Università degli Studi di Padova

Dipartimento di Fisica e Astronomia
Corso di studi in Fisica

Classificazione di strutture proteiche per mezzo della loro dinamica intrinseca

Laureando:
Luigi Zanovello

Relatore:
Ch.mo Prof. Antonio Trovato

Anno accademico 2015/2016

Sommario

Modelli di rete elastica sono molto utilizzati per descrivere la dinamica intrinseca di una proteina attraverso l'analisi dei modi normali di oscillazione. Nonostante la loro semplicità permettono infatti di riprodurre e predire movimenti collettivi fondamentali per la funzionalità biologica della proteina. In questo lavoro di tesi si utilizzano le proprietà della dinamica intrinseca per confrontare fra loro e classificare diverse strutture proteiche, includendo fra queste anche strutture "artificiali" recentemente generate per mezzo di simulazioni numeriche.

Indice

1 Struttura delle proteine **7**
1.1 Le proteine 7
1.2 Stati nativi 8
1.3 Riproduzione "in silico" di strutture simili a catene proteiche: VAL60 9
1.4 Confronto fra strutture VAL60 e Protein Data Bank 11

2 Proprietà dinamiche delle proteine **14**
2.1 Dinamica interna e modelli a rete elastica 14
2.2 Gaussian Network Model 16
2.3 Dimensione spettrale 18

3 Proprietà dinamiche delle strutture di VAL60 **21**
3.1 Strutture utilizzate e lavoro preliminare 21
3.2 Localizzazione dei modi 21
3.3 Correlazione tra Contact Order e localizzazione media 23
3.4 Distribuzione delle frequenze dei modi di oscillazione e dimensione spettrale 25

4 Conclusione **29**

A Bibliografia **32**

Capitolo 1

Struttura delle proteine

1.1 Le proteine

Le proteine sono presenti in tutte le cellule e sono le macromolecole più abbondanti negli organismi, infatti la maggior parte delle reazioni biologiche che avvengono in una cellula sono operate dalle proteine[1]; esse hanno un grandissimo numero di funzioni, che spaziano dalla protezione immunitaria fino alla trasmissione di impulsi nervosi, dal controllo della crescita alla regolazione interna.

La struttura delle proteine può essere considerata a diversi livelli; essa è infatti caratterizzata da una struttura primaria, una struttura secondaria, una terziaria e a volte una quaternaria. [2]

Le proteine sono dei polimeri formati da 20 diversi tipi di amminoacidi (detti anche residui) disposti a formare una catena lineare: la diversità di una sequenza di amminoacidi risiede nelle diverse combinazioni di amminoacidi concatenati. Questi amminoacidi sono caratterizzati dalla presenza di un atomo di carbonio centrale (detto carbonio-α), di un gruppo carbossilico, uno amminico e una catena laterale, diversa per ciascun amminoacido (tale catena laterale può essere idrofobica, non polare o carica). Essi sono legati tra loro dalla presenza di legami covalenti (che per gli amminoacidi sono detti peptidici) tra le molecole. La sequenza di amminoacidi forma la struttura primaria.

La struttura secondaria è invece dovuta al ripiegamento della catena peptidica a causa delle interazioni tra ciascun amminoacido e gli amminoacidi vicini. Esistono diversi tipi di struttura secondaria, tra cui i principali sono:

- α-elica (α-helix): questa struttura è di tipo elicoidale ed è la struttura secondaria più semplice. Lo scheletro carbonioso della catena si avvolge attorno all'asse centrale con le catene laterali protese verso l'esterno. Essa è tenuta assieme dai legami idrogeno tra un gruppo carbossilico e il gruppo amminico del quarto amminoacido successivo lungo la catena.

- β-foglietto (β-sheet): è caratterizzato dalla presenza di legami idrogeno tra i gruppi carbossilici e quelli amminici di residui anche lontani tra loro lungo la sequenza.

La struttura terziaria è una combinazione di strutture secondarie tra loro; nel caso in cui la proteina sia costituita da una sola catena amminoacidica, la struttura della proteina stessa è data dalla struttura terziaria. Una struttura terziaria è ad esempio il motivo "coiled coil", caratterizzato da un fascio di α-eliche che si superavvolgono tra di loro. La struttura terziaria si ottiene quindi tramite un ripiegamento complessivo della catena proteica dovuto a interazioni tra gli amminoacidi di due diversi elementi di struttura secondaria (comunque appartenenti alla stessa catena).

Le proteine che ripiegano in solvente acquoso possono essere suddivise in due categorie a seconda della loro struttura terziaria:

- proteine fibrose, aventi catene peptidiche disposte in lunghi filamenti e in generale caratterizzate da un singolo tipo di struttura secondaria;

- proteine globulari le cui catene peptidiche sono ripiegate in forme sferiche o globulari, costituite generalmente da diversi tipi di strutture secondarie. Tali proteine sono caratterizzate da un nocciolo idrofobico circondato da strutture α e β, mentre i loop irregolari sono spinti verso l'esterno della struttura, in quanto non essendo coinvolti in strutture secondarie possono interagire tramite legami idrogeno con l'acqua stabilizzando la proteina ed evitando interazioni disastrose per la struttura stessa tra le molecole d'acqua e gli amminoacidi idrofobici della regione centrale. [3]

Infine, la struttura quaternaria è presente solo nelle proteine dotate di più subunità (catene peptidiche) che possono essere uguali o diverse. Queste subunità sono associate grazie a forze deboli (per esempio forze di Van Der Waals, ioniche, legami idrogeno e interazioni idrofobiche).

1.2 Stati nativi

Generalmente le strutture proteiche si trovano in conformazioni che sono date dall'insieme di strutture primaria, secondaria, terziaria e quaternaria (quando presente) caratteristico di quella proteina e sono dette stati nativi.

Essi sono le conformazioni in condizioni termodinamiche più stabili, ossia quelle che minimizzano l'energia libera di Gibbs G.

Una proteina verrà detta stabile se essa tende a mantenere il suo stato nativo. Gli effetti che favoriscono o ostacolano la stabilità di una proteina sono molteplici: quando essa non è ripiegata, le possibili conformazioni che può assumere sono numerose, e dunque tale stato sarà dotato di una maggiore entropia. Questo fatto, insieme ai legami idrogeno che gli amminoacidi formano con il solvente acquoso in cui sono immersi, tende a mantenere la proteina in uno stato non ripiegato. Gli effetti che si oppongono a ciò e che fanno sì che la proteina assuma la caratteristica struttura ripiegata sono invece le interazioni covalenti tra amminoacidi (legame disolfuro) e le interazioni non covalenti, più deboli delle precedenti. Esse sono però più numerose, e apportano dunque un contributo maggiore alla stabilità.

Gli stati nativi sono importanti in quanto una proteina nel suo stato nativo svolge correttamente la sua funzione biologica. Quando la proteina perde la sua conformazione nativa, ossia avviene un processo di denaturazione (causato da variazioni di temperatura, pH o altri fattori), non è più in grado di svolgere la sua funzione biologica.

Tramite un processo, detto di rinaturazione, è stato possibile dimostrare un fatto molto importante: la struttura terziaria di una proteina (e quindi anche la secondaria) è univocamente determinata dalla sua sequenza di amminoacidi. Questa conclusione deriva da alcuni esperimenti (il primo dei quali è dovuto a Christian Anfinsen nel 1950 su denaturazione e rinaturazione della ribonucleasi[2]) che mostrarono come una proteina, precedentemente denaturata, se posta nelle condizioni in cui il suo stato nativo è stabile riguadagni la sua funzionalità biologica e una struttura analoga a quella nativa. Dunque la sola sequenza di residui è sufficiente a determinare la struttura tridimensionale della proteina stessa.

