UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA "GALILEO GALILEI"
Corso di laurea in Fisica

CALCOLO DI DESCRITTORI DI DINAMICA MOLECOLARE E LORO APPLICAZIONE AI CANALI DI CONNESSINA

Relatore:
Prof. MARIO BORTOLOZZI

Laureando:
MATTEO ABBATE

Correlatore:
Dott. DAMIANO BURATTO

ANNO ACCADEMICO 2013/2014

Uno dei primi giorni di lezione del corso di Fisica, il professore ci disse che un buon fisico non è colui che sa trovare immediatamente tutte le risposte, ma colui che si sa porre domande: non esistono domande giuste o sbagliate, forse domande più o meno opportune. Non sono assolutamente certo di essere un buon fisico al momento, ma senza dubbio in questi anni mi sono posto a modo mio molte domande.

Dedico questa tesi a tutti coloro i quali le hanno sapute, con pazienza e con affetto, ascoltare: grazie!

Indice

Introduzione

100 000 miliardi di cellule costituiscono il corpo umano. Una cellula è l'unità funzionale che caratterizza tutti gli esseri viventi, dai primi organismi monocellulari agli organismi pluricellulari più complessi, in cui ogni cellula si specializza, dando origine ai tessuti, agli organi e agli apparati. Una cellula ha le dimensioni di qualche micrometro ed è racchiusa da una membrana plasmatica al cui interno è presente il citoplasma, una sostanza salina, il citoscheletro, una struttura alla quale sono ancorati vari organuli che adempiono alle diverse funzioni della cellula, e il nucleo, in cui è conservata l'informazione genetica, necessaria alla riproduzione della cellula stessa.
La membrana plasmatica ha quindi il fondamentale compito di separare e proteggere la cellula dall'ambiente esterno, ma anche quello di mettere in comunicazione la cellula con ciò che la circonda, permettendo lo scambio di sostanze quali ioni, metaboliti, nucleotidi e piccoli peptidi. Ciò avviene tramite la creazione di stretti canali di comunicazione tra cellule differenti, che possono così scambiare i propri prodotti e regolare i propri potenziali in modo tale da adempiere alle funzioni delle strutture che costituiscono.
I canali più comuni prendono il nome di canali di giunzione gap: ogni canale è costituito da due unità esameriche chiamate emicanali o *connessoni*, ciascuna delle quali si origina in una delle due cellule, si integra nella membrana plasmatica e si collega all'altra nello spazio extracellulare. Ogni emicanale è a sua volta costituito da sei connessine, ossia delle proteine, molecole biologiche a loro volta caratterizzate da una particolare sequenza di unità monomeriche dette amminoacidi.

Una mutazione genetica che comporta la sostituzione di uno solo di questi amminoacidi nelle connessine può causare diverse malattie nell'uomo: malattie neurodegenerative, malattie dermatologiche, sordità sindromiche e non, anomalie nello sviluppo. Per capire come ciò accade può rivelarsi utile analizzare non solo come varia la sequenza amminoacidica o la struttura delle connessine, metodi con i quali tradizionalmente si classificano le proteine, ma anche come muta la loro dinamica. A causa del moto di agitazione termica e ad altri potenziali cui sono soggetti i singoli atomi, infatti, ogni proteina ha una sua dinamica, caratterizzata da una mobilità nello spazio limitata e movimenti delle strutture interne ricorrenti. Le leggi che governano tale moto sono di tipo browniano.
Negli ultimi anni, è stato proposto un metodo del tutto innovativo per riuscire ad analizzare tale dinamica nella sua globalità ed in modo universale, metodo che ha ancora poche applicazioni e che consiste nel calcolo di 34 variabili caratteristiche di vari aspetti della dinamica stessa. Il set delle 34 variabili prende il nome di *dinasoma* ed è stato possibile classificare le proteine in base alle loro dinamiche proprio rappresentandole come vettori in tale spazio 34-dimensionale dei dinasomi.
Lo scopo di questa tesi è quello di studiare tale metodo da un punto di vista teorico e di sviluppare dei metodi computazionali per poter calcolare, attraverso anche l'uso di pacchetti di programmi di analisi di dinamica molecolare, ciascuna delle 34 variabili del dinasoma nel caso degli emicanali di membrana.

Applicando tali metodi alla dinamica dell'emicanale costituito dalle connessine 26 e a quella dell'emicanale costituito da alcune forme mutate della stessa connessina, se ne farà un confronto. Si osserveranno delle significative differenze non tanto nei profili di energia della dinamica o nelle fluttuazioni relative ai singoli atomi del connessone, quanto piuttosto nelle grandezze che

caratterizzano i moti d'insieme degli atomi. Il calcolo del dinasoma si rivelerà quindi essere un metodo sufficientemente sensibile, da un punto di vista biofisico, per comprendere che effettivamente una variazione della dinamica esiste e per darne una descrizione qualitativa e quantitativa.
Queste prime osservazioni potranno poi essere un primo risultato per comprendere come tali mutazioni influiscono sulla funzionalità del canale e quindi dell'intera cellula, fino a diventare le maggiori cause della sordità genetica e di altre malattie, inficiando gravemente lo sviluppo salutare, personale e sociale dell'individuo.

Primo capitolo

Struttura e classificazione delle proteine

Le proteine sono una classe di molecole biologiche coinvolte in molti processi cellulari. Esse giocano ruoli fondamentali nel trasporto, nella struttura, nell'attività enzimatica e altre funzioni della cellula. Formano le basi per le maggiori componenti strutturali dei tessuti animali e umani e sono eteropolimeri naturali le cui unità fondamentali sono gli amminoacidi,
Gli amminoacidi sono composti da un gruppo amminico, da un carbonio centrale detto Carbonio-α (C^{α}), da un gruppo carbossilico (i cui atomi di Carbonio e Ossigeno indicheremo d'ora in poi con C' e O) e da un residuo. Un residuo è una catena che si collega al Carbonio-α e caratterizza il tipo specifico e le proprietà dell'amminoacido. La classificazione degli amminoacidi si basa quindi sul residuo e sulle sue proprietà chimico-fisiche. Nonostante siano conosciuti centinaia di amminoacidi diversi, solo una piccola parte di loro (circa venti) si trova realmente nelle proteine.
La sintesi delle proteine avviene in due fasi: l'informazione genetica, contenuta nel DNA, viene prima trascritta in un RNA messagero (mRNA) all'interno del nucleo e successivamente viene tradotta in sequenza amminoacidica nei ribosomi. Gli amminoacidi vengono legati a molecole di RNA di trasporto (tRNA), che entrano in una regione del ribosoma e si legano alla sequenza dell'RNA messaggero. Gli amminoacidi così fissati vengono poi legati insieme tra loro. Il ribosoma si muove lungo l'mRNA, “leggendo” la sua sequenza e producendo una catena di amminoacidi. Il DNA è una sequenza costituita da quattro differenti basi (adenina, citosina, guanina, tirosina) e una sequenza tre basi (*codone*) codifica per un singolo amminoacido. Anche se un codice di tre basi potrebbe produrre potenzialmente sessantaquattro differenti sequenze, il codice è ridondante, ossia più codici codificano per lo stesso amminoacido e ciò spiega la presenza di soli venti diversi amminoacidi nelle proteine (Figura 1.1). Si ritiene che ciò contribuisca a rendere il codice del DNA meno sensibile a una mutazione random. È possibile trovare anche dei residui “non-standard” nelle proteine, dovuti alle modifiche che avvengono nel ribosoma dopo la traduzione del codice e che potrebbero dipendere dal tipo di amminoacido e dalla sua posizione. Una lista dei venti amminoacidi standard è riportata in Figura 1.2. Solitamente, ciascun amminoacido è identificato con una o tre lettere.

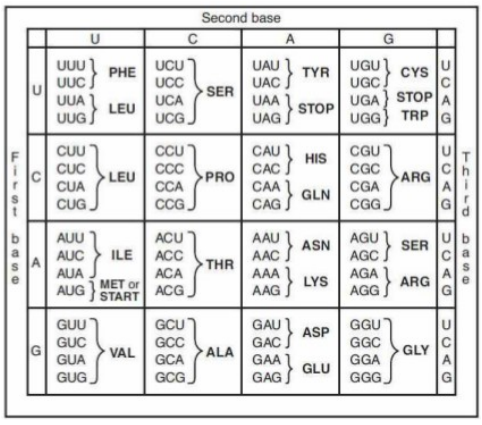

Figura 1.1. *Tabella del codice genetico: codoni di RNA e amminoacidi per i quali codificano.*

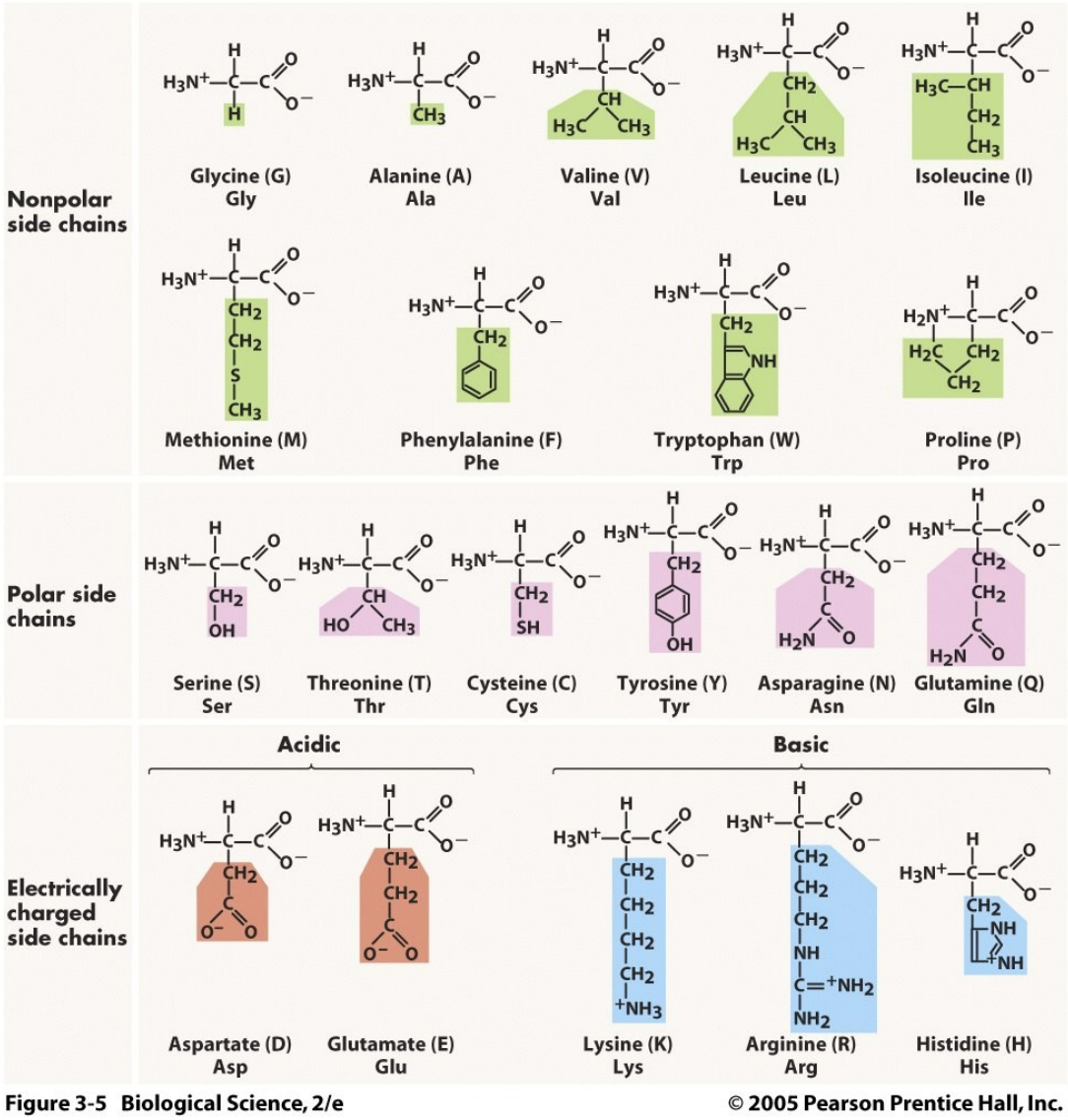

Figura 1.2. *I venti amminoacidi standard: i residui sono evidenziati con colori diversi a seconda delle loro proprietà.*

