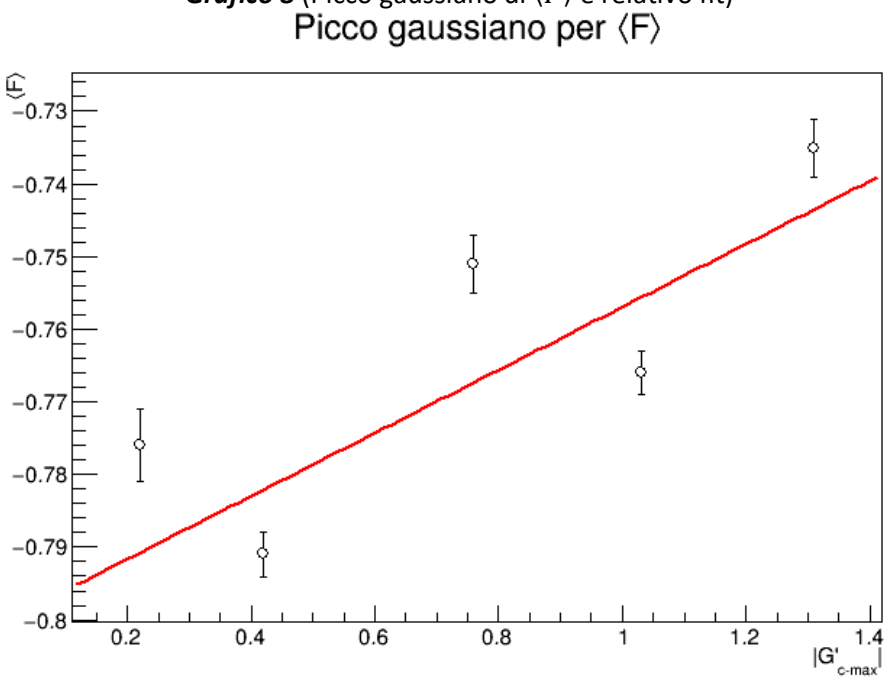


Tabella 4 (Risultati fit di μ_g per $\langle F \rangle$)

Fit	μ_g
a	-0,800±0,004
b	0,043±0,004
χ^2	50/3
r	0,81

Studiando il pattern di $\langle F \rangle$ si è riscontrato un altro risultato molto interessante: la dispersione del picco tende a diminuire all'aumentare dell'*entanglement*. La relazione tra σ_g del fit e $|G'_{c-max}|$ è ben descritta da una retta. Questo significa che nel processo di *folding* i vari parametri sono tenuti sotto controllo: ad un aumento di $|G'_{c-max}|$ corrisponde un aumento di F. Tuttavia, F non può aumentare a dismisura, altrimenti la struttura ripiegata non sarebbe più stabile. La progressiva diminuzione di σ_g può dunque essere interpretata come una progressiva focalizzazione attorno al valore ottimale di F che garantisce un giusto compromesso tra stabilità ed *entanglement*.

Grafico 9 (σ_g per $\langle F \rangle$ e relativo fit)