È importante notare però come l'esperimento di Anfinsen non abbia messo in luce una corrispondenza 1:1 tra le strutture proteiche e le catene di amminoacidi, corrispondenza che di fatto non c'è. Si è infatti scoperto che la mappa

$$\textit{sequenza di amminoacidi} \longrightarrow \textit{struttura della proteina}$$

risulta essere degenere. Dunque, mentre a una sequenza di amminoacidi è associata necessariamente una struttura 3-dimensionale ben definita, può accadere che a diverse sequenze di amminoacidi corrisponda la "stessa" struttura (non è banale definire una buona misura di distanza fra due strutture proteiche tridimensionali ed è arbitrario, entro certi limiti, decidere che due strutture sono sufficientemente simili da essere considerate come un'unica struttura). Notiamo in particolare che, nel confronto tra sequenze di amminoacidi, si distinguono due principali categorie: sequenze omologhe (con più del 20-30% di residui uguali) che si dicono appartenenti alla stessa famiglia e discendenti dalla stessa proteina progenitore, e sequenze non omologhe (che possono comunque avere la stessa struttura e in tal caso potrebbero costituire o un esempio di omologia remota e non rilevabile, oppure uno di evoluzione convergente verso la stessa struttura e la stessa funzione, a partire da diverse proteine progenitrici). Come vedremo nel secondo capitolo, per studiare relazioni di tipo evoluzionistico tra proteine diverse ci si è basati anche su altri studi con un approccio non soltanto strutturale.

1.3 Riproduzione "in silico" di strutture simili a catene proteiche: VAL60

Ci concentreremo ora su un lavoro [4] riguardante la riproduzione di strutture simili a proteine tramite simulazione, dai cui risultati parte il presente lavoro di tesi.

In tale esperimento è stato possibile generare $\sim$ 30000 omoproteine di polivalina (VAL60), costituite da 60 amminoacidi uguali tra loro e caratterizzate energeticamente da un potenziale che tiene conto in maniera accurata di tutte le interazioni tra atomi: il potenziale utilizzato è un potenziale classico che usa come particelle interagenti gli atomi pesanti della catena peptidica (escludendo quindi gli atomi di idrogeno), mentre le interazioni con il solvente sono trattate implicitamente (senza considerare nel potenziale in maniera esplicita le molecole del solvente); infine i parametri di tale potenziale sono ricavati empiricamente, cercando di riprodurre risultati noti per piccoli peptidi. Come si è potuto verificare, tali strutture corrispondono in maniera molto vicina a minimi locali dell'energia potenziale di VAL60 e sono dotate di quantità strutturali molto simili a quelle delle strutture proteiche esistenti in natura. Da ciò si è quindi dedotto che tali strutture sono molto simili a strutture proteiche di lunghezza comparabile (non si sono costruiti set con più amminoacidi in quanto 60 era il limite di amminoacidi computazionalmente studiabili *in silico*), e che è dunque possibile riprodurre strutture proteiche esistenti in natura da una catena formata da un solo tipo di amminoacido (e quindi senza bisogno dell'informazione dovuta alla sequenza). Non solo, come viene descritto nella sezione seguente, si è dimostrato che le strutture ottenute tramite simulazione sono in grado di riprodurre tutte le strutture esistenti in natura.

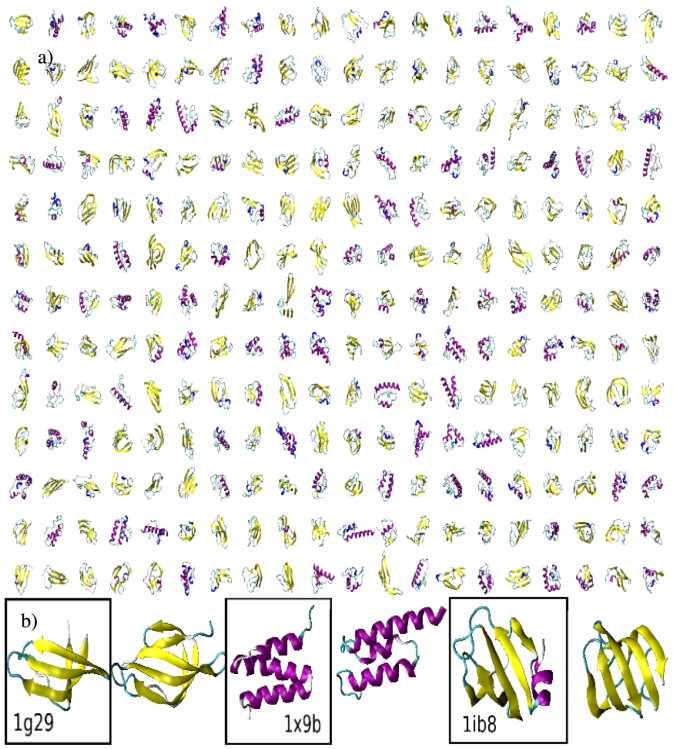

Figura 1.1: *(a): una selezione di 260 delle 30000 strutture simulate. (b): alcuni esempi di strutture simulate corrispondenti ad alcune esistenti in natura presentate insieme. Le strutture secondarie sono mostrate in maniera pittorica.*

1.4 Confronto fra strutture VAL60 e Protein Data Bank

Nel corso dell'analisi delle strutture prodotte nelle simulazioni descritte nella sezione precedente, si è osservato un fatto molto interessante. La comparazione tramite una tecnica detta *TM-align* tra strutture proteiche esistenti in natura (catalogate nel Protein Data Bank e di cui viene usata la classificazione nota come CATH), di lunghezza paragonabile alle strutture simulate, e le strutture di VAL60 ha portato a notare che mentre tutte le proteine esistenti in natura di lunghezza ~ 60 residui sono presenti (o ne esistono di molto simili) nel set di strutture prodotte artificialmente, ne esistono molte altre nel set di VAL60 che non riscontrano una corrispondenza in natura. Questo dunque porta a chiedersi se le sequenze proteiche presenti in natura esistano perché sono state scelte in maniera casuale tra tutte quelle possibili o se invece hanno qualcosa di speciale che faccia sì che alle proteine "convenga", dal punto di vista energetico o per altri motivi, assumere una conformazione piuttosto che un'altra.

In particolare si è notato che, mentre le strutture di VAL60 sono prevalentemente formate da β-foglietti paralleli, tale conformazione è molto più rara nell'insieme di proteine esistenti in natura. Tale osservazione può essere quantificata introducendo una coppia di grandezze, il contact order (CO), che verrà trattato anche più avanti nei prossimi capitoli e la contact locality (CL).

Definiremo come contact order di una struttura proteica la media della distanza all'interno della sequenza tra residui in contatto, divisa per la lunghezza totale della catena.

Dunque, una volta fissata una distanza per cui due residui possono considerarsi in contatto ($R_{C_{CO}}$ da qui in avanti) si potrà definire il contact order come segue:

$$CO = \frac{1}{N_r} \cdot \frac{\sum_n \mid i - j \mid}{\sum_n 1} \tag{1.1}$$

dove N_r è il numero totale di residui nella catena, n è il numero di coppie di amminoacidi che sono considerati in contatto tra di loro a seconda della misura di $R_{C_{CO}}$ inserita, i è l'indice di posizione dell'i-esimo amminoacido della catena e j l'indice del j-esimo.

La contact locality invece è una caratteristica strutturale che calcola la frazione di residui in contatto all'interno della stessa metà della catena. Utilizzando n con lo stesso significato visto sopra, si avrà $n = n_N + n_C + n_{CN}$, con n_N che indica il numero di coppie di residui in contatto in cui entrambi gli amminoacidi appartengono alla metà della catena più vicina all'estremità N-terminale (ossia terminante con un gruppo amminico), n_C il numero di coppie di residui in contatto in cui entrambi appartengano alla metà più vicina all'estremità C-terminale (terminante in un gruppo carbossilico) e n_{CN} il numero di coppie in contatto in cui i due residui appartengono a metà diverse.

A questo punto la contact locality sarà definibile come:

$$CL = \frac{(n_N + n_C)}{n} \tag{1.2}$$

A questo punto la differenza strutturale tra le proteine di VAL60 e quelle catalogate nel Protein Data Bank è esprimibile tramite queste due quantità: le strutture del PDB sono caratterizzate da un CO più basso rispetto a quelle di VAL60 e da un CL più alto, come si può verificare in *fig. 1.2*.

Questo risultato suggerisce che, tra tutte le possibili conformazioni che una catena proteica può assumere, le strutture reali sembrano essere orientate verso il mantenimento di un CO più basso. Da tali elementi si è quindi sviluppato il lavoro svolto nell'analisi di strutture proteiche di questa tesi.