Essi sono solitamente divisi in quattro classi dipendenti dalla polarità e dalla carica:

1. Amminoacidi carichi negativamente:
   - Acido aspartico (Asp, D)
   - Acido glutammico (Glu, E)

2. Amminoacidi carichi positivamente:
   - Lisina (Lys, K)
   - Arginina (Arg, R)
   - Istidina (His, H)

3. Amminoacidi polari:

- Asparagina (Asn, N)
- Cisteina (Cys, C)
- Glutammina (Gln, Q)
- Serina (Ser, S)
- Tirosina (Tyr, Y)
- Treonina (Thr, T)

4. Amminoacidi apolari:
- Alanina (Ala, A)
- Fenalanina (Phe, F)
- Glicina (Gly, G)
- Isoleucina (Ile, I)
- Leucina (Leu, L)
- Metionina (Met, M)
- Prolina (Pro, P)
- Triptofano (Trp, W)
- Valina (Val, V)

Con l'eccezione della glicina, tutti gli amminoacidi sono monomeri chirali. In tutte le proteine biologiche è presente solo la forma levogira.
Il gruppo amminico di un amminoacido può reagire col gruppo carbossilico di un altro amminoacido rilasciando una molecola d'acqua e formando un legame chiamato *legame peptidico o ammidico.* Il C^{α}, il C' e l'N formano quindi una catena che prende il nome di *scheletro carbonioso* o *backbone* della proteina.
L'energia necessaria per la formazione del legame dipende dagli amminoacidi coinvolti e il legame di un singolo amminoacido ad una catena già formata è più favorito rispetto al legame a un altro singolo amminoacido, risultando minore l'energia di attivazione della reazione. Valori tipici sono dell'ordine di una frazione di alcune Kcal/mol [1]. La lunghezza di legame è stata determinata da tecniche di cristallografia a raggi X e si aggira intorno a 1,325 °A, significativamente più corta rispetto a quella di un singolo legame N-C [2].
Il gruppo ammidico ha due forme di risonanza (Figura 1.3), che gli conferiscono molte importanti proprietà. La risonanza suggerisce che il gruppo ammidico abbia un caratteristico doppio legame parziale: in condizioni tipiche, il legame è singolo e non carico con una probabilità di circa il 60%, mentre si trova un doppio legame con una probabilità di circa il 40%. Il gruppo peptidico è non carico a tutti i normali valori di pH, ma la sua forma di doppio legame risonante gli conferisce un considerevole e insolito momento di dipolo, di approssimativamente 3,5 Debye (0,7 electron-angstrom). Questi momenti di dipolo possono allinearsi in specifiche strutture secondarie (come le α-eliche, discusse di seguito), producendo un grande e netto dipolo.

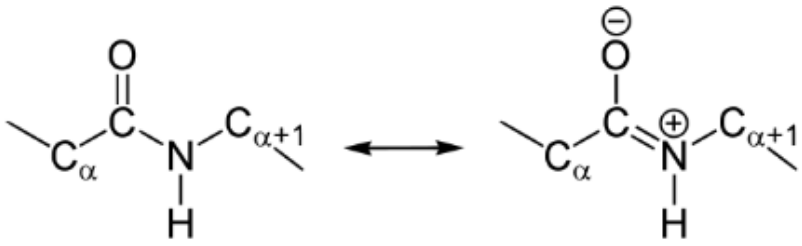

Figura 1.3. *Forma di risonanza del legame peptidico: il legame è una sovrapposizione dei due stati, così che gli atomi di O, C, N e H giacciono sullo stesso piano.*

Il caratteristico doppio legame parziale può essere rafforzato o indebolito da modificazioni che favoriscono una forma di risonanza a un'altra. Per esempio, la forma doppiamente legata è sfavorita in ambienti idrofobici, a causa della sua carica.
Il doppio legame parziale rende il gruppo ammidico quasi planare. Il cosiddetto angolo diedrico del legame peptidico (ω), definito dagli atomi C^{α} - C' - N - C^{α}, non differisce mai tanto da 180°. Questa conformazione è chiamata forma *trans* ed è fortemente favorita. C'è un secondo importante minimo nell'energia corrispondente a ω = 0°, chiamato isoforma. Una transizione tra le due forme può avvenire, ma è energeticamente difficile poiché la barriera di potenziale è stimata essere di circa 20 kcal/mol.
Questa rigidità riduce significativamente i gradi di libertà della catena, permettendo rotazioni libere solo intorno ai legami N- C^{α} e C^{α}-C'. Queste rotazioni sono parametrizzate da due angoli, detti angoli di Ramachandran e indicati rispettivamente con le lettere Φ e Ψ (Figura 1.4). Quindi, solo 2N parametri sono lasciati liberi in una catena di N amminoacidi. La presenza del residuo tuttavia limita i valori assumibili dagli angoli di Ramachandran, riducendo quindi definitivamente la configurazione spaziale accessibile.

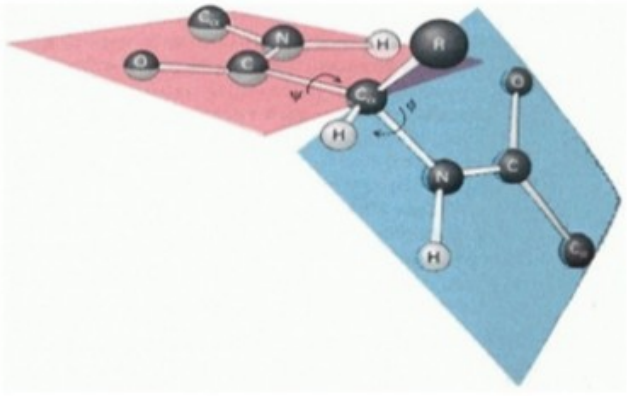

Figura 1.4. *Angoli di Ramachandran.*

Struttura delle proteine

La sequenza degli amminoacidi in una proteina è indicata come la sua struttura primaria. Per convenzione, si riporta partendo dal terminale amminico (N) e terminando in quello carbossilico

(C). Essa non dà alcuna informazione sulla posizione spaziale dei residui o degli atomi.
Spesso accade che gruppi di residui si organizzino localmente tra loro in strutture come eliche o foglietti. Queste strutture sono rese stabili da legami idrogeno e le più frequenti sono le α-eliche e i foglietti β.
I monomeri nelle α-eliche (Figura 1.5) hanno una distanza assiale *p* di 1,5°A e il rapporto tra il passo dell'elica e *p* è 3,6. Sono quindi necessari cinque giri e diciotto monomeri prima che due monomeri esattamente sovrapposti si incontrino nuovamente. L'angolo di rotazione tra due successive unità monomeriche è 100° e i valori degli angoli di Ramachandran sono gli stessi per tutti i residui formanti l'elica. La struttura a elica è stabilizzata da legami idrogeno paralleli agli assi tra il gruppo C = O dell'i-esimo monomero e il gruppo N – H dell' (i+4)-esimo monomero.

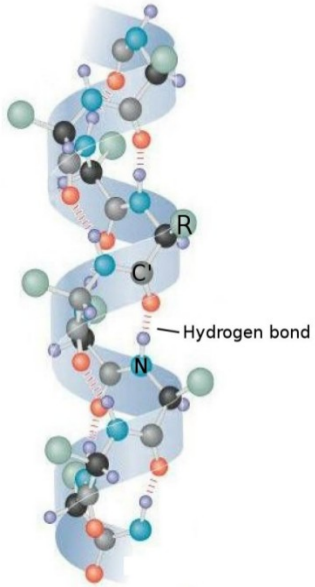

***Figura 1.5**. Struttura ad alpha elica: atomi di diverso tipo sono indicati con colori diversi (i Carbonio-alpha sono rappresentati in nero). I residui sono mostrati schematicamente come un singolo atomo R. I due tipi di legami idrogeno nelle alpha eliche sono rappresentati da linee tratteggiate.*

I foglietti β (Figura 1.6) sono strutture in cui due o più parti della catena sono legati da legami idrogeno, formando quindi dei piani. Le singole parti sono note come filamenti β e possono essere organizzati anche in direzione parallela o anti-parallela. I residui si organizzano alternativamente in uno dei lati del foglietto.

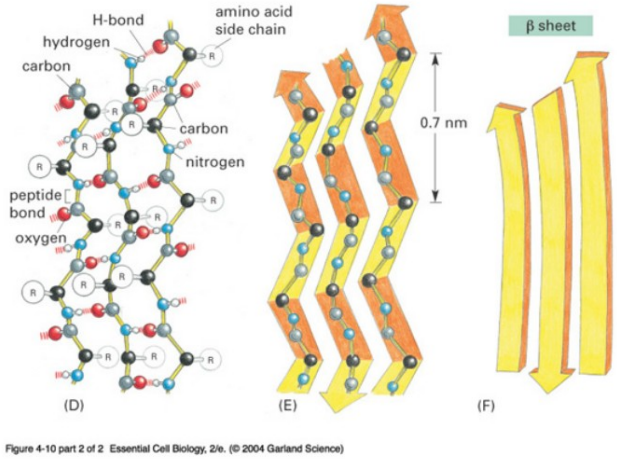

Figura 1.6. *Schematica rappresentazione di foglietti beta antiparalleli. I filamenti beta sono rappresentati da frecce.*

Inoltre, sono classificate anche altre strutture meno comuni come eliche 3-10, β-turns, etc. Regioni non strutturate sono dette conformazioni casuali. La sequenza di questo tipo di strutture (eliche, foglietti, conformazioni casuali) nella proteina è indicata come struttura secondaria. La globale organizzazione tridimensionale delle posizioni degli atomi è chiamata struttura terziaria.
La relazione tra struttura primaria e terziaria è molto complessa: sebbene siano strettamente dipendenti, molte strutture primarie, a volte molto differenti l'una dall'altra, possono riprodurre lo stesso modello tridimensionale e viceversa. Analogamente, la relazione tra la funzione e la struttura è controversa, sebbene siano state a lungo considerate profondamente legate tra loro. La questione del ripiegamento, cioè predire la struttura terziaria dalla sequenza amminoacidica, può essere considerata uno dei più importanti problemi aperti nella biologia odierna. Finora, i migliori risultati si stanno ottenendo da metodi empirico-statistici e non consentono nessun approccio da parte dell'ingegneria proteica. La soluzione a questo tipo di problema o di quello inverso sarebbe di importanza storica nella biologia molecolare e un punto di svolta fondamentale nell'industria farmaceutica.
In alcune proteine la struttura tridimensionale può mostrare importanti regioni strutturate, connesse attraverso catene relativamente flessibili. In questo caso, le regioni rigide sono spesso chiamate domini.
Infine, diverse proteine possono agganciarsi insieme per formare un'unica unità fondamentale. Questa è chiamata struttura quaternaria. L'emicanale *connessone* studiato in questa tesi, che è composto da sei proteine legate insieme dette *connessine*, è un esempio di struttura quaternaria.