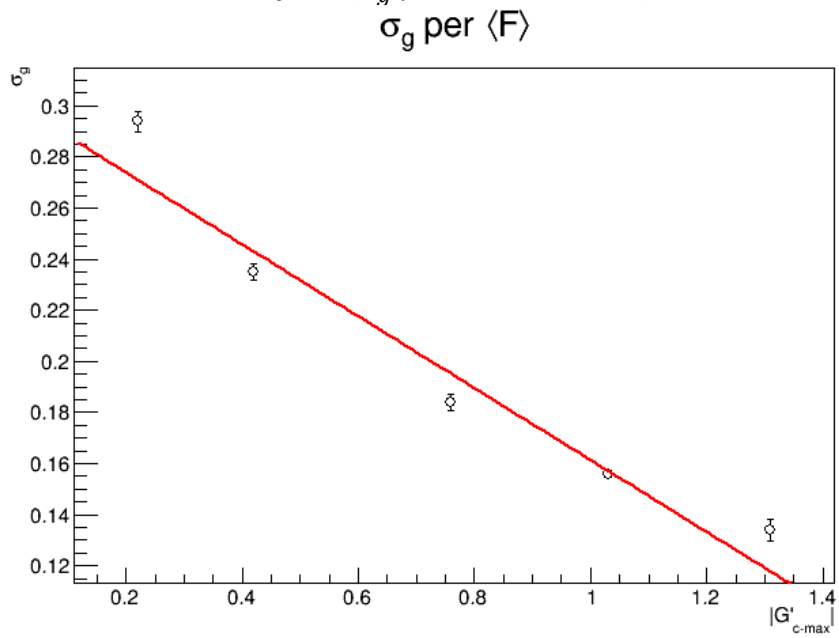


Tabella 5 (Risultati fit di σ_g per $\langle F \rangle$)

Fit	σ_{gaus}
a	0,302±0,003
b	-0,141±0,004
χ^2	70/3
r	0,97

Un altro valore analizzato per studiare l'andamento globale delle proteine è F_{max}, cioè la frustrazione massima per i contatti per ogni proteina. Il pattern tipico di questo parametro al variare di $|G'_{c-max}|$ è riportato al ***Grafico 10***. Nonostante la forma suggerisca la sovrapposizione di due gaussiane, l'andamento in realtà è più complesso (all'aumentare di $|G'_{c-max}|$ sembra che le gaussiane si scambino tra loro), e si è preferito fare un fit con gaussiana singola (che comunque approssima molto bene l'istogramma). I risultati alla ***Tabella 6*** mostrano chiaramente un andamento di F_{max} simile a quello di $\langle F \rangle$. Il fit è lineare con buona approssimazione, e conferma quanto già detto prima: all'aumentare dell'*entanglement* aumenta la frustrazione, non solo mediamente, ma anche in alcuni siti particolari. Anche in questo caso è stata fatta un'analisi sulla dispersione che mostra di nuovo un'ottima tendenza lineare.

Grafico 10 (Fit a gaussiana singola per F_{max} a valori estremi di $|G'_{c-max}|$)

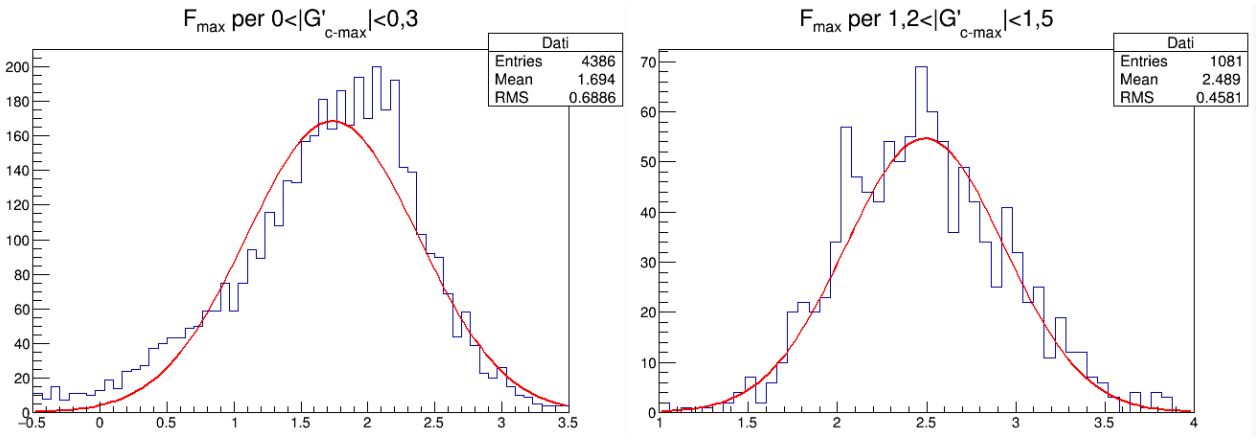


Tabella 6 (Risultati del fit a gaussiana singola per F_{max})