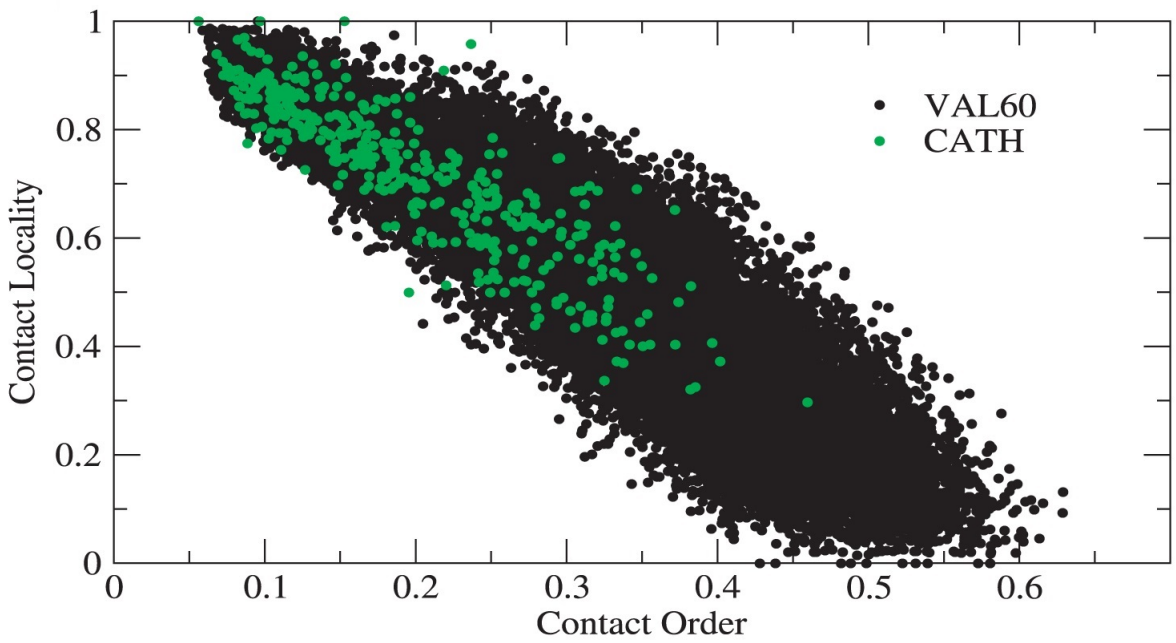

Figura 1.2: *Rappresentazione delle strutture del PDB con lunghezza compresa tra* 40 *e* 75 *residui e delle VAL*60 *in funzione dei rispettivi CO e CL.*

Capitolo 2

Proprietà dinamiche delle proteine

2.1 Dinamica interna e modelli a rete elastica

Negli ultimi decenni diversi studi [5] hanno portato alla consapevolezza che non solo la struttura statica ma anche le caratteristiche dinamiche delle proteine influenzano la loro funzione. In particolare, si è scoperto come confronti tra proteine mediante le loro proprietà dinamiche possano essere utili a scoprire proprietà funzionali e tratti simili tra strutture diverse che risultano elusivi a un approccio di tipo statico.

Nel precedente capitolo abbiamo descritto la degenerazione sequenza - struttura. Essa è risultata evidente imponendo che tra tutti i cambiamenti di struttura possibili avvenissero soltanto quelli evoluzionisticamente accessibili, ovvero oggetto di un ripiegamento stabile e riproducibile, ottenendo così un numero di strutture possibili molto minore delle sequenze di partenza. Da questo è emerso che per studiare relazioni evoluzionistiche tra strutture proteiche diverse fosse più conveniente utilizzare metodi basati sullo studio delle strutture proteiche piuttosto che sullo studio delle loro sequenze. Recentemente si è trovato anche che alcune caratteristiche specifiche della dinamica interna delle proteine, che influenzano la loro attività biologica, sono soggette a conservazione nel processo di evoluzione. In analogia con il caso sequenza-struttura, si può dunque avanzare l'ipotesi che metodi quantitativi atti a confrontare caratteristiche funzionali tra diverse strutture possano essere utilizzati per scoprire relazioni di tipo evoluzionistico che potrebbero invece sfuggire ad approcci basati sullo studio delle strutture statiche e delle sequenze di amminoacidi.

Visto che spesso le funzioni proteiche sono influenzate dalle loro proprietà dinamiche, negli anni passati sono stati sviluppati molti modelli teorici necessari per riuscire a trattare in maniera approssimata ma soddisfacente le proprietà della dinamica interna delle proteine.

Il modello utilizzato in questo lavoro di tesi rientra nel gruppo dei modelli a rete elastica, che deve il suo sviluppo agli studi innovativi compiuti da Tirion [5]. Sebbene tali modelli siano caratterizzati da una descrizione semplificata della struttura proteica e delle interazioni tra i suoi amminoacidi [6], sono in

grado di trattare in maniera accurata le caratteristiche dinamiche di proteine globulari con un consumo contenuto di risorse computazionali.

In tali modelli si procede a costruire una matrice di correlazione degli spostamenti (disposizioni) di coppie di amminoacidi dalla loro posizione di equilibrio, in grado di verificare se coppie di fluttuazioni di amminoacidi sono correlate tra loro e dunque se esse fanno parte di un moto collettivo (per approfondimenti vedere [7]), e il cui spettro sarà in grado di fornire informazioni sui principali gradi di libertà che influenzano la dinamica della struttura.

Dunque, la dinamica interna di una struttura proteica in questo tipo di modelli sarà descrivibile tramite le variazioni delle posizioni degli amminoacidi da quelle di equilibrio (della struttura nativa). A tal proposito introduciamo la matrice di correlazione delle disposizioni di coppia, C, una cui generica entrata è definita come

$$C_{ij,\mu\nu} = \langle \delta r_{i,\mu}(t) \cdot \delta r_{j,\nu}(t) \rangle \tag{2.1}$$

ove $\langle\rangle$ indica la media sul tempo di simulazione e $\delta r_{i,\mu}(t)$ indica la componente cartesiana μ-esima del vettore tridimensionale $\delta \mathbf{r}_i(t) \equiv \mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle$ (detto vettore di spostamento dell'amminoacido dalla sua posizione media nel tempo di simulazione). Per proteine aventi N amminoacidi, la matrice C ha dimensione 3N.

Notiamo [6] che l'entrata $C_{ij,\mu\nu}$ della matrice delle disposizioni di coppia può essere scritta anche nella maniera seguente:

$$C_{ij,\mu\nu} = \sum_{l=1}^{3N} \lambda_l v^l_{i,\mu} v^l_{j,\nu} \tag{2.2}$$

con λ_l (che ha dimensioni fisiche di una lunghezza al quadrato) che indica gli autovalori della matrice C ordinati in maniera decrescente e $\mathbf{v}^l$ i corrispondenti autovettori ortonormali.

Ora, visto che la fluttuazione quadratica media dell'intera struttura proteica è data da

$$\sum_{i,\mu} \langle \delta r_{i,\mu}(t)^2 \rangle = \sum_{i,\mu} C_{ii,\mu\mu} = \sum_l \lambda_l \tag{2.3}$$

si ha che gli autovettori con autovalore più grande rappresentano i gradi di libertà principali che influiscono sulla dinamica della struttura. Infatti, per proteine di dimensioni dell'ordine di grandezza di $\sim$100 amminoacidi, possono essere sufficienti i primi 10 autovettori per studiare in maniera abbastanza esatta le fluttuazioni della struttura. Tali autovettori sono generalmente indice di fluttuazioni non di singole coppie di residui, bensì di intere porzioni della proteina.

Generalmente, per modelli a rete elastica si usa una descrizione di ciascun amminoacido tramite uno o pochi centroidi (ad esempio quelli dei principali gruppi che costituiscono gli amminoacidi), e la costruzione dell'energia potenziale della struttura in modo che tenga conto in maniera quadratica dell'allon-

tanamento dai valori delle distanze tra coppie di residui in contatto nello stato nativo (in maniera simile a un potenziale elastico).