Secondo capitolo

Studio della dinamica di una proteina: il dinasoma

Come già osservato, lo studio e la classificazione delle proteine si basano sulla loro sequenza amminoacidica, la loro struttura e la loro funzione. Diversi studi hanno messo in evidenza come la struttura proteica dipenda strettamente dalla sequenza amminoacidica[3].
Nell'analisi della funzione delle proteine è stato però verificato che proteine con sequenze simili possono avere funzioni diverse e, viceversa, proteine con stessa funzione e simile struttura possono avere sequenze amminoacidiche differenti. Allo stesso modo, l'individuazione della struttura, che può essere vista come un intermediario tra la sequenza e la funzione, non permette di predire con esattezza la funzione della proteina [4].
A causa del moto di agitazione termica e ad altri potenziali cui sono soggetti i singoli atomi, ogni proteina ha una dinamica, caratterizzata da una mobilità nello spazio limitata e movimenti delle strutture secondaria e terziaria ricorrenti. Di recente, sta assumendo sempre più importanza lo studio delle relazioni tra la dinamica della proteina e la sua funzione e le conseguenti corrispondenze tra le classificazioni basate sulla struttura delle proteine e le possibili classificazioni dovute alla loro dinamica.
L'analisi della dinamica di una proteina si può studiare attraverso simulazioni computazionali di dinamica molecolare. Per ottenere una simulazione della dinamica della proteina, sono stati sviluppati diversi metodi, divisibili in tre classi: quantum-mechanical simulations (QM), che descrive le traiettorie delle particelle associando a ciascuna di esse una funzione d'onda ed è quindi più dettagliato, ma ha un costo computazionale maggiore (tra questi metodi hanno particolare importanza gli *ab initio methods*, tra cui il metodo Car-Parrinello, che studiano la dinamica dei singoli elettroni di valenza); classical molecular dynamics (MD), ove ogni atomo è considerato come un'unità singola con un comportamento classico, soggetta a campi di forze e la cui dinamica è ricavata dall'integrazione delle equazioni del moto newtoniane; coarse grained simulations (CG), che utilizza il raggruppamento di atomi e strutture, mappando una dinamica complessiva. Spesso vengono inoltre utilizzati metodi ibridi, che utilizzano in parte ciascuno dei metodi appena descritti.
In questo lavoro di tesi useremo traiettorie ricavate tramite metodi di classical molecular dynamics, per cui nella discussione successiva considereremo solo questi.
Ogni simulazione di dinamica molecolare richiede come input un set delle coordinate iniziali e, facoltativamente, delle velocità iniziali di tutte le particelle coinvolte. Le coordinate sono spesso ottenute da tecniche di cristallografia a raggi X o da simulazioni quanto-meccaniche. Se la configurazione iniziale è lontana dall'equilibrio è necessario applicare un processo di minimizzazione dell'energia. Quando l'input viene letto, il potenziale di interazione è calcolato come una funzione della posizione degli atomi e successivamente le forze sono calcolate come il gradiente del potenziale, rendendo possibile risolvere numericamente l'equazione del moto di Newton per piccoli intervalli di tempo. La parametrizzazione delle interazioni prende il nome di *force field*. La forza su ciascun atomo $F_i = -\frac{\partial V}{\partial r_i}$ viene calcolata come somma delle singole forze *non-bonded*, tra le coppie di atomi non legati $F_i = \sum_j F_{ij}$ (forze elettromagnetiche e interazioni Van der Waals), più le forze *bonded*, presenti tra le coppie di atomi legati, che possono essere causate da 1, 2, 3 o 4 atomi e dipendono dalla lunghezza del legame, un angolo di legame ed uno di torsione e da un improprio termine *diedrico*. Inoltre, è possibile applicare forze esterne per

riprodurre dei particolari processi (si parla in questo caso di *steered molecular dynamic*). Essendo le dimensioni del sistema simulato limitate, è necessario imporre delle condizioni periodiche ai bordi, supponendo che il sistema stesso si ripeta periodicamente intorno a sé. Se il sistema è cristallino, queste condizioni sono quelle desiderate, se è un liquido o una soluzione, invece, tale approssimazione causa degli errori, che risultano comunque meno gravi di quelli che deriverebbero da un innaturale confinamento nel vuoto.
Anche se l'uso diretto della dinamica molecolare darebbe luogo a delle simulazioni in cui si considera un ensemble microcanonico di particelle, il più delle volte l'ensemble canonico è più adatto per simulare sistemi biologici realistici. In esso vengono mantenuti costanti il numero di particelle, il volume e la temperatura, calcolata a partire dall'energia cinetica e controllata da un termostato virtuale, oppure la pressione, calcolata a partire dalla stima del tensore degli sforzi e controllata da un barostato virtuale.
Per il calcolo delle forze *non-bonded* si considera che ogni particella può interagire con tutte le altre, dando luogo a $N^2/2$ interazioni e per ottimizzare il calcolo dei potenziali, in caso di condizioni periodiche ai bordi, è possibile usare il metodo Particle Mesh Ewald, un metodo basato sull'idea centrale di dividere il potenziale in due componenti, una a corto e l'altra a lungo raggio d'azione: per raggiungere una velocità computazionale maggiore, la prima viene calcolata nello spazio diretto, la seconda nello spazio reciproco, composto dalle trasformate di Fourier delle funzioni utilizzate.
Una simulazione di questo tipo di una singola proteina può dare numerose informazioni riguardo le interazioni tra i suoi atomi. In questo modo vengono però omesse le interazioni con l'ambiente esterno, mentre gli effetti della solvatazione potrebbero essere importanti allo stesso modo. Per riprodurre tali effetti, si utilizzano delle modellizzazioni dell'acqua, che possono essere esplicite o implicite: le prime introducono effettivamente molecole d'acqua nel sistema, mentre le seconde prevedono di introdurre la proteina in un mezzo dielettrico continuo.
In questo modo è possibile quindi simulare computazionalmente una dinamica completa delle proteine in un ambiente realistico.
Solo di recente, nel 2012, è stato proposto da un gruppo di ricercatori tedesco un metodo del tutto innovativo per riuscire ad analizzare tale dinamica nella sua globalità ed in modo universale [5]: esso consiste nel calcolo di 34 variabili che rappresentano vari aspetti della dinamica e che sono paragonabili per ogni coppia di proteine, anche se molto differenti tra loro per struttura o dimensione. Il set delle 34 variabili prende il nome di *dinasoma* e la dinamica della proteina è così rappresentata da un vettore 34-dimensionale nello spazio dei dinasomi, in cui la differenza tra due vettori qualsiasi dà conseguentemente una misura ragionevole della differenza tra le corrispondenti dinamiche delle proteine.
Grazie a tale metodo è stato dimostrato che è possibile classificare le proteine in classi ben separate in base al proprio dinasoma, ossia in base alla propria dinamica. Inoltre, è stato provato che la struttura delle proteine è strettamente correlata con la loro dinamica e che proteine con funzioni simili presentano dinamiche simili.
Il dinasoma risulta quindi un metodo efficace per l'analisi della dinamica di una proteina. Ci concentreremo su di esso per studiare la dinamica di una particolare struttura proteica della membrana delle cellule, un emicanale, costituito da proteine dette connessine 26. Ci si prefigge di verificare, inoltre, se tale metodo è sufficientemente sensibile per determinare delle significative differenze con emicanali, costituiti da forme mutate della connessina 26 che possono causare malattie genetiche. I metodi tramite i quali ciò è stato fatto verranno discussi nel capitolo successivo. In questo capitolo ci concentreremo in una descrizione dettagliata, da un punto di vista più teorico, del *dinasoma* e delle 34 variabili di cui è costituito. Per chiarezza, le classificheremo in quattro gruppi, che analizzano rispettivamente lo spettro degli autovalori della matrice di covarianza, le componenti principali della traiettoria, la rugosità del profilo di energia libera e le

fluttuazioni atomiche.

i) Variabili 1-7: caratterizzazione dello spettro degli autovalori della proteina

Indichiamo con $\vec{x}$ la 3N-upla costituita dalle 3 coordinate cartesiane di ciascuno degli N atomi di Carbonio-α.
Date 3N variabili x_i , si costruisce quindi la matrice di covarianza C, di coefficienti:

$$C_{ij}=\langle (x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle$$

e la si diagonalizza tramite la matrice ortonormale R, le cui colonne sono gli autovettori di C, detti anche *modi principali*:

$$R^{-1}C\,R=diag(\lambda_{1,}\ldots,\lambda_{3N})$$

Gli autovalori λ_i , tutti positivi, rappresentano la fluttuazione del modo principale corrispondente.
Essendo una proteina costituita da un numero elevato di atomi, la cui posizione è conosciuta per un grande numero di istanti temporali, il numero di dati da analizzare risulta elevato. Allo scopo di ridurre tale numero, si utilizza la *Principal Component Analysis (PCA)*, una tecnica matematica per l'analisi di set di dati di grandi dimensioni. Essa consente di definire un nuovo set di variabili con la proprietà che la covarianza tra due qualsiasi di esse è nulla e dunque non risultano avere correlazione.
A tale scopo, è sufficiente disporre gli autovalori della matrice di covarianza in ordine decrescente e ricavare le proiezioni (dette *componenti principali* del moto) degli scarti $x-\langle x\rangle$ nello spazio degli autovettori $p=R^{-1}(x-\langle x\rangle)$: le prime coordinate di questo nuovo spazio sono quelle ad avere le varianze, ovvero le fluttuazioni, più significative e quindi le uniche rilevanti per l'analisi successiva. Si considerano i primi cinque autovalori della matrice di covarianza come *variabili 1-5* del dinasoma.
Poiché il sistema obbedisce alle leggi di Newton, per ottenere una *PCA* significativa dal punto di vista fisico ciascuna coordinata dovrà essere però pesata con la radice della massa dell'atomo corrispondente.
Il moto complessivo consiste però nel movimento ininterrotto e irregolare compiuto dagli atomi in sospensione in un fluido: la dinamica delle proteine è quindi interpretabile come un caso di ensemble canonico di particelle soggette a moto di diffusione random, o moto browniano, sotto un potenziale quasi armonico [6].
Come descritto di seguito, le *variabili 6-7* sono state scelte in modo da stimare quanto lo spettro degli autovalori è compatibile [7] con lo spettro atteso da un moto di diffusione random.

Nel caso generale in cui il tempo è continuo e non discreto, una diffusione random di un ensemble canonico di N particelle può essere descritta da un sistema di N equazioni differenziali stocastiche del tipo:

$$\frac{dq_i(t)}{dt}=r_i(t)$$

con $q_i(t)$ traiettoria della i-esima particella, $r_i(t)$ termine di rumore tale che $\langle r_i(t)\rangle=0$ e

$\langle r_i(t) r_j(t+\tau)\rangle = 2D\delta_{ij}\delta(\tau)$ ove δ_{ij} è una δ di Kronecker, $\delta(\tau)$ è una δ di Dirac e D è il coefficiente di diffusione, definito da:

$$D = \frac{1}{\tau}\int_{-\infty}^{+\infty} \frac{\Delta^2}{2}\Phi(\Delta)\, d\Delta$$

con $\Phi(\Delta)$ funzione pari che rappresenta la probabilità che una particella compia uno spostamento Δ nel tempo τ .
Le equazioni del moto sono del tipo:

$$m_i \frac{dr_i}{dt} = -f_i r_i + F_i(t) + K_i(q)$$

ove f_i è il coefficiente di frizione, $F_i(t)$ la forza totale esercitata sull'i-esima particella, dovuta a un campo di forze esterne e alle particelle microscopiche del fluido, e $K_i(q)$ è un termine dovuto a un potenziale armonico.