$\langle \lvert G'_{c-max} \rvert \rangle$	μ_g	σ_g	χ^2/dof
0,22±0,06	1,74±0,01	0,64±0,01	259/57
0,42±0,08	2,052±0,008	0,529±0,007	128/56
0,76±0,09	2,26±0,01	0,492±0,008	90/54
1,03±0,09	2,326±0,009	0,474±0,007	70/55
1,31±0,08	2,49±0,01	0,45±0,01	64/41

Grafico 11 (Picco gaussiano per F_{max} e relativo fit)

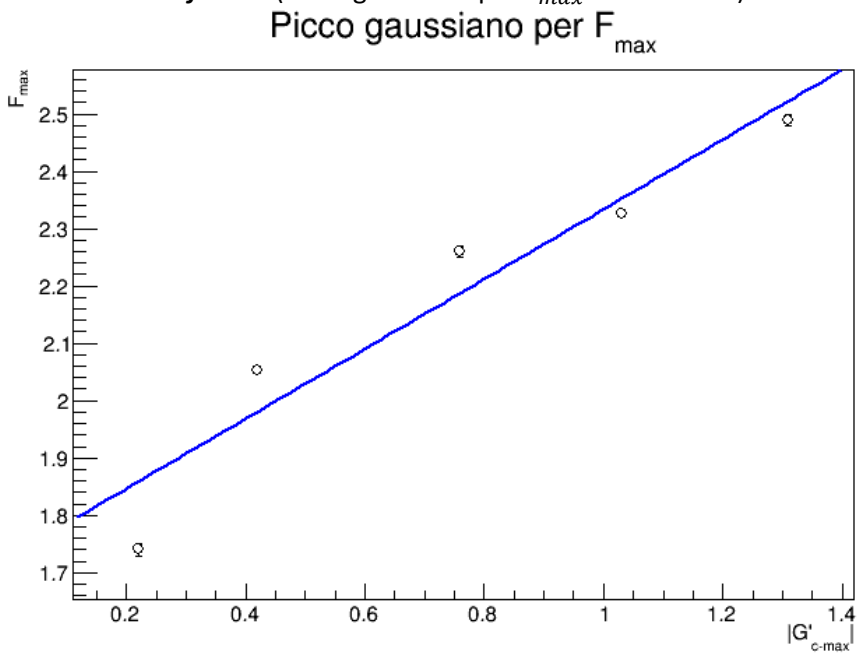


Tabella 7 (Risultati fit di μ per F_{max})

Fit	μ_g
a	1,725±0,009
b	0,61±0,01
χ^2	290/3
r	0,96

Grafico 13 (σ_g per F_{max} e relativo fit)

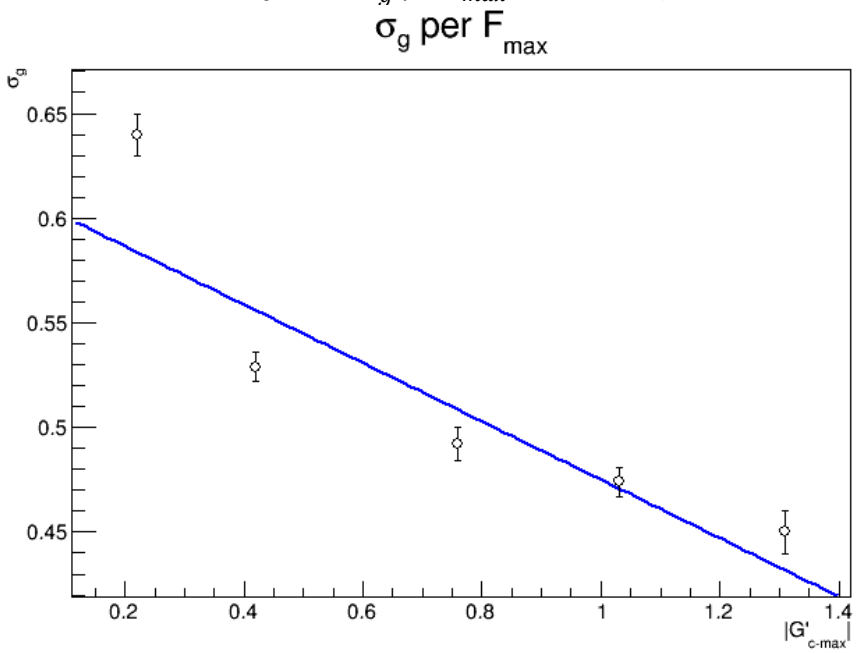


Tabella 8 (Risultati fit di σ_g per F_{max})