Nei modelli a rete elastica quindi l'energia potenziale per una struttura con N amminoacidi avrà la forma seguente:

$$U = \frac{1}{2} \cdot \sum_{ij,\mu\nu} \delta r_{i,\mu} M_{ij,\mu\nu} \delta r_{j,\nu} \tag{2.4}$$

con $\delta r_{i,\mu}$ che indica la differenza tra la coordinata μ del residuo i nella struttura per cui si valuta l'energia potenziale e la coordinata μ del residuo i nello stato nativo, e M una matrice simmetrica di dimensione $3N \cdot 3N$ i cui coefficienti cambiano a seconda delle coppie di amminoacidi considerate. Indichiamo ora con $\tau_1 ... \tau_{3N}$ gli autovalori di M ordinati in maniera crescente e con $\mathbf{w}_1 ... \mathbf{w}_{3N}$ i corrispondenti autovettori ortonormali: si avrà che gli autovettori $\{\tau_l\}$ sono tutti positivi a parte i primi sei che sono nulli e sono attribuiti alle traslazioni della molecola lungo i tre assi e alle rotazioni attorno agli stessi.

In particolare, notiamo che un'entrata generica della matrice di covarianza C può essere scritta per i modelli elastici come segue:

$$C_{ij,\mu\nu} = \kappa_B T \widetilde{M}^{-1}_{ij,\mu\nu} \tag{2.5}$$

con $\kappa_B T$ l'energia termica alla temperatura d'interesse e $\widetilde{M}^{-1}_{ij,\mu\nu}$ la matrice ottenuta da M rimuovendo lo spazio dei primi sei autovalori nulli e invertendola. In maniera analoga, un'entrata della matrice C potrà essere scritta nel seguente modo:

$$C_{ij,\mu\nu} = \widetilde{\sum_l} \frac{\kappa_B T}{\tau_l} w^l_{i,\mu} w^l_{j,\nu} \tag{2.6}$$

con il carattere $\sim$ sulla somma che indica che gli elementi relativi ai primi sei autovalori nulli sono stati omessi dal calcolo.

Prestiamo particolare attenzione al fatto che in precedenza abbiamo visto che i principali modi normali di oscillazione della struttura sono quelli relativi agli autovalori di C più alti, e che questi si traducono negli autovalori di M più bassi. Infatti per i τ_l più piccoli (ossia quelli che richiedono meno energia) si avranno i contributi maggiori alla matrice di correlazione. Come si può intuire dall'equazione (2.4), M è la matrice che quantifica il tipo di costante elastica per l'interazione tra diverse coppie di residui; dunque i suoi autovalori τ_l hanno dimensione di costanti elastiche appunto, risultando proporzionali al quadrato di una frequenza. Da questo segue che i modi di oscillazione caratterizzati da un valore di τ_l più piccolo saranno i più lenti.

2.2 Gaussian Network Model

Vediamo ora in maniera più approfondita il modello utilizzato in questo lavoro. Il Gaussian Network Model (da qui in poi GNM) è un modello a rete elastica in cui [8] si considerano costanti elastiche identiche per tutte le coppie di residui

interagenti e fluttuazioni isotrope, cosa che riduce gli autovettori dei modi di oscillazione da $3N-6$ a $N-1$ (ossia si avrà un solo autovalore nullo).

In questo modello ciascun amminoacido viene rappresentato come un singolo punto posizionato alle coordinate del suo carbonio-α, C^{α} [1]. Se una coppia di tali punti è posta al di sotto di una distanza di *cutoff* $R_{C_{EN}}$, i due residui sono considerati in contatto e sono collegati da una molla di costante elastica γ.

In tale modello il potenziale di interazione di due residui assumerà la forma:

$$V_{GNM}(i,j) = \frac{\gamma}{2} \cdot \{(\Delta x_{ij})^2 + (\Delta y_{ij})^2 + (\Delta z_{ij})^2\} \tag{2.7}$$

La caratteristica essenziale di questo modello è una matrice di Kirchhoff Γ (analoga alla matrice M del caso generico), che nel caso di una catena formata da N residui sarà una matrice $N \times N$ simmetrica e con determinante nullo, definita come segue.

$$\Gamma_{ij} = \begin{cases} -1 & se \quad i \neq j \quad e \quad R_{ij} \leq R_{C_{EN}} \\ 0 & se \quad i \neq j \quad e \quad R_{ij} > R_{C_{EN}} \\ -\sum_{i=1, i\neq j}^{N} \Gamma_{ij} & se \quad i = j \end{cases} \tag{2.8}$$

con R_{ij} distanza tra i residui i e j. Inoltre, dall'isotropia dello spazio imposta in tale modello, la forma del potenziale può essere ulteriormente semplificata ottenendo:

$$V_{GNM} = \frac{\gamma}{2} \cdot [\Delta \mathbf{R}^T \Gamma \Delta \mathbf{R}] \tag{2.9}$$

In questo caso [8], gli elementi della matrice di correlazione delle fluttuazioni saranno nella forma:

$$C_{ij} = \frac{3\kappa_B T}{\gamma} \cdot [\widetilde{\Gamma}^{-1}]_{ij} \tag{2.10}$$

con $\widetilde{\Gamma}^{-1}$ calcolata dagli N-1 autovettori non nulli τ_k di Γ e dai corrispondenti autovettori come segue:

$$[\widetilde{\Gamma}^{-1}]_{ij} = \widetilde{\sum_k} \frac{1}{\tau_k} \cdot [w_i^k w_j^k] \tag{2.11}$$

dove il simbolo $\sim$ sulla somma indica che nel calcolo è omesso il termine relativo all'autovalore nullo.

Come nel caso dei modelli generici a rete elastica dunque, anche qui si avrà che ad autovalori di Γ più piccoli (associati a frequenze più basse dei modi di oscillazione della struttura) corrisponderanno i contributi maggiori alla dinamica interna della struttura e i corrispondenti autovettori descriveranno la forma delle oscillazioni. Facciamo notare che nel caso del GNM la matrice di Kirchhoff e quindi i suoi autovalori risultano essere adimensionali, come si può notare confrontando l'equazione (2.5) con la (2.10) e notando come in quest'ultima compaia la costante elastica γ a denominatore.

Definiamo ora un ultimo parametro che verrà usato nel corso di questo lavoro, la localizzazione ρ. Questo parametro in sostanza è utile per descrivere se un modo di oscillazione è localizzato (ossia interessa soltanto una piccola parte della struttura) o è delocalizzato (e quindi tale oscillazione influenza gran parte della struttura proteica)[9].

$$\rho = \sum_{i=1}^{N} w_i^4 \tag{2.12}$$

I casi limite per tale parametro saranno $\rho = 1$ che indica un modo di oscillazione completamente localizzato su un singolo residuo i_0, caratterizzato da $w_i = \delta_{i,i_0}$, e $\rho = \frac{1}{N}$ che indica invece un modo completamente delocalizzato, caratterizzato da $w_i = \frac{1}{\sqrt{N}}$ per ogni residuo.

2.3 Dimensione spettrale

Dedichiamo ora una breve sezione per mostrare i risultati di un lavoro [10] che introduce un'altra delle grandezze analizzate nel corso di questa tesi, la dimensione spettrale d.

Essa è sostanzialmente un parametro in grado di descrivere il comportamento a basse frequenze della densità dei modi normali di oscillazione dati dal GNM, che risulta essere fortemente correlato al numero di amminoacidi presenti nella catena proteica.

In generale si ha che l'equilibrio di una struttura con un bagno termico può influenzarne in maniera molto pesante le possibili disposizioni topologiche; nel caso specifico delle proteine è stata dimostrata [10] l'esistenza di una dimensione critica per la stabilità e la sua dipendenza da un parametro topologico (la dimensione spettrale).

Tale parametro viene definito tramite il comportamento asintotico della densità dei modi di oscillazione della struttura a basse frequenze: indicando con $f(\omega)$ la funzione di densità di probabilità (PDF) dei modi con frequenza ω (che si trova come radice quadrata degli autovalori della matrice di Kirchhoff Γ), si ha che tale funzione ha il seguente andamento:

$$f(\omega) \sim \omega^{d-1} \tag{2.13}$$

nel limite $\omega \longrightarrow 0$ (basse frequenze). La dimensione spettrale coincide con la dimensione euclidea per strutture ordinate con interazioni a corto raggio fra primi vicini, mentre per quanto riguarda lo studio di strutture disordinate (come le proteine) generalmente assume valori compresi tra 1 e 3 per interazioni a corto raggio. Essa costituisce un ottimo modo di misurare il livello di connessione di strutture geometriche complesse su larga scala, infatti un alto valore di d è associato a alti livelli di connessione topologica.