Le soluzioni di questo sistema $q_i(t)$ sono tali che $\langle q_i(t+\tau) - q_i(t)\rangle = 0$ e $\langle (q_i(t+\tau) - q_i(t))(q_j(t+\tau) - q_j(t))\rangle = 2D\tau\delta_{ij}$. Le medie fin qui espresse sono medie di ensemble.
Per sistemi che si muovono in un potenziale quasi armonico, ovvero in cui l'energia potenziale *V(q)* possiede un minimo "quadratico" *q** in una configurazione pesata sulla massa , l'Hessiana *V''(q)* risulta definita positiva: tutti i suoi autovalori sono strettamente positivi ed è possibile dimostrare che ogni soluzione *q(t)-q** è combinazione lineare di 3N soluzioni periodiche, dette *modi normali di oscillazione*, proporzionali agli autovettori dell'Hessiana. I modi principali ottenuti con la *PCA* sono molto simili ai modi normali.
E' possibile provare che gli autovalori della matrice di covarianza sono inversamente proporzionali ai propri indici [Appendice A], secondo la relazione:

$$\lambda_k' = \frac{2NDT}{\pi^2 k^2}$$

ove T è la temperatura del sistema.

Nel caso di diffusione random, si ha quindi una dipendenza dell'autovalore dall'inverso del quadrato del proprio indice. A partire da questo risultato, è possibile dimostrare [8] che tale relazione continua ad essere valida in buona approssimazione anche nel caso in cui il tempo non è continuo, ma discreto.
Per fittare la legge $\lambda_k = a\, k^b$ si può considerare il logaritmo di entrambi i membri. Effettuando una regressione lineare, si sceglie il coefficiente b come sesto parametro del dinasoma.
Come settimo parametro si considera il coefficiente di determinazione R^2 del fit, parametro che quantifica la bontà del fit stesso. Nel caso di una regressione lineare esso è definito da:

$$R^2 = \frac{\sum_0^N (\log\hat{\lambda}_i - \langle\log\lambda\rangle)^2}{\sum_0^N (\log\lambda_i - \langle\log\lambda\rangle)^2}$$

ove $\log\hat{\lambda}_i$ è il valore atteso dal fit del logaritmo dell'i-esimo autovalore.

ii) Variabili 8-20: analisi delle componenti principali della traiettoria

L'analisi delle componenti principali, nelle quali ci si è proiettati utilizzando la PCA, dà quindi una buona stima delle caratteristiche del moto strutturale della proteina.

Le *variabili 8-12* sono dei coefficienti, detti *cosine contents*, che stimano quanto è verificata l'attesa teorica che le prime componenti principali di una diffusione random sono coseni [manuale gromacs]: come mostrato in Appendice A, è possibile infatti sviluppare lo scarto $x_i(t)-\langle x_i\rangle$ in una base di funzioni $f_k(t)$ proporzionali alle componenti principali del moto che risultano avere un andamento cosinusoidale:

$$x_i(t)-\langle x_i\rangle \approx \sum_0^\infty c_i^k f_k'(t) = \sqrt{\frac{2}{T}} \sum_{k=1}^\infty c_i^k \cos\left(\frac{\pi k t}{T}\right) \quad \text{con} \quad c_i^k = \sqrt{\frac{2}{T}} \int_0^T x_i(t) \cos\left(\frac{\pi k t}{T}\right) dt \quad .$$

Avendo definito un autovettore $\vec{e}_k$ della matrice di covarianza come il vettore di norma unitaria tale che, proiettandosi su di esso, la fluttuazione quadratica media corrispondente è massima, ossia quello che massimizza l'autovalore:

$$\lambda_k \equiv max_{\|e\|=1} \frac{1}{T} \int_0^T \left(\vec{e}_k \cdot (\vec{x}(t)-\langle \vec{x}\rangle)\right)^2 dt$$

da tale risultato si può dedurre che, se gli atomi della proteina seguono l'andamento di una diffusione random, gli autovalori corrispondenti sono pari a:

$$\lambda_k \approx \frac{1}{T}\frac{2}{T}\left(\sum_{t=1}^T p_k(t) \cos\left(\frac{\pi k t}{T}\right)\right)^2$$

ove $p_k(t)$ è ancora la proiezione sul k-esimo autovettore della matrice di covarianza dello scarto $\vec{x}(t)-\langle \vec{x}\rangle$ al tempo t:

$$p_k(t) = \vec{e}_k \cdot (\vec{x}(t)-\langle \vec{x}\rangle)$$

Si sceglie di stimare il rapporto tra tale valore e quello ottenuto direttamente dalle proiezioni sui modi principali per le sole prime cinque coordinate essenziali, con le varianze più significative. Tali rapporti prendono il nome di *cosine contents* e vengono considerati come parametri del dinasoma:

$$\cos_i = \frac{2}{T} \frac{\left(\sum_{t=1}^T \cos\left(\frac{i}{T}\pi t\right) p_i(t)\right)^2}{\sum_{t=1}^T p_i^2(t)} \quad .$$

Le *variabili 13-15* analizzano la distribuzione casuale dei valori assunti dalle prime tre coordinate principali nei vari istanti temporali. Effettuando un binning dei valori che ciascuna di esse assume in ciascuno degli M frames, si può calcolare un fit gaussiano dei risultati ottenuti. Si considerano come parametri del dinasoma i *coefficienti di determinazione* di ciascun fit. Detto b un parametro caratteristico della distribuzione di probabilità delle variabili considerate, si definisce *coefficiente di determinazione* la quantità [9]:

$$R^2 = 1 - \left(\frac{L(0)}{L(b)} \right)^{\frac{2}{n}}$$

ove *L(0)* è la likelihood della distribuzione di probabilità delle variabili valutata *b=0* , *L(b)* quella stimata per il valore di b atteso dal fit. Nel caso del fit gaussiano, essendo la likelihood della j-esima coordinata principale p^j pari a [10]:

$$L(\mu^j) = \prod_{i=1}^{M} f_i(p^j(t_i); \mu^j) = \frac{1}{\sigma^j \sqrt{2\pi}} e^{\frac{-\sum_{i=1}^{M} (p^j(t_i) - \mu^j)^2}{2\sigma^{j2}}}$$

esso diventa:

$$R^2 = 1 - \left(\frac{L(0)}{L(\mu^j)} \right)^{\frac{2}{n}} = 1 - e^{\frac{-\mu^j}{n\sigma^{j2}} (\mu^j n - 2\sum p^j(t_i))}$$

ove μ^j e σ^j sono rispettivamente la media e lo scarto della gaussiana trovata tramite l'interpolazione.
Le *variabili 16-20* sono le costanti di frizione del moto proiettato nelle prime cinque componenti principali. Esse si ricavano da un fit della funzione di autocorrelazione ACF di tali componenti, stimabile come:

$$ACF_i(\Delta t) = \frac{\langle p_i(t) \cdot p_i(t + \Delta t) \rangle}{\lambda_i} = \frac{\sum_{1}^{T - \Delta t} (p_i(t) \cdot p_i(t + \Delta t))}{(T - \Delta t)} \frac{1}{\lambda_i}$$

per ogni possibile intervallo Δt tra due frames.
Secondo la teoria proposta da Ornstein e Uhlenbeck [11], nel processo di diffusione di una particella sotto un potenziale armonico, la funzione di autocorrelazione assume una particolare forma analitica: considerata la generica soluzione dell'equazione del moto browniano $x(t + \Delta t)$ ad un istante successivo a quello iniziale *t*, la funzione di autocorrelazione è ottenuta prendendone una media di ensemble. Infatti, supposte soddisfatte le condizioni di validità del teorema ergodico, una media temporale del prodotto $(x(t + \Delta t) - \langle x \rangle)(x(t) - \langle x \rangle)$ è equivalente ad una media su un sub-ensemble di particelle aventi all'istante t uno scarto pari a un $x_0 = (x(t) - \langle x \rangle)$ e velocità arbitraria.
In questo modo la funzione di autocorrelazione risulta:

$$ACF(\Delta t) = e^{\frac{-\beta \Delta t}{2}} \left(\cos(\omega \Delta t) + \frac{\beta}{2\omega} \sin(\omega \Delta t) \right)$$

ove $\beta = \frac{f}{m}$ con *f* coefficiente di frizione e *m* massa della particella, ω è la frequenza del potenziale e Δt l'intervallo temporale tra due frames.

Per ciascuna delle prime 5 coordinate principali, considerati tutti i possibili intervalli temporali Δt e i relativi ACF, si effettua un fit tra di essi e si assumono le costanti β così ricavate come parametri del dinasoma.

iii) Variabili 21-23: rugosità del profilo di energia libera

Le *variabili 21-23* sono parametri che caratterizzano il profilo di energia libera di Helmholtz della dinamica delle proteine e sono strettamente legati alla definizione di un coefficiente di *rugosità*.

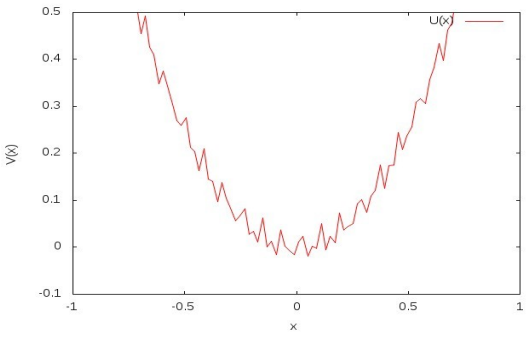

Figura 2.1. *Esempio di profilo di energia libera rugoso.*

Come mostrato da diverse ricerche [12][13], il profilo di energia libera principale *V(q)* di una proteina risulta *rugoso* (Figura 2.1), ossia costituito da una serie di barriere di potenziale di ampiezza molto più piccola, dovute ad un termine correttivo del potenziale, proporzionale alla temperatura del sistema e alla costante di Boltzmann: ciò comporta un moto più "smorzato" rispetto a quello che ci si aspetterebbe con un profilo liscio. Tale rugosità risulta di notevole interesse soprattutto perché a basse temperature è la causa di una drastica riduzione della diffusione.

Per riprodurre tale rugosità, si può supporre che l'altezza ΔF delle barriere di cui è costituito il potenziale armonico aumenti logaritmicamente con la mutua distanza Δx tra due barriere consecutive (Figura 2.2).

Stabilita una generica unità di lunghezza L per i Δx e un corrispondente coefficiente di diffusione D_0 del sistema, si definisce *coefficiente di rugosità del profilo di energia libera*, il parametro γ tale che:

$$\Delta F = \gamma k_B T \ln\left(\frac{\Delta x}{L}\right)$$

ove k_B è la costante di Boltzmann e T la temperatura fissata.
Esso non misura l'altezza delle barriere, ma piuttosto quanto velocemente esse crescano con l'aumentare della distanza tra esse.

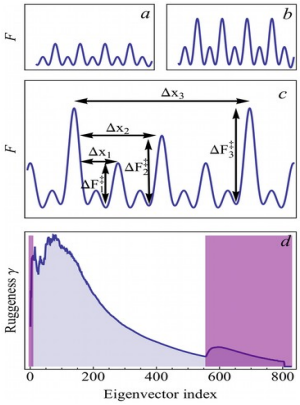

Figura 2.2. a) – c) *esempi schematici dei profili di energia libera;* ***d)*** *tipico profilo della rugosità in funzione dell'ordine dell'autovalore.*

Una forma analitica del quadrato dello spostamento medio delle particelle σ, ricavata da Einstein [6], è $\sigma^2 = 2tD$.

Assumendo $\sigma \approx \Delta x$, utilizzando l'equazione di Arrhenius $D_{eff} = D_0 e^{\frac{-\Delta F}{kT}}$ e sostituendo si trova:

$$\sigma^2 = \left(2\, t\, D_0\, L\right)^{\frac{2}{(2+\gamma)}}$$

Assumendo il quadrato dello spostamento medio pari alla varianza delle particelle, ossia gli autovalori della matrice di covarianza, si trova la dipendenza dell'autovalore dalla potenza del

tempo $t^{\frac{2}{(2+\gamma)}}$, dipendenza che si può trattare come lineare prendendone il logaritmo. Da tale relazione si può ricavare il coefficiente di rugosità per ogni autovalore.
La media dei coefficienti trovati per ciascun autovalore è la ventunesima variabile del dinasoma. I coefficienti di asimmetria e di curtosi sono rispettivamente la ventiduesima e la ventitreesima variabile. Essi sono definiti rispettivamente come:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \quad \text{e} \quad \gamma_2 = \frac{\mu_4}{\mu_2^2}$$

ove μ_k è il momento di ordine k rispetto alla media, definito come $\mu_k = \langle (\gamma - \langle \gamma \rangle)^k \rangle$.

iv) Variabili 24-34: fluttuazioni atomiche

Le ulitme variabili del dinasoma tengono conto della dinamica della proteina e delle sue strutture secondarie.