Fit	σ_g
a	0,614±0,008
b	-0,14±0,01
χ^2	54/3
r	-0,90

4. Conclusioni

4.1 Risultati e prospettive

Partendo dalla struttura base delle proteine, sono state descritte le principali dinamiche che regolano il loro ripiegamento. Il problema del *folding*, ben sintetizzato dal paradosso di Levinthal, è una questione ancora attuale e che stimola la ricerca di modelli in grado di predire la corretta struttura 3D di una proteina. Il concetto di frustrazione, sviluppato inizialmente per i vetri di spin, ha avuto un'importante applicazione a livello biologico, portando alla teorizzazione della cosiddetta *Energy Landscape Theory* ed alla formulazione del principio di minima frustrazione. Le proteine tendono cioè, durante il percorso di *folding*, ad abbassare la loro energia (ottimizzando le interazioni reciproche tra gli amminoacidi) e la loro entropia (ottimizzando la disposizione spaziale, dando vita a strutture simili a cristalli).

Negli ultimi anni, studi recenti[13,15] hanno mostrato l'importanza di strutture topologicamente complesse (come nodi e *loop entangled*) all'interno delle proteine. Sempre più dati sperimentali mostrano la crescente presenza di nodi, e questo complica ancora di più la comprensione della struttura. Sorgono spontanee domande sulla stabilità di queste strutture e del loro ruolo effettivo all'interno del *folding*.

La presente tesi, partendo da alcuni risultati sperimentali[15], cerca di mostrare esplicitamente il legame tra la frustrazione energetica e la complessità topologica di queste strutture. Nello specifico sono state prese in considerazione le strutture *loop-thread* descritte al ***Paragrafo 2.2-3*** e ne è stata calcolata la frustrazione. I risultati sono stati diversi e possono essere riassunti come segue:

1) la frustrazione dei singoli *loop* cresce con la complessità topologica;
2) la frustrazione media e massima all'interno di una proteina aumentano con la complessità topologica della proteina stessa (data da G'_{c-max});
3) la frustrazione media e massima tendono a localizzarsi vicino ad un valore limite all'aumentare della complessità topologica, diminuendone la dispersione.

I primi due risultati fanno vedere che la frustrazione aumenta per soli effetti di complessità topologica. Questo risultato è di per sé molto interessante; infatti, le proteine sono sistemi estremamente eterogenei e complessi, con un'incredibile varietà di interazioni atomiche e molecolari. Non è quindi scontato riuscire a trovare un andamento come quello mostrato nell'analisi. Mediamente allora, al crescere dell'*entanglement* all'interno delle proteine si ha una corrispettiva crescita della frustrazione.

Il terzo risultato è molto particolare; si tratta infatti di un'analisi fatta su parametri globali per una proteina. La frustrazione massima della proteina aumenta, come detto sopra, all'aumentare del valore di *entanglement* per quella particolare proteina. Tuttavia, l'analisi su $\langle F \rangle$ mostra anche che la proteina tiene sotto controllo questo aumento, tramite la presenza di contatti meno frustrati che abbassano la media globale. Questo mostra sia che la proteina necessita di un sito altamente *entangled* e frustrato (altrimenti tali strutture non sarebbero presenti), ma anche che la frustrazione globale è ottimizzata (attraverso la formazione di sottostrutture molto stabili) ad un valore circa costante (la pendenza di $\langle F \rangle$ al crescere di $|G'_{c-max}|$ è di 0,043 contro lo 0,12 della frustrazione globale o 0,60 di F_{max}).