Nel caso di un sistema posto in condizioni di instabilità termica, dal lavoro sopracitato è emerso che il valor medio del quadrato degli spostamenti degli atomi (o amminoacidi per le proteine) dalle loro posizioni di equilibrio ($\langle r^2 \rangle$) diverge per $N \longrightarrow \infty$, ossia per catene costituite da un grande numero di

amminoacidi, ma lo fa in modo diverso a seconda della dimensione spettrale d considerata. Ad esempio per valori di $d \leq 2$ si ha

$$\langle r^2 \rangle \propto \frac{\kappa_B T}{\gamma} N^{\frac{2}{d}-1} \tag{2.14}$$

mentre per $d = 2$ si ha

$$\langle r^2 \rangle \propto \frac{\kappa_B T}{\gamma} ln(N) \tag{2.15}$$

Fissato dunque un valore massimo per le fluttuazioni $\langle r^2 \rangle$, oltrepassato il quale la struttura diventa instabile, per strutture poste a una temperatura T esisterà un corrispondente valore di soglia $N(T)$ al di sopra del quale (ossia per catene più lunghe) la struttura risulterà instabile se sottoposta a oscillazioni termiche.

Visto che una proteina per mantenere la sua funzione biologica deve essere caratterizzata da una struttura e da modi di oscillazione ben specifici, che si perderebbero in caso di instabilità strutturale, si ha che è presente una correlazione tra la dimensione spettrale e la lunghezza delle proteine.

Ora che si è spiegata l'importanza della dimensione spettrale d per quanto riguarda la lunghezza e la stabilità della proteina, possiamo descrivere brevemente il modo in cui d è stata calcolata. Dopo aver ottenuto il set di autovalori $\{\omega_1^2, ..., \omega_N^2\}$ per la matrice Γ di ciascuna struttura analizzata, si è costruita una PDF $f(\omega)$ istogrammando i valori delle frequenze dei modi (come descritto nel capitolo successivo) e dalla conoscenza della PDF si è provveduto a ottenerne la cumulativa $G(\omega)$ dalla quale si è stimata la dimensione spettrale d tramite la relazione seguente:

$$G(\omega) \sim \omega^d \tag{2.16}$$

È infine importante considerare le conclusioni del lavoro [10], che hanno dimostrato come per diverse strutture proteiche il comportamento a basse frequenze della cumulativa $G(\omega)$ segua effettivamente la legge espressa nell'equazione (2.16) e che sussiste una dipendenza logaritmica non scontata tra la lunghezza della catena e la sua dimensione spettrale (di cui abbiamo parlato a inizio sezione) tramite la seguente relazione:

$$\frac{2}{d} = 1 + \frac{b}{ln(N)} \tag{2.17}$$

in cui b è una costante di proporzionalità che dipende dallo spazio a disposizione in media tra gli amminoacidi, dalla costante elastica γ e dalla temperatura T.

Inoltre, all'aumentare del valore di $R_{C_{EN}}$ considerato in tale lavoro, si è riscontrato anche un aumento del valore della dimensione spettrale d, che combacia con quanto trovato in questa tesi, come vedremo nel prossimo capitolo. Questo è la conseguenza dell'aumento del numero di connessioni topologiche fra coppie di residui della proteina all'aumentare di $R_{C_{EN}}$.

Capitolo 3

Proprietà dinamiche delle strutture di VAL60

3.1 Strutture utilizzate e lavoro preliminare

Nei capitoli precedenti abbiamo visto da quali risultati parte questo lavoro e quali sono i principali parametri valutati. Vediamo ora cosa quali sono stati gli obiettivi e i risultati a cui si è giunti.

Innanzi tutto i dati presentati di seguito sono stati ottenuti considerando 300 strutture di $VAL60$, scelte a caso tra quelle prodotte nel corso della ricerca trattata in [4], e alcune strutture proteiche scaricate dal Protein Data Bank.

Si è proceduto scrivendo un algoritmo in grado di leggere le posizioni dei C^{α} delle varie strutture in input e da queste costruire la matrice di Kirchhoff del GNM, illustrato in precedenza, per ognuna di esse. Si è quindi proseguito calcolando autovalori e autovettori della matrice Γ, dei quali si è escluso il primo (autovalore nullo riferito a un autovettore costante che indica una traslazione della struttura nella direzione (1,1,1), non interessante per lo studio delle proprietà dinamiche della struttura stessa) e dagli autovalori si sono stimate le frequenze dei modi di oscillazione per ciascuna struttura.

3.2 Localizzazione dei modi

Si sono poi utilizzate le frequenze dei modi normali di oscillazione per stimare la localizzazione per ciascun modo, e si sono costruiti dei grafici che riportassero in ascissa le frequenze e in ordinata i valori della localizzazione. I valori (adimensionali) delle frequenze in questa parte del lavoro sono stati calcolati più volte variando $R_{C_{EN}}$ tra i valori di $8, 10$ e $12\mathring{A}$ in modo da osservare i cambiamenti della localizzazione al variare di tale valore. L'andamento tipico è riportato nel seguente grafico.

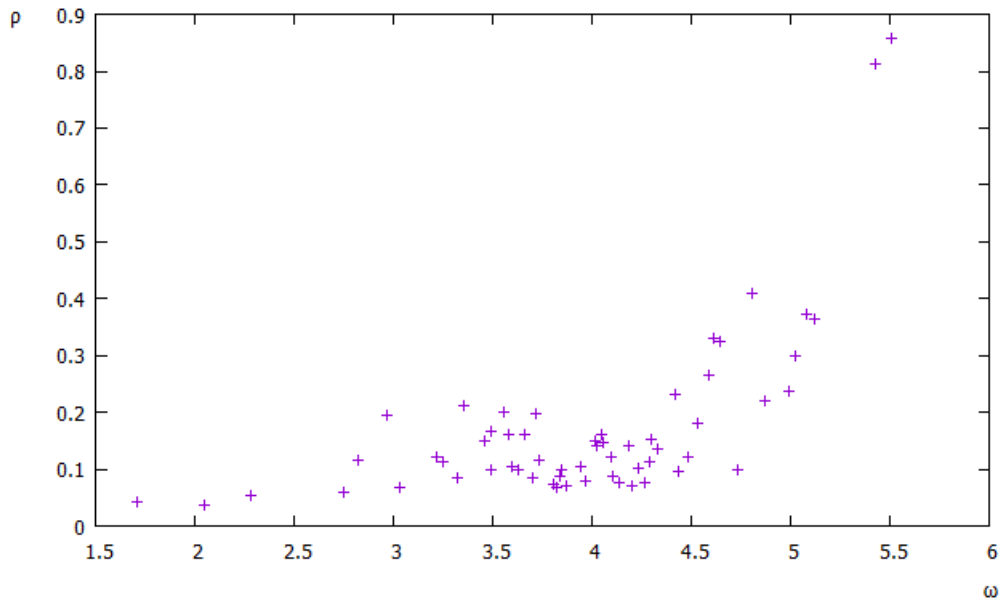

Figura 3.1: *Grafico della localizzazione ρ per la struttura di VAL60"VAL60_1" ottenuto con un valore di $R_{C_{EN}}$ pari a $10\mathring{A}$.*

Come si può notare i modi più lenti sono meno localizzati, come ci si aspetta: essi influenzano larghe parti della struttura come visto in *fig. 3.2* e dunque risultano essere poco localizzati. I modi con frequenza maggiore arrivano invece ad essere quasi completamente localizzati, come si può vedere in *fig. 3.3*.