Si definisce *root mean square deviation (RMSD)* la quantità:

$$RMSD(t) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} \|x_i(t) - x_i(0)\|^2}$$

ove $M = \sum m_i$, N è il numero *totale* di atomi e $x_i(t)$ è la posizione dell'i-esimo atomo al tempo t. Essa stabilisce la fluttuazione media degli atomi rispetto a una configurazione di riferimento (al tempo t=0). La media su tutti i tempi dei valori trovati μ^{RMSD} e la deviazione standard relativa alla media $c^{RMSD} = \frac{\sigma^{RMSD}}{\mu^{RMSD}}$ sono le *variabili 24-25*.

La *variabile 26* è la media su tutti gli atomi della *root mean square fluctuation (RMSF)*, ovvero le fluttuazioni rispetto alla struttura media:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{i=1}^{T} \|x_i(t) - \bar{x}_i\|^2}$$

La *variabile 27* è la deviazione standard relativa alla media su tutti i tempi del *raggio giratore*, definito come:

$$R_g = \sqrt{\frac{\sum_{1}^{N} \|\vec{x_i^{CM}}\|^2 m_i}{\sum_{1}^{N} m_i}}$$

ove m_i è la massa dell'i-esimo atomo e $\vec{x_i^{CM}}$ la sua posizione rispetto al centro di massa. Esso quantifica i moti di "*respirazione*" della proteina.

Le *variabili 28-31* danno informazioni sulla struttura secondaria delle proteine. E' possibile ricavare tali informazioni seguendo l'algoritmo proposto da Kabsch e Sanders [15] per riuscire ad ottenere un metodo oggettivo e accurato per distinguere le strutture secondarie delle proteine, note le

coordinate di tutti gli atomi.
Per ciascun residuo si stima la percentuale di legami idrogeno appartenenti ad una delle strutture secondarie interessate: struttura secondaria complessiva, α eliche, foglietti β e *turns*. Si considerano come variabili del dinasoma le deviazioni standard relative alla media su tutti i residui di tali percentuali.

Tramite il medesimo algoritmo si può stimare la *solvent accessible surface (SAS)*, che fisicamente rappresenta il numero di molecole d'acqua in diretto contatto con la proteina o con una precisa parte di essa. Le *variabili 32-33* sono il valor medio e la deviazione standard rispetto alla media dei valori ottenuti per ogni frame.

Infine, come ultima *variabile 34* si stima l'entropia legata alla distribuzione delle RMSF medie di ciascun residuo.
Calcolata e mediata sul tempo la *root mean square fluctuation* di ciascun residuo $\langle x_i \rangle_{RMSF}$, se ne stima la fluttuazione relativa come $x_i^{rel} = \frac{\langle x_i \rangle_{RMSF}}{\sum_i^{n_{res}} \langle x_i \rangle_{RMSF}}$. Si costruisce in tal modo un set di dati interpretabile come una distribuzione discreta finita di probabilità delle n alternative che il residuo i-esimo abbia una determinata fluttuazione: si può associare ad essa una funzione $S_n(x_1^{rel},\ldots,x_n^{rel})$ che ne descrive l'incertezza, come dimostrato dal teorema di Shannon [15] [Appendice B]. Essa avrà la forma:

$$S_n(x_1^{rel}\ldots x_n^{rel}) = -\sum_{k=1}^{n} x_k^{rel} \ln x_k^{rel}$$

e si sceglie come variabile del dinasoma.

	Variabile	Descrizione	Simbolo
Caratterizzazione dello spettro degli autovalori	1	*Primo autovalore matrice di covarianza dei Carbonio α*	λ_1
	2	*Secondo autovalore matrice di covarianza dei Carbonio α*	λ_2
	3	*Terzo autovalore matrice di covarianza dei Carbonio α*	λ_3
	4	*Quarto autovalore matrice di covarianza dei Carbonio α*	λ_4
	5	*Quinto autovalore matrice di covarianza dei Carbonio α*	λ_5
	6	*Esponente della legge di potenza degli autovalori*	b_λ
	7	*Coeff. di determinazione della regressione lineare*	R^2_{lin}
Analisi delle componenti principali della traiettoria	8	*Cosine content della prima componente principale*	cos_1
	9	*Cosine content della seconda componente principale*	cos_2
	10	*Cosine content della terza componente principale*	cos_3
	11	*Cosine content della quarta componente principale*	cos_4
	12	*Cosine content della quinta componente principale*	cos_5
	13	*Coeff. di determinazione del fit gaussiano della prima componente principale*	$R^2_{gauss}1$
	14	*Coeff. di determinazione del fit gaussiano della prima componente principale*	$R^2_{gauss}2$
	15	*Coeff. di determinazione del fit gaussiano della prima componente principale*	$R^2_{gauss}3$
	16	*Coeff. di frizione della prima componente principale*	f_1
	17	*Coeff. di frizione della seconda componente principale*	f_2
	18	*Coeff. di frizione della terza componente principale*	f_3
	19	*Coeff. di frizione della quarta componente principale*	f_4
	20	*Coeff. di frizione della quinta componente principale*	f_5
Rugosità del profilo di energia libera	21	*Coefficiente di rugosità medio*	μ^γ
	22	*Asimmetria del coefficiente di rugosità*	as_γ
	23	*Cùrtosi del coefficiente di rugosità*	$curt_\gamma$
Fluttuazioni atomiche	24	*RMSD medio*	μ^{RMSD}
	25	*Deviazione standard relativa RMSD*	c^{RMSD}
	26	*RMSF*	μ^{RMSF}
	27	*Deviazione standard relativa del raggio giratore*	c^{gyr}
	28	*Deviazione standard relativa della percentuale di struttura secondaria complessiva*	c^{comp}
	29	*Deviazione standard relativa della percentuale di α eliche*	c^{α}
	30	*Deviazione standard relativa della percentuale di foglietti β*	c^{β}
	31	*Deviazione standard relativa della percentuale di turns*	c^{γ}
	32	*Solvent accessible surface media*	μ^{SAS}
	33	*Deviazione standard relativa della SAS*	c^{SAS}
	34	*Entropia di Shannon*	S

Tabella 1. *Tabella riassuntiva delle 34 variabili del dinasoma.*

Terzo capitolo

Calcolo del dinasoma per un connessone

La connessione diretta tra i citoplasmi di due cellule vicine avviene, nella maggior parte degli esseri viventi, tramite delle particolari strutture dette *canali di giunzione gap*[16]. Questi canali sono estremamente importanti e permettono il passaggio di molecole come ioni, metaboliti, nucleotidi e piccoli peptidi. Ogni canale di giunzione gap è costituito da due unità esameriche chiamate emicanali o *connessoni*, ciascuna delle quali si origina in una delle due cellule, si integra nella membrana plasmatica e si collega all'altra nello spazio extracellulare. Ciascun connessone è a sua volta costituito da sei connessine, una famiglia di proteine che include circa 20 isoforme, aventi circa il 40% della propria sequenza identico, in particolare nelle zone trans-membrana ed extracellulari.
Nonostante i canali formati da ciascuna isoforma abbiano caratteristiche differenti, come conduttanza, permeabilità, sensibilità di potenziale, tutte le isoforme presentano importanti proprietà strutturali in comune. La nomenclatura più utilizzata per distinguere tali isoforme è basata sulla loro massa molecolare, espressa in kilodalton: ad esempio, la *hCx26* è una connessina umana di massa circa 26 kDa. Per identificare, invece, una mutazione presente nella proteina, si utilizza un suffisso che indica l'amminoacido originale, la posizione e il residuo mutato. Ad esempio, la *hCx30T5M* è la connessina umana di 30 kDa in cui la treonina nella quinta posizione è sostituita da una metionina.
Ciascuna connessina è costituita da 4 eliche transmembrana (TM1-TM4), collegate da due anelli extracellulari (EC1-EC2) aventi una struttura a foglietto β, ed uno citoplasmatico (CL).
Ogni connessone ha una lunghezza compresa tra i 7 e i 7,5 nm ed è in grado di aprire e chiudere il poro che le connessine formano al loro interno tramite diversi meccanismi di regolazione dovuti alla presenza di un potenziale transgiunzionale, ad un potenziale di membrana o fattori chimici come la fosforilazione, il pH o ioni Ca^{+2} .

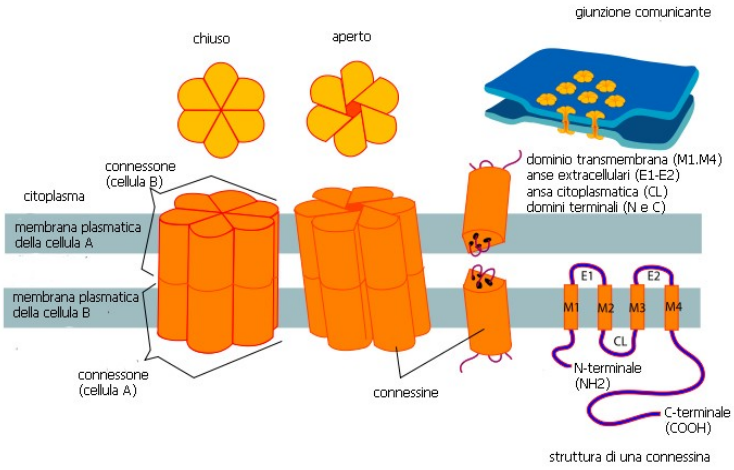

Figura 3.1. *Struttura schematica della gap junction nello stato di chiusura e di apertura; struttura schematica di una connessina.*

Le *gap junctions* hanno ruoli cruciali in molti processi biologici, come lo sviluppo, la differenziazione, la sincronizzazione cellulare, l'attività neuronale e risposte immunitarie. Mutazioni nelle connessine nell'uomo possono quindi causare diverse malattie, come malattie neurodegenerative, dermatologiche, sordità sindromiche e non e anomalie nello sviluppo.

In particolare, il gene GJB2 codificante la connessina umana 26, è stato il primo gene ad essere collegato ad una forma autosomica recessiva ed altre rare forme dominanti di sordità e sono state individuate più di 90 mutazioni recessive differenti di tale gene, che causano la presenza di connessina 26 mutata in cellule non sensoriali dell'intero orecchio.
È possibile allora applicare il modello del dinasoma alla connessina 26 wild type, non mutata, e ad alcune sue forme mutate per verificare se tali mutazioni strutturali comportano delle sensibili differenze nella dinamica delle proteine.

Nel Protein Data Bank sono disponibili le strutture cristallografiche ottenute per diffrazione di raggi x di un canale intercellulare formato da connessine 26 umana con una risoluzione di 3,5 A. La connessina 26 risulta avere 226 residui, per un totale di 22380 atomi.

La simulazione della dinamica della connessina 26 è stata ottenuta mediante metodi di dinamica molecolare classica [17], supponendola immersa in un ambiente realistico, comprendente fosfolipidi di membrana, molecole del solvente quali molecole d'acqua e un sufficiente numero di ioni per simulare le forze ioniche normali. La dinamica viene seguita per 100 ns e divisa in 834 frames temporali, per ciascuno dei quali si può ricavare la posizione di ogni atomo della connessina.

Le informazioni relative alla struttura e alle traiettoria della proteina vengono analizzate tramite programmi di visualizzazione di dinamica molecolare come VMD, l'utilizzo di *tools* del pacchetto di dinamica molecolare GROMACS [7] e del programma *cpptraj* di *AmberTools14*. Per il calcolo delle 34 variabili del dinasoma, di seguito descritto nel dettaglio, si rende necessario anche l'utilizzo di programmi appositamente scritti in linguaggio computazionale come C++.