Questo lavoro di tesi ha quindi confermato le ipotesi sulla relazione tra frustrazione ed *entanglement*, dando nuove conferme a riguardo (interessanti gli esiti su $\langle F \rangle$ e F_{max}).

Ulteriori studi potrebbero essere fatti andando ad analizzare le proteine singolarmente alla ricerca di un legame più esplicito, migliorando così i risultati ottenuti. Guardando ai casi particolari è possibile dare un peso e un significato ad alcune anomalie o risultati globali che risultano altrimenti incomprensibili. Infine, analisi dettagliate sulla topologia dei *loop* e dei *thread* potrebbero regalare risultati nuovi ed interessanti.

Riferimenti bibliografici

(1) Voet D., Voet J. G., Pratt C. W., *Fondamenti di biochimica*, Zanichelli, Bologna, **2013**, Cap. 5
(2) Voet D., Voet J. G., Pratt C. W., *Fondamenti di biochimica*, Zanichelli, Bologna, **2013**, Cap. 4
(3) Finkelstein A. V., Ptitsyn O., *Protein Physics*, Academic Press, Londra, **2002**, Cap. 3
(4) Finkelstein A. V., Ptitsyn O., *Protein Physics*, Academic Press, Londra, **2002**, Cap. 2
(5) Voet D., Voet J. G., Pratt C. W., *Fondamenti di biochimica*, Zanichelli, Bologna, **2013**, Cap. 6
(6) Finkelstein A. V., Ptitsyn O., *Protein Physics*, Academic Press, Londra, **2002**, Cap. 4
(7) Finkelstein A. V., Ptitsyn O., *Protein Physics*, Academic Press, Londra, **2002**, Cap. 1
(8) Whitford P. C., Sanbonmatsu K. Y., Onuchic J. N., Biomolecular dynamics: order-disorder transitions and energy landscapes, *Rep. Prog. Phys* **75** (2012)
(9) Finkelstein A. V., Ptitsyn O., *Protein Physics*, Academic Press, Londra, **2002**, Cap. 5
(10) Kessel A., Ben-Tal N., *Introduction to Proteins*, CRC Press, Boca Raton, **2018**, Cap. 5
(11) Dill K. A., MacCallum J. L., The Protein-Folding Problem, 50 Years On, *Science* **338**, 1042 (2012)
(12) Ferreiro D. U., Komives E. A., Wolynes P. G., Frustration in biomolecules, *Quarterly Reviews of Biophysics* **47**, 4 (2014)
(13) Lim N. C. H., Jackson S. E., Molecular knots in biology and chemistry, *J. Phys.: Condens. Matter* **27** (2015)
(14) Jackson S. E., Suma A., Micheletti C., How to fold intricately: using theory and experiments to unravel the properties of knotted proteins, *Current Opinion in Structural Biology* **42** (2017)
(15) Baiesi M., Orlandini E., Seno F., Trovato A., Sequence and structural patterns detected in entangled proteins reveal the importance of co-traslational folding, *Sci. Rep.* **9** (2019)
(16) https://www.youtube.com/watch?v=cAJQbSLlonI
(17) http://frustratometer.qb.fcen.uba.ar/localizing_frustration
(18) https://root.cern.ch/downloading-root

Ringraziamenti

Sono giunto alla fine di questo percorso, ma a breve ne inizierà un altro. Volevo ringraziare alcune persone che mi hanno aiutato a raggiungere questo obiettivo.
Grazie al Professor Antonio Trovato per i suggerimenti, le correzioni, e la disponibilità.
Grazie alla mia famiglia, Michela, Franco, Enrico, Elisa, che mi ha cresciuto negli anni e mi ha sostenuto durante il percorso.
Grazie ai compagni, Daniele, Marco, Piero e Luca, che hanno alleggerito le lezioni e i laboratori con racconti, risate e battute.
Infine, un grande Grazie ad Haidi, che con il suo Amore e la sua determinazione mi ha sempre spinto a credere in me stesso e a non mollare.