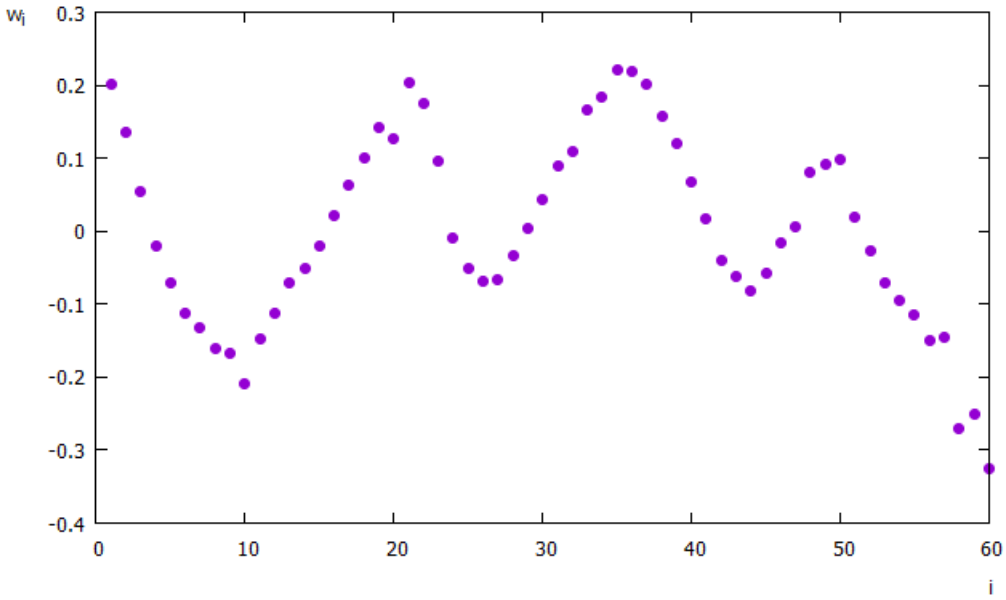

Figura 3.2: *Grafico dei valori delle componenti $\boldsymbol{w}_i$ dell'autovettore relativo al modo di oscillazione con frequenza più bassa, in funzione del residuo i corrispondente a ciascuno di esse, per la struttura di VAL60"VAL60_1" analizzata con un valore di $R_{C_{EN}}$ pari a $10\mathring{A}$.*

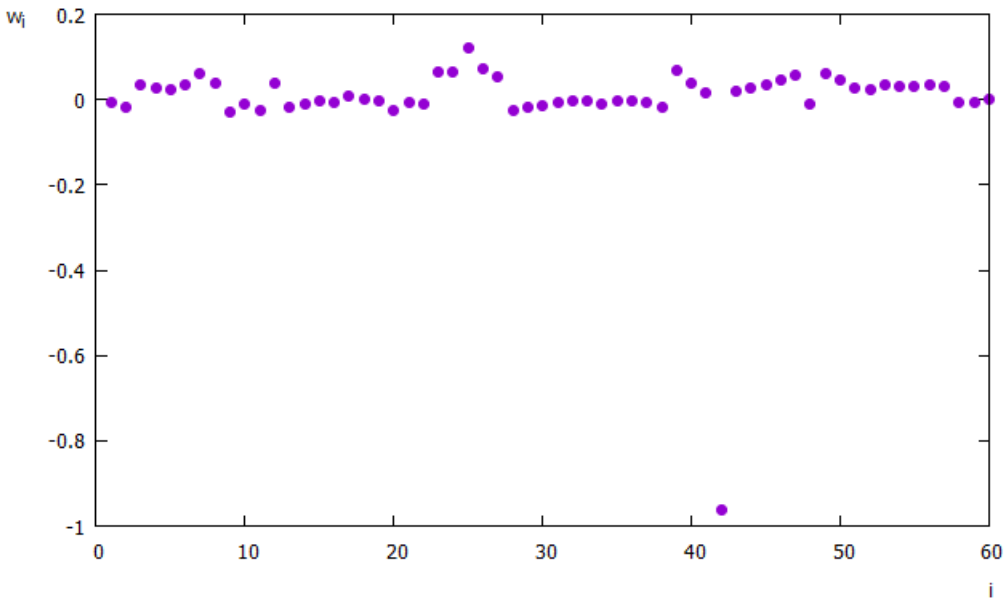

Figura 3.3: *Grafico dei valori delle componenti* $\boldsymbol{w}_i$ *dell'autovettore relativo al modo di oscillazione con frequenza più alta, in funzione del residuo i corrispondente a ciascuno di esse, per la struttura di VAL60 "VAL60_1" analizzata con un valore di* $R_{C_{EN}}$ *pari a* $10\mathring{A}$.

3.3 Correlazione tra Contact Order e localizzazione media

Si è poi proceduto con un'analisi quantitativa operata su tutte le strutture di VAL60 utilizzate nel corso dello studio, in cui si è scelto inizialmente di calcolare il *CO* per ciascuna di esse oltre che un valore di localizzazione mediato su tutti i suoi autovettori, graficandolo poi in funzione del Contact Order di ciascuna struttura. Da tale studio è emersa una possibile correlazione tra il *CO* e la localizzazione media per ciascuna struttura e dunque si sono variati i parametri di $R_{C_{EN}}$, $R_{C_{CO}}$ e il numero di autovettori su cui effettuare la media della localizzazione per trovare la configurazione in cui tale correlazione risultasse più evidente.

I valori utilizzati sono stati di 8, 10 e 12 $\mathring{A}$ per $R_{C_{EN}}$, 6, 7 e 8 $\mathring{A}$ per $R_{C_{CO}}$ e si sono calcolate le medie delle localizzazioni su 5, 10, 20 e tutti e 59 gli autovalori. Le combinazioni dei tre parametri sono state provate tutte in maniera automatizzata all'interno del programma.

Per ciascuna di queste combinazioni si sono calcolati i valori del coefficiente di Pearson, definito come:

$$r_{xy} = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} x_i y_i - \frac{1}{N^2} \cdot \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{(\frac{1}{N} \cdot \sum_{i=1}^{N} x_i^2 - \frac{1}{N^2} \cdot (\sum_{i=1}^{N} x_i)^2) \cdot (\frac{1}{N} \cdot \sum_{i=1}^{N} y_i^2 - \frac{1}{N^2} \cdot (\sum_{i=1}^{N} y_i)^2)}} \tag{3.1}$$

con x_i che nel nostro caso indica il CO dell'i-esima struttura, y_i il valore medio della localizzazione per l'i-esima struttura e N che indica il numero delle strutture su cui si sono calcolati tali dati (300 nel nostro caso).

Oltre al coefficiente di Pearson, per ogni combinazione si è calcolato anche l'errore relativo della pendenza della retta interpolante $y = \hat{m} \cdot x + \hat{q}$, calcolato come segue:

$$s_{\hat{m}_r} = \frac{s_{\hat{m}}}{\hat{m}} \tag{3.2}$$

con $\hat{m}$ pendenza della retta interpolante, calcolata come segue:

$$\hat{m} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N} \cdot (\sum_{i=1}^{N} x_i)^2} \tag{3.3}$$

e con $s_{\hat{m}}$ l'errore standard sulla pendenza, definito come

$$s_{\hat{m}} = \sqrt{\frac{\frac{1}{N-2} \cdot \sum_{i=1}^{N} \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{N} (x_i - \frac{1}{N} \cdot \sum_{i=1}^{N} x_i)^2}} \tag{3.4}$$

dove gli $\hat{\varepsilon}_i$ sono i residui, calcolati con la seguente formula:

$$\hat{\varepsilon}_i = y_i - (x_i \cdot \hat{m} + \hat{q}) \tag{3.5}$$

con $\hat{q}$ calcolato come

$$\hat{q} = \frac{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N \cdot \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \tag{3.6}$$

Tra tutte le combinazioni si sono quindi calcolati i valori di questi due parametri e si è cercata quella che massimizzasse il modulo di r_{xy} (per il quale si ottengono valori negativi e dunque una correlazione inversa tra CO e il valore medio della localizzazione) e minimizzasse il modulo di $s_{\hat{m}_r}$. Tale configurazione è risultata essere quella con $R_{C_{EN}}$ pari a 10 $\mathring{A}$, $R_{C_{CO}}$ pari a 8 $\mathring{A}$ e una media della localizzazione effettuata sui primi 20 autovettori.

In questa configurazione si sono ottenuti il valore più alto per r_{xy}, di 0,37022 (che è indice di una discreta correlazione tra i due parametri), e più basso per $s_{\hat{m}_r}$, di 0,145352.