- **Variabili 1-7: caratterizzazione dello spettro degli autovalori della proteina.** La diagonalizzazione della matrice di covarianza delle tre coordinate di ciascuno dei 1356 atomi di Carbonio-α della connessina 26 si calcola tramite il programma *g_covar* di *GROMACS*. Tale matrice avrà quindi dimensione 4068x4068.
  Alcuni degli autovalori ottenuti in tal modo, però, a causa delle approssimazioni dovute al calcolo numerico, possono risultare negativi e in valore assoluto piccoli. In tal caso, questi si escluderanno dall'analisi successiva. Normalizzando i restanti autovalori e ponendoli in ordine decrescente si possono ricavare i primi 5 parametri del dinasoma.
  Considerando, invece, il logaritmo dell'intero set di dati, attraverso il programma *xmgrace* si seleziona una regione dei dati, che corrisponde approssimativamente a quella centrale, in cui l'andamento è lineare e non è influenzato dai fattori correttivi dovuti alla discretizzazione del tempo. Di tali dati si effettua una regressione lineare semplice e se ne considera il coefficiente angolare e il fattore di determinazione.

- **Variabili 8-20: analisi delle componenti principali della traiettoria.** Per ottenere gli autovettori si utilizza il programma *g_covar* all'interno del pacchetto di *GROMACS* e, utilizzando il file così ottenuto, è possibile ricavare le componenti principali tramite il programma *g_anaeig* e quindi i *cosine contents* tramite *g_analyze*.

I fit gaussiani, calcolabili tramite *ROOT*, si possono ottenere effettuando, con un apposito programma, un binning dei valori assunti dalla componente negli 834 frames: si sceglie di utilizzare canali di ampiezza 0,25. E' conveniente, tra i vari risultati ottenibili dai fit al variare dei parametri iniziali, scegliere quelli per i quali si ha un migliore valore del chi quadro ridotto. Successivamente, è possibile stimare computazionalmente il coefficiente di determinazione.
I valori della funzione di autocorrelazione si stimano computazionalmente dalla formula teorica per tutti i possibili valori di distanza tra due frames. Tali valori risultano uguali a quelli calcolati tramite il programma *g_analyze* di GROMACS. Si può quindi effettuare un fit non lineare della forma analitica attesa dal modello di Uhlenbeck-Ornstein e ricavarne le costanti di frizione.

- **Variabili 21-23: rugosità del profilo di energia libera.** I coefficienti di rugosità si ottengono dal coefficiente angolare della relazione lineare tra il logaritmo degli autovalori in funzione del logaritmo del tempo:

$$\log \lambda^2 = \frac{2}{2+\gamma} \log T + \log \left(D_0 L^{\gamma} \right)^{\frac{2}{2+\gamma}}$$

E' quindi necessario calcolare la matrice di covarianza, e dunque gli autovalori, per traiettorie di diverse lunghezze temporali. Per ridurre le fluttuazioni statistiche, si sceglie di dividere la lunghezza temporale della traiettoria completa di 100 ns in 5 finestre di 20 ns l'una. Ciascuna di esse si divide in 10 intervalli temporali logaritmicamente equispaziati e si calcolano le 10 matrici di covarianza corrispondenti, per una lunghezza temporale che va dall'istante iniziale della finestra all'intervallo scelto. Scelti i soli autovalori positivi e normalizzati, si effettua una media tra i valori corrispondenti a traiettorie di pari lunghezze nelle 5 diverse finestre. Si ottengono in questo modo 10 set di autovalori. Considerando i soli autovalori che risultano positivi per tutti i 10 set, per ciascun autovalore si effettua un fit lineare con le 10 lunghezze di traiettoria corrispondenti.
Dei vari valori di γ così ottenuti si calcola la media, il coefficiente di asimmetria e quello di curtosi attraverso un programma in C++ appositamente scritto.

- **Variabili 24-34: fluttuazioni atomiche.** I programmi *g_rms*, *g_gyrate* e *gmx sasa* di GROMACS restituiscono rispettivamente il valore della RMSD, del raggio giratore e della *solvent accessible surface* ad ogni frame: di tali valori si calcolano media e deviazione standard relativa.
Il programma *g_rmsf* calcola invece la RMSF per ogni atomo. Di questi valori va calcolata la media.
Per il calcolo dei parametri relativi alla struttura secondaria della proteina è necessario l'utilizzo del tool *cpptraj* del pacchetto *AmberTools14*. Tramite la funzione secstruct di *cpptraj* si ottengono, per ciascun residuo, le percentuali di legami idrogeno che contribuiscono a formare foglietti β paralleli e antiparalleli, α eliche, *turns*, eliche Pi ed eliche 3-10. La somma di tutte queste dà la percentuale, per residuo, di legami idrogeno che contribuiscono a formare la struttura secondaria totale. Di queste percentuali si calcola la deviazione standard relativa. Allo stesso modo, si calcola quella delle catena α , dei foglietti β , dati dalla somma per residuo delle percentuali degli anti e dei para, e dei

turns.
Tramite la funzione *atomicfluct* di *cpptraj* si ottengono le fluttuazioni medie di ciascun atomo: la loro media restituisce correttamente il valore della RMSF, ma essi sono utilizzabili per il calcolo dell'entropia tramite un programma scritto in C++.

Tramite tali metodi di calcolo è possibile ricavare le 34 variabili del dinasoma. Tali metodi sono del tutto generali e applicabili a diverse mutazioni della connessina 26.
In particolare, si è scelto di analizzare, oltre alla forma *wild type*, le mutazioni V84L, C169Y ed una mutazione che causa la mancanza di cisteina nella parte extracellulare della connessina e che indicheremo d'ora in avanti con la notazione *NoCys_EL*. È stato dimostrato che la mutazione V84L è patogena e causa sordità [18], mentre per la mutazione C169Y, ritenuta fino a poco tempo fa un polimorfismo, la questione è controversa. La privazione delle sei cisteine nella parte extracellulare della connessina è invece una mutazione non esistente in natura, ma ottenuta in laboratorio: è stato osservato sperimentalmente che eliminando la cisteina di una delle sei connessine di canale per volta, non si ha la formazione della gap junction. Le cisteine sono infatti dei particolari amminoacidi in grado di legarsi in maniera molto forte tra di loro tramite la formazione di un ponte disolfuro. Si è deciso quindi di simulare computazionalmente il comportamento di un canale nel quale tutte e sei le connessine avessero tale mutazione contemporaneamente per studiarne le caratteristiche. I dati così ottenuti sono stati analizzati nel capitolo successivo.

Quarto capitolo

Risultati

Abbiamo applicato i metodi di calcolo esposti nel precedente capitolo all'emicanale composto da connessine 26 e a quelli in cui sono presenti le forme mutate V84L, C169Y e NoCys_EL.
In questo modo si trovano i valori riportati in Tabella 2.

			V84L	C169Y	NoCys_EL	WT	Media
1	Primo autovalore (nm^2)	λ_1	0,29	0,38	0,39	0,33	0,35
2	Secondo autovalore (nm^2)	λ_2	0,13	0,11	0,12	0,14	0,13
3	Terzo autovalore (nm^2)	λ_3	0,089	0,061	0,083	0,064	0,074
4	Quarto autovalore (nm^2)	λ_4	0,063	0,054	0,037	0,054	0,052
5	Quinto autovalore (nm^2)	λ_5	0,050	0,030	0,026	0,048	0,039
6	b del fit di potenza	b_1	-1,6	-1,6	-1,5	-1,5	-1,5
7	R2 fit potenza	R^2_{lin}	1,00	0,99	0,99	0,99	0,99
8	Primo cosine	cos_1	0,96	0,94	0,96	0,94	0,95
9	Secondo cosine	cos_2	0,66	0,80	0,83	0,62	0,73
10	Terzo cosine	cos_3	5,08E-04	0,016	0,74	0,42	0,29
11	Quarto cosine	cos_4	2,82E-03	7,60E-03	0,66	0,40	0,27
12	Quinto cosine	cos_5	0,65	0,76	0,11	0,22	0,43
13	Primo R2 gauss	R^2_{gauss1}	1,00	-1,5	-1,8	1,00	-0,33
14	Secondo R2 gauss	R^2_{gauss2}	0,67	0,27	-0,36	-0,030	0,14
15	Terzo R2 gauss	R^2_{gauss3}	1,27E-04	0,095	0,29	0,68	0,26
16	Primo coeff frizione (ps^{-1})	f_1	8,63E-06	-1,11E-05	-1,58E-05	-1,33E-05	-7,89E-06
17	Secondo coeff frizione (ps^{-1})	f_2	-2,27E-05	-7,93E-06	-8,95E-06	-7,27E-06	-1,17E-05
18	Terzo coeff frizione (ps^{-1})	f_3	1,77E-05	1,75E-06	-2,12E-06	1,02E-05	6,88E-06
19	Quarto coeff frizione (ps^{-1})	f_4	9,85E-06	4,41E-06	-1,26E-05	5,53E-05	1,42E-05
20	Quinto coeff frizione (ps^{-1})	f_5	2,25E-05	-1,90E-07	2,53E-05	-1,09E-05	9,17E-06
21	Gamma medio	μ^{γ}	1,2	1,5	1,4	1,1	1,3
22	Skew gamma	as_{γ}	-1,2	-1,00	-1,2	-1,4	-1,2
23	Kurt gamma	$curt_{\gamma}$	3,5	2,9	3,7	4,1	3,6
24	RMSD medio (nm)	μ^{RMSD}	0,35	0,35	0,35	0,34	0,35
25	dev relativa RMSD	c^{RMSD}	0,17	0,18	0,19	0,13	0,17
26	RMSF medio (nm)	μ^{RMSF}	0,18	0,18	0,18	0,18	0,18
27	Dev relativa Gyrate	c^{gyr}	4,25E-03	5,45E-03	2,44E-03	4,55E-03	4,17E-03
28	Dev relativa strutt sec tot	c^{comp}	0,54	0,55	0,55	0,55	0,55
29	Dev relativa alfa	c^{α}	0,80	0,79	0,81	0,80	0,80
30	Dev relativa beta	c^{β}	4,4	4,9	4,4	4,6	4,6
31	Dev relativa turn	c^{t}	2,2	2,2	2,2	2,2	2,2
32	SAS media (nm^2)	μ^{SAS}	694,1	693,9	678,4	671,7	684,5
33	Dev relativa SAS	c^{SAS}	0,012	0,014	0,016	0,011	0,013
34	Entropia	s	10,0	9,8	9,8	9,8	9,9

Tabella 2. *Valori delle 34 variabili del dinasoma per l'emicanale composto dalla connessina 26 wild type e mutazioni V84L, C169Y, NoCys_EL. L'ultimacolonna riporta la media dei 4 valori assunti dalla corrispondente variabile.*

E' possibile fare una prima approssimativa analisi delle differenze esistenti tra le dinamiche, dividendo ciascuna variabile per il valore medio assunto nei quattro casi studiati e calcolando lo scarto delle forme mutate dalla forma *wild type*: si trova in tal modo che gli scarti più significativi

sono quelli relativi alle variabili del secondo gruppo, che analizza i modi principali della dinamica.

Una conferma di tale risultato si ottiene effettuando una PCA sulle 34 variabili del dinasoma. Per farlo si costruisce la matrice di covarianza 34x34 dei valori assunti dalle variabili, divisi per la media, dove la media temporale è stata sostituita con la media sui 4 casi studiati. Successivamente si diagonalizza e si trova che i primi due autovalori coprono da soli rispettivamente il 64,5% e il 26% delle fluttuazioni totali.