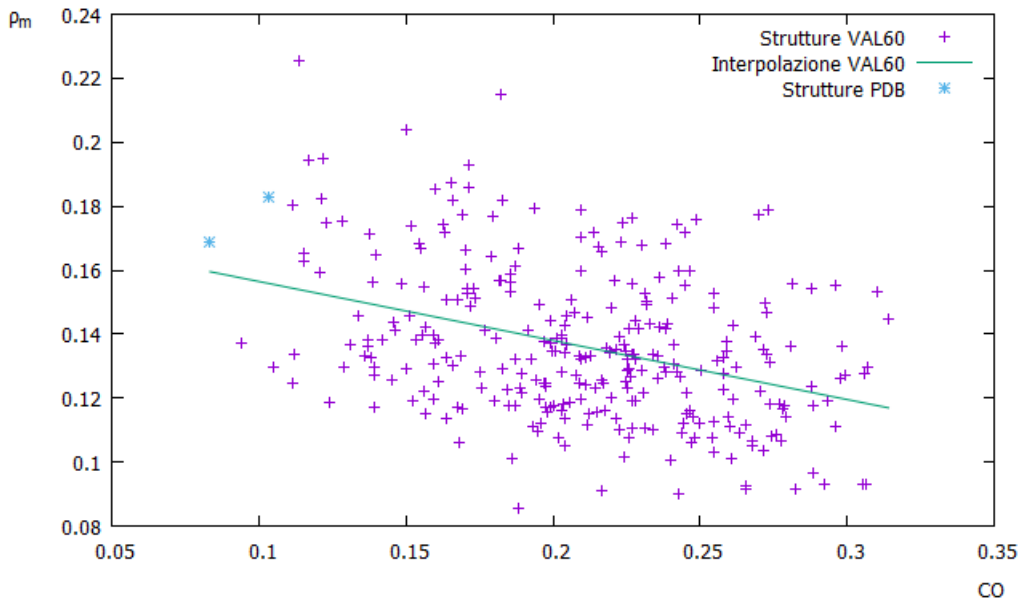

Figura 3.4: *Grafico dei valori di CO e localizzazione media per le 300 strutture di VAL60 e per le due strutture del PDB "1rq6A00" e "1vq0A02", studiate nella configurazione con $R_{C_{EN}}$ pari a 10 Å, $R_{C_{CO}}$ pari a 8 Å e 20 autovettori considerati per il calcolo della localizzazione. In grafico è mostrata anche la retta interpolante per i valori delle strutture di VAL60.*

Facciamo notare infine un ultimo elemento emerso nello studio del *CO* delle varie strutture. Mentre le strutture di polivalina hanno un *CO* abbastanza alto (generalmente compreso tra 0,1 e 0,3 come si può notare in *fig. 3.4*) i valori di Contact Order per strutture proteiche prese dal PDB analizzate in questo lavoro risultano essere più bassi, attorno allo 0,1 o inferiori. Visto che questo combacia con i risultati ottenuti in [4] e con le aspettative della teoria (come visto nel *Capitolo 1*), può essere preso come un indice della bontà dei calcoli svolti nello stimare il *CO* nel presente lavoro.

3.4 Distribuzione delle frequenze dei modi di oscillazione e dimensione spettrale

Nell'ultima parte del lavoro svolto si sono infine analizzate le densità di modi normali di oscillazione per le strutture considerate, al fine di stimarne la dimensione spettrale.

Dopo aver calcolato le frequenze dei modi normali di oscillazione, esse sono state istogrammate suddividendo i valori compresi tra il modo con frequenza minima e quello con frequenza massima in una ventina di intervalli, sufficienti ad avere qualche entrata per canale, e da tale istogramma si è costruita la PDF delle frequenze ($f(\omega)$, funzione che rappresenta il numero di entrate per ogni canale). Dalla PDF si è risaliti alla cumulativa CDF ($G(\omega)$, che calcola per ogni intervallo il numero di frequenze che ricadono nello stesso, sommate al numero

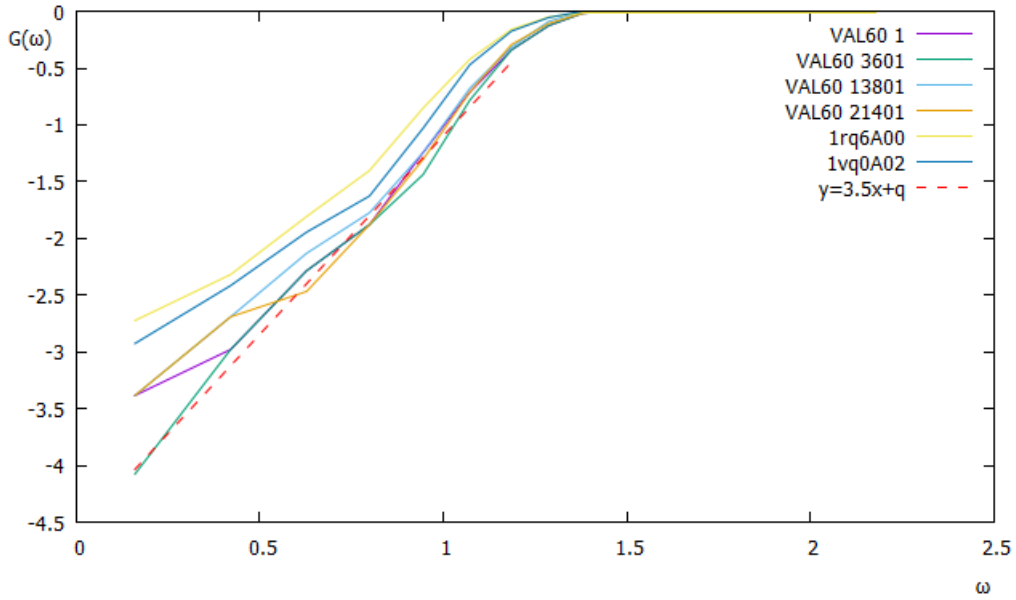

Figura 3.5: *Andamento delle $G(\omega)$ per alcune strutture. Le due curve più alte rispetto alle altre sono relative a due strutture del PDB "1rq6A00" e "1vq0A02", mentre le altre quattro sono le cumulative delle strutture di VAL60 "VAL60_1", "VAL60_3601", "VAL60_13801" e "VAL60_21401". In questo caso le CDF sono calcolate utilizzando $R_{C_{EN}}$ pari a 8 Å. Insieme alle $G(\omega)$ è rappresentata anche una retta con pendenza pari a 3.5 che come si può vedere ha un andamento molto simile alle cumulative delle strutture.*

di entrate dei canali precedenti), dalla quale si è potuta stimare la dimensione spettrale d per ciascuna struttura. In queste pagine alcuni grafici delle $G(\omega)$ ottenute. In tali grafici sono state utilizzate scale log-log per entrambi gli assi; l'asse x indica le frequenze, mentre l'asse y il valore di $G(\omega)$ per ogni intervallo. Come si può notare, la pendenza delle rette aumenta all'aumentare di $R_{C_{EN}}$ (infatti passa da un valore di circa 3.5 per $R_{C_{EN}} = 8\mathring{A}$ a un valore di 6 per $R_{C_{EN}} = 12\mathring{A}$). Ciò significa che la dimensione spettrale d aumenta all'aumentare del valore di $R_{C_{EN}}$ considerato. Tali risultati sono in accordo con l'andamento riscontrato in [10] per valori di $R_{C_{EN}}$ più piccoli.