Per capire quali variabili corrispondono a queste fluttuazioni e in che quantità, si analizzano le componenti degli autovettori relativi a questi autovalori (Figura 4.1a). Tutte le variabili coinvolte in maggiore misura sono relative alle componenti principali, in particolare ai *cosine contents*, che stimano quanto il campione analizzato è buono e segue l'atteso moto browniano, e i coefficienti di frizione, che danno una stima dell'attrito relativo alla dinamica dei modi principali corrispondenti.
Proiettando il dinasoma dei quattro canali nello spazio bidimensionale di tali autovettori si rappresentano le loro dinamiche come dei vettori a due dimensioni, evidenziando che la dinamica del canale NoCys_EL risulta simile a quella del C169Y e quella del V84L simile alla forma *wild type* (Figura 4.1b). Si noti che entrambe le mutazioni NoCys_EL e C169Y sono caratterizzate da modificazioni nella loop extra-cellulare (EL), dovute rispettivamente all'assenza di tutte le cisteine della connessina e all'assenza di una sola di esse, la C169: ciò causa la non formazione di alcuni ponti disolfuro, fattore che potrebbe dunque essere importante nella determinazione della loro dinamica.

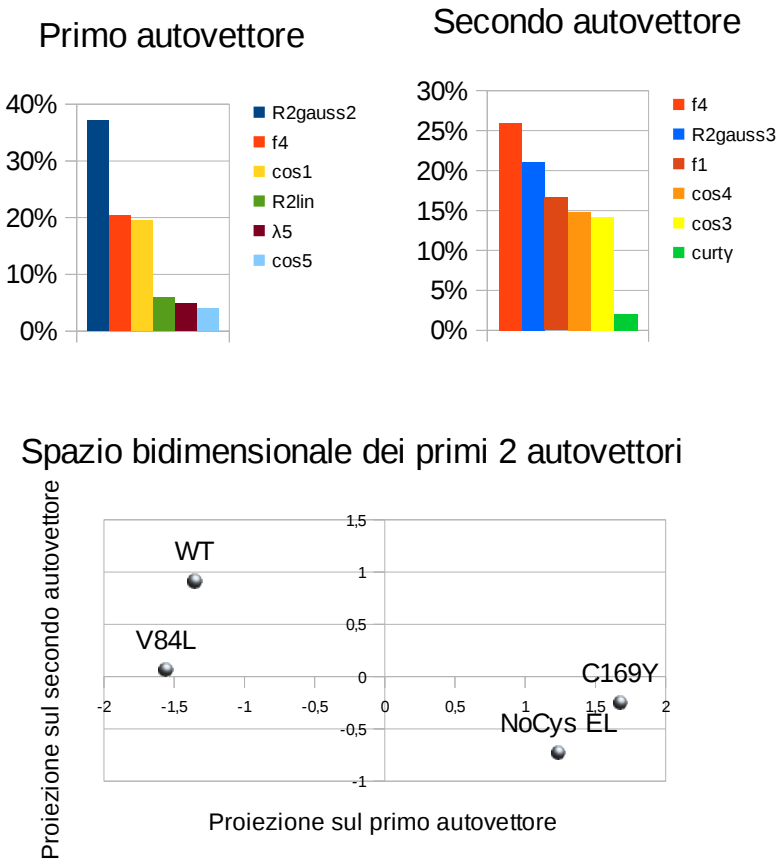

Figura 4.1. a) *Grafico delle percentuali delle componenti dei primi autovettori della matrice di covarianza costruita sulle 34 variabili del dinasoma.* ***b)*** *Proiezione dei dinasomi nello spazio bidimensionale dei primi 2 autovettori.*

Discutiamo tali risultati in maniera più dettagliata, analizzando ciascuno dei quattro gruppi di variabili separatamente, lasciando per ultima la descrizione più approfondita delle variabili relative all'analisi dei modi principali.

- **Variabili 1-7: caratterizzazione dello spettro degli autovalori della proteina**

In Figura 4.2 sono riportati i valori ottenuti per i 5 primi autovalori normalizzati, in Figura 4.3 invece sono riportati i grafici della legge di potenza prevista dal fit effettuato.

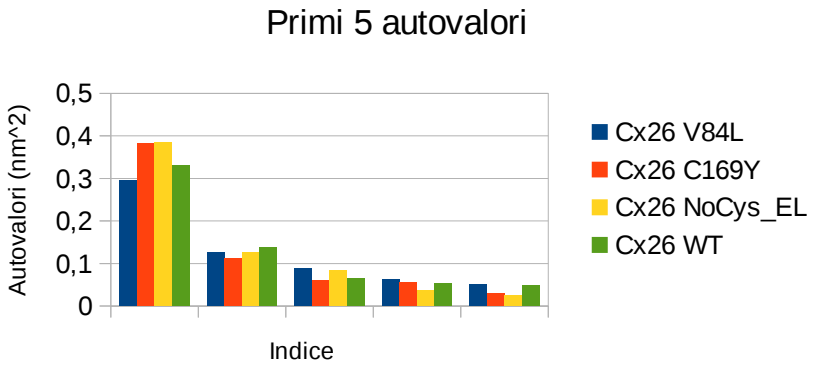

Figura 4.2. *Grafico dei primi 5 autovalori ottenuti tramite la PCA.*

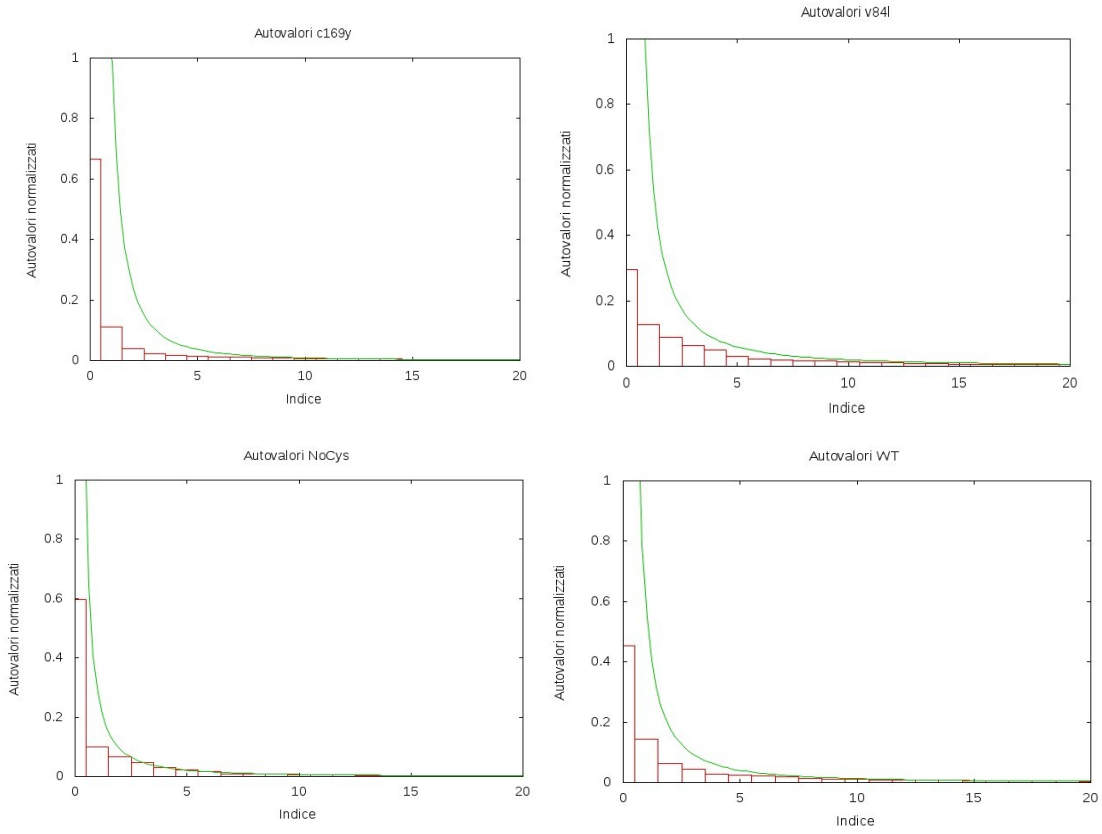

Figura 4.3. *Grafici dei primi autovalori di ciascun connessone e della curva interpolante prevista dal fit effettuato.*

La curva del fit di potenza non interpola bene i dati iniziali: infatti questi dati sono stati esclusi per il calcolo della regressione, in quanto in tale regione l'andamento poteva essere

influenzato da fattori correttivi dovuti alla discretizzazione del tempo. Il valore del coefficiente di determinazione (R^2) suggerisce che il fit di potenza è accettabile in tutti e 4 i casi analizzati. Inoltre, tutti i valori dell'esponente del fit di potenza risultano compatibili, nei limiti della statistica utilizzata, con il valore teorico (*b = -2*). Questo conferma che l'andamento seguito dagli autovalori è sempre lo stesso.

- **Variabili 21-23: rugosità del profilo di energia libera**

In Figura 4.4 sono riportati i grafici dei coefficienti di rugosità in funzione dell'indice degli autovalori.

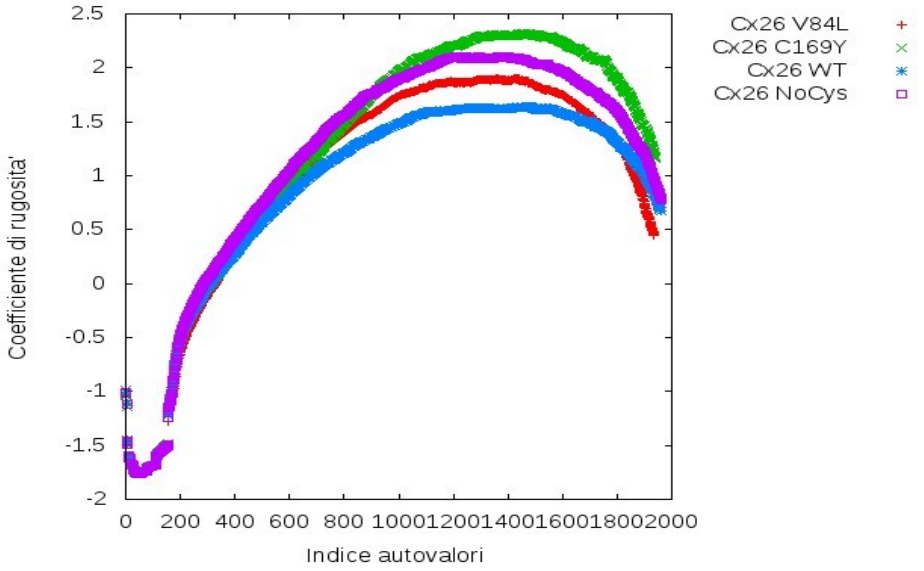

Figura 4.4. *Andamento dei coefficienti di rugosità in funzione dell'indice degli autovettori.*

Tutti i grafici presentano nella parte iniziale uno stesso minimo della rugosità che caratterizza la dinamica di tutti gli emicanali. Per questo motivo, volendo mettere in evidenza le differenze tra le varie dinamiche, tale regione è stata esclusa dal calcolo delle tre variabili del dinasoma relative a questa grandezza. La rugosità mantiene sempre lo stesso profilo, ma assume valori mediamente sempre più alti nell'ordine *wild type* - *V84L*- *NoCys_EL – C169Y* . Questo significa che le barriere di energia di cui si compone il profilo di energia libera crescono più velocemente. Ne risulta un moto più smorzato, a causa della presenza di buche di potenziale più frequenti e più profonde.

- **Variabili 24-34: fluttuazioni atomiche.**

Le variabili relative alle fluttuazioni atomiche analizzano sotto vari aspetti le fluttuazioni di ogni singolo atomo e della struttura secondaria e sono quelle che presentano le fluttuazioni minori (Figura 4.5).

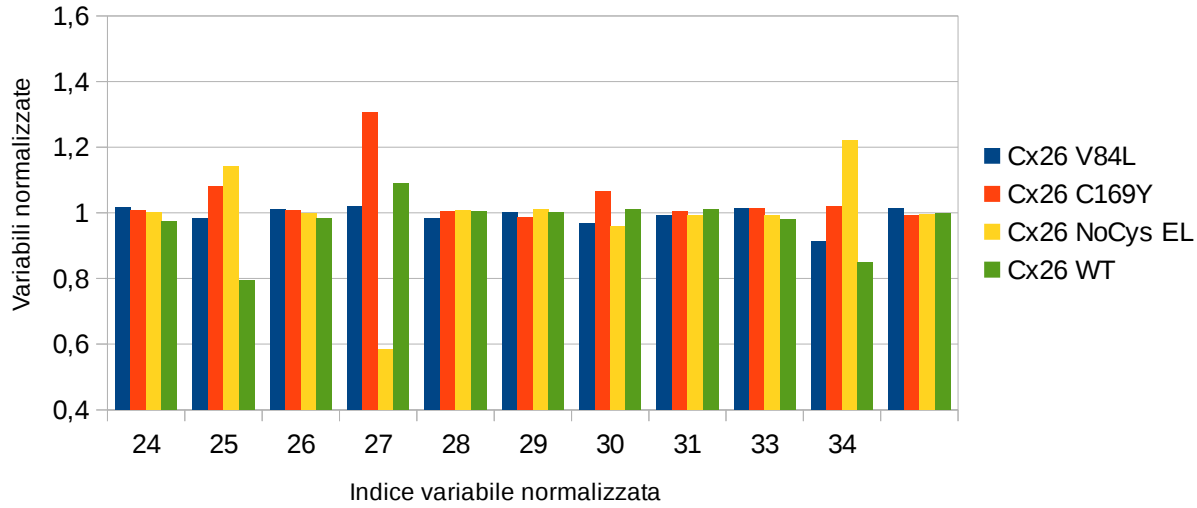

Figura 4.5. *Grafico delle variabili relative alle fluttuazioni atomiche divise per la loro media.*

Ciò non sorprende, poiché le mutazioni considerate coinvolgono pochi residui per volta e quindi solo pochi atomi variano nella sequenza amminoacidica dell'intero connessone. Non è quindi la dinamica delle singole particelle a mutare, ovvero quanto un atomo fluttui mediamente nel tempo o la quantità di legami idrogeno che forma con particolari strutture secondarie, ma piuttosto la dinamica della struttura nel complesso e del moto d'insieme dell'ensemble considerato.

- **Variabili 8-20: analisi delle componenti principali della traiettoria.**

  La prova del fatto che le maggiori variazioni della dinamica tra una mutazione e la forma wild type del connessone coinvolgano i moti strutturali delle proteine è data dall'analisi delle componenti principali, che sono combinazioni lineari dei moti di quegli specifici atomi che presentano le fluttuazioni maggiori. In Figura 4.6 sono riportati i valori ottenuti per i cosine contents e per i coefficienti di frizione di tali componenti principali.

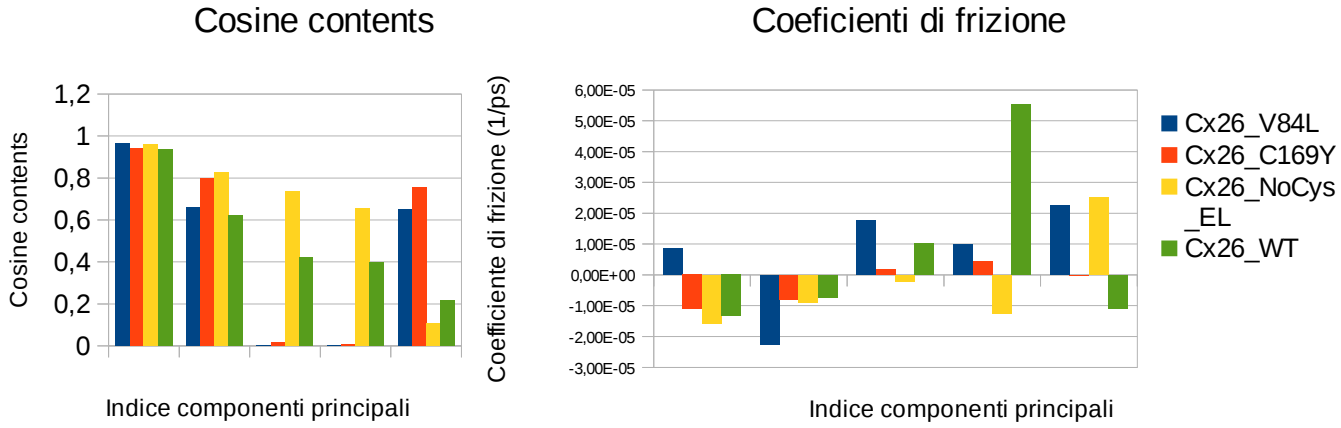

Figura 4.6. *Grafici dei cosine contents e dei coefficienti di frizione relativi alle prime 5 componenti principali.*

Come già osservato in precedenza, queste variabili presentano le fluttuazioni maggiori, sia rispetto alla media che rispetto al valore assunto dalla forma wild type.
L'andamento nel tempo delle componenti principali risulta diverso per ciascuna mutazione. Per studiare l'andamento delle componenti principali, si analizzano gli autovettori delle fluttuazioni per ciascun emicanale.
Infatti, l'i-esima componente principale è calcolata come

$$p_i(t) = \vec{e}_i \cdot (\vec{x}(t) - \langle \vec{x}(t) \rangle)$$

ove $\vec{x}(t)$ è la la 3N-upla costituita dalle 3 coordinate cartesiane di ciascuno degli N atomi di Carbonio-α della proteina ed $\vec{e}_i$ l'autovettore che andremo a studiare. Se le componenti dell'autovettore delle fluttuazioni sono diverse nei diversi connessoni significa che diverso sarà il peso che si darà alle fluttuazioni degli atomi corrispondenti a quelle componenti: più sono grandi, più il moto di quegli atomi contribuirà ai primi modi principali.
Nel caso in esame, nei primi 5 autovettori le componenti seguono lo stesso andamento nei 4 canali, con ampiezze che non risultano significativamente diverse (Figura 4.7). Ciò significa che ci si attende un moto degli stessi atomi, o degli stessi residui, con fluttuazioni rispetto alla posizione media che risultano simili.

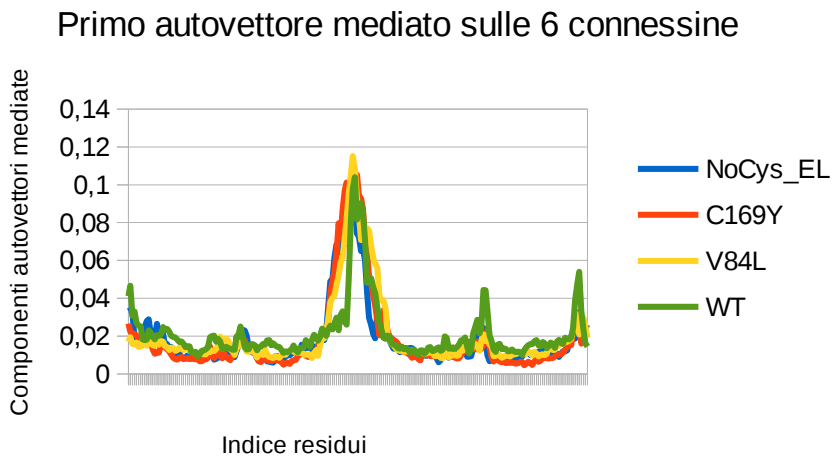

Figura 4.7. *Grafico delle componenti del primo autovettore mediate sulle sei diverse connessine dello stesso connessone: sono sovrapposti quelli relativi a connessoni diversi, evidenziati con colori differenti.*

Le differenze presenti in tali variabili sono quelle che comportano le maggiori fluttuazioni sia delle forme mutate rispetto alla wild type, sia di tutti e 4 gli emicanali rispetto ad una loro media. Come già visto precedentemente, ciò comporta una netta divisione in due gruppi nello spazio bidimensionale dei dinasomi, in cui le forme NoCys_EL e C169Y sono significativamente separate rispetto alle altre. Per confermare tale suddivisione e per accertarsi della sensibilità dello strumento utilizzato, si può prendere in considerazione il dinasoma di un emicanale completamente differente e confrontarlo con quelli già calcolati. Abbiamo analizzato l'emicanale costituito da connessina 32

R220stop, ovvero una connessina 32 con una mutazione che causa l'interruzione della proteina dopo il residuo 219. Grazie a tale mutazione, la catena amminoacidica della connessina risulta di lunghezza molto simile a quella della connessina 26. Tale canale è costituito da 1314 residui e 21336 atomi totali. La dinamica a disposizione è di solamente 50 ns, la metà rispetto alle traiettorie usate con la connessina 26. Ciò ha comportato qualche modifica nel calcolo, ad esempio, dei coefficienti di rugosità e potrebbe risultare un tempo non sufficiente per la corretta convergenza di tutte le variabili.
Le 34 variabili del dinasoma sono riportate in Tabella 3.

			Cx32 R220stop	Cx26 WT
1	Primo autovalore (nm^2)	λ_1	0,59	0,33
2	Secondo autovalore (nm^2)	λ_2	0,12	0,14
3	Terzo autovalore (nm^2)	λ_3	0,048	0,064
4	Quarto autovalore (nm^2)	λ_4	0,034	0,054
5	Quinto autovalore (nm^2)	λ_5	0,025	0,048
6	b del fit di potenza	b_λ	-2,0	-1,5
7	R2 fit potenza	R^2_{lin}	0,99	0,99
8	Primo cosine	cos_1	0,31	0,94
9	Secondo cosine	cos_2	0,014	0,62
10	Terzo cosine	cos_3	0,024	0,42
11	Quarto cosine	cos_4	1,76E-03	0,40
12	Quinto cosine	cos_5	0,031	0,22
13	Primo R2 gauss	R^2_{gauss1}	1,0	1,0
14	Secondo R2 gauss	R^2_{gauss2}	0,72	-0,030
15	Terzo R2 gauss	R^2_{gauss3}	1,00	0,68
16	Primo coeff frizione (ps^{-1})	f_1	9,89E-06	-1,33E-05
17	Secondo coeff frizione (ps^{-1})	f_2	2,09E-03	-7,27E-06
18	Terzo coeff frizione (ps^{-1})	f_3	1,82E-05	1,02E-05
19	Quarto coeff frizione (ps^{-1})	f_4	4,68E-03	5,53E-05
20	Quinto coeff frizione (ps^{-1})	f_5	4,40E-05	-1,09E-05
21	Gamma medio	μ^γ	3,2	1,1
22	Skew gamma	as_γ	-0,98	-1,4
23	Kurt gamma	$curt_\gamma$	2,7	4,1
24	RMSD medio (nm)	μ^{RMSD}	0,63	0,34
25	dev relativa RMSD	c^{RMSD}	0,58	0,13
26	RMSF medio (nm)	μ^{RMSF}	0,31	0,18
27	Dev relativa Gyrate	c^{gyr}	0,010	4,55E-03
28	Dev relativa strutt sec tot	c^{comp}	0,55	0,55
29	Dev relativa alfa	c^{α}	0,87	0,80
30	Dev relativa beta	c^{β}	4,3	4,6
31	Dev relativa turn	c^{γ}	2,1	2,2
32	SAS media (nm^2)	μ^{SAS}	660,8	671,7
33	Dev relativa SAS	c^{SAS}	0,015	0,011
34	Entropia	S	9,9	9,8

Tabella 3. *Valori delle 34 variabili del dinasoma per l'emicanale composto dalla connessina 32 R220stop e per quello composto dalla connessina 26 wild type.*

La maggior parte delle 34 variabili della Cx32R220stop ha uno scarto maggiore dalla