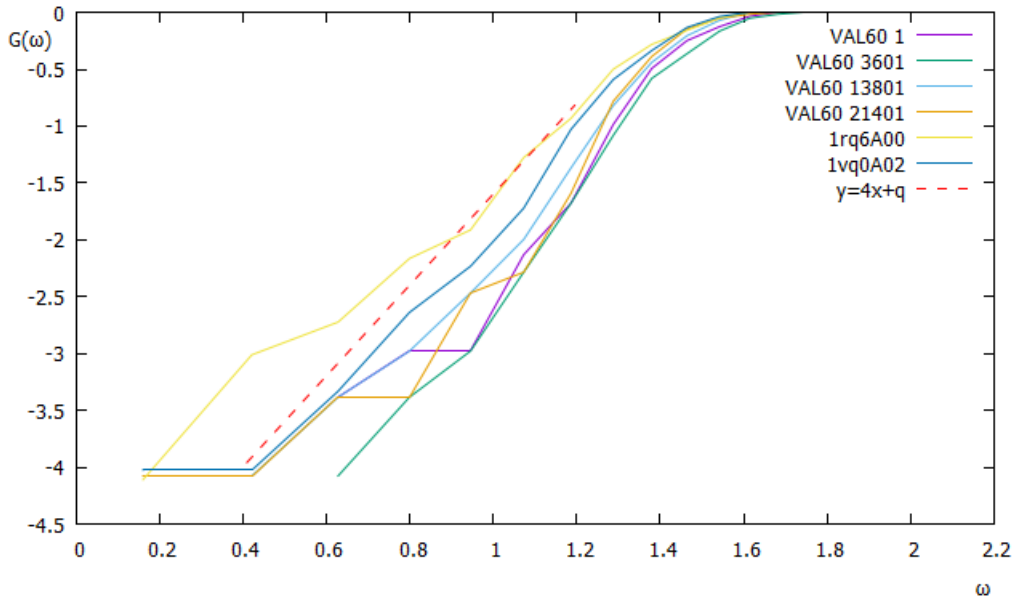

Figura 3.6: *Andamento delle $G(\omega)$ per le stesse strutture viste sopra. In questo caso le CDF sono calcolate utilizzando $R_{C_{EN}}$ pari a 10 Å, più interessante del precedente visto che questo valore di $R_{C_{EN}}$ era risultato quello migliore per la correlazione tra CO e localizzazione media. Come si può notare, la pendenza della retta è cambiata per adattarsi alle nuove $G(\omega)$.*

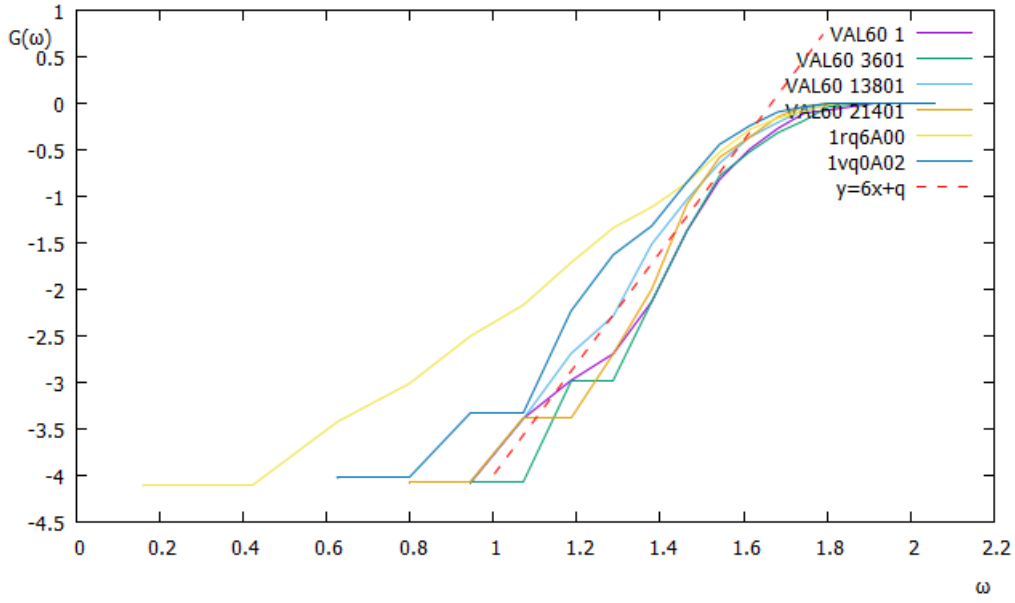

Figura 3.7: *Andamento delle $G(\omega)$ per le medesime strutture. In questo caso le CDF sono calcolate utilizzando $R_{C_{EN}}$ pari a 12 Å. La pendenza cresce ancora.*

Capitolo 4

Conclusione

Dall'analisi svolta sulle strutture considerate in questo lavoro sono emersi dunque risultati interessanti.

Come visto nel capitolo precedente, in accordo con le aspettative teoriche si sono osservati valori della localizzazione ρ più bassi per modi normali con frequenza più bassa, segno che il metodo utilizzato per calcolare ρ è efficace in quanto fornisce risultati compatibili con le aspettative: modi di oscillazione che interessano grandi porzioni della struttura (quindi con autovalori più bassi e dunque frequenze più basse) risultano essere meno localizzati dei modi con frequenza più alta.

Assai più importanti sono i risultati ottenuti dallo studio della localizzazione media in funzione del CO per le strutture di VAL60. Come visto è stata evidenziata la possibile presenza di una correlazione inversa tra ρ_m e CO, con la conseguenza che a strutture con valori più bassi di CO corrispondono valori di ρ_m più alti (ossia strutture con Contact Order più basso hanno modi di oscillazione più localizzati).

Questo risulta controintuitivo: a modi lenti delocalizzati corrispondono importanti funzionalità biologiche delle proteine esistenti in natura, e le proteine del PDB compaiono tra le strutture di VAL60 con valore di CO più basso (e dunque dotati di modi di oscillazione più localizzati). Ciò lascia spunti per analisi future che studino il legame tra CO e localizzazione media per strutture del PDB in modo da verificare o smentire quanto trovato per strutture di VAL60, ed eventualmente che relazione possa esserci tra strutture di VAL60 e strutture del PDB.

Infine è stata stimata la dimensione spettrale per le strutture di VAL60, verificando un andamento già proposto per strutture proteiche esistenti: all'aumentare del valore di $R_{C_{EN}}$ si osserva anche l'aumentare della dimensione spettrale d. Da notare come per una struttura del PDB (1rq6A00) la dimensione spettrale aumenti con la connettività topologica in maniera meno marcata rispetto alle strutture di VAL60 considerate. In ogni caso $N = 60$ è una lunghezza molto piccola che rende difficile l'analisi del comportamento della $G(\omega)$ per frequenze piccole ed in particolare la possibile verifica di Eq. (2.17).

Da quanto emerso pare quindi che il comportamento delle strutture di VAL60 sia analogo a quello delle strutture esistenti in natura per quanto ri-

guarda il comportamento della localizzazione al variare delle frequenze dei modi normali e per quanto concerne l'andamento della dimensione spettrale d al variare di $R_{C_{EN}}$; interessante sembra essere invece il rapporto tra localizzazione media e CO delle strutture considerate, cosa che spinge a proseguire gli studi con gli stessi metodi finora applicati all'analisi di strutture del PDB.

Appendice A

Bibliografia

[1]Dror Tobi, "Dynamics alignment: Comparison of protein dynamics in the SCOP database", Wiley Periodicals, Proteins 2012, volume 80:1167-1176.
[2] David L. Nelson, Michael M. Cox, "Lehninger Principles of Biochemistry", W. H. Freeman, 4th edition (April 23, 2004).
[3] Alexei V. Finkelstein, Oleg B. Ptitsyn, "Protein Phisycs A Course of Lectures", Academic press, 2002.
[4] Antonio Trovato *et* al, "Exploring the Universe of Protein Structures beyond the Protein Data Bank", PLoS Computational Biology, November 4, 2010, volume 6, issue 11.
[5] Tirion MM. "Large amplitude elastic motions in protein from a single-parameter, atomic analysis", Physical Review Letters, 1996.
[6] Cristian Micheletti, "Comparing proteins by their internal dynamics: Exploring structure-function relationships beyond static structural alignments", Physics of Life reviews 10 (2013),1-26, SciVerse ScienceDirect.
[7] J. Andrew McCammon, Stephen C. Harvey, "Dynamics of proteins and nucleic acids", Cambridge University Press, 1987.
[8] I. Bahar *et* al, "Proteins with Similar Architecture Exhibit Similar Large-Scale Dynamic Behavior", Biophisical Journal Volume 78, April 2000, 2093-2106.
[9]Micheletti C, Lattanzi G, Maritan A, Journal of Molecular Biology 321, 909-921 (2002). "Elastic Properties of Proteins: Insight on the Folding Process and Evolutionary Selection of Native Structures".
[10] Fabio Cecconi *et* al, "Topological Thermal Instability and Lenght of Proteins", PROTEINS: Structure, Function, and Bioinformatics, 2004, volume 55:529-535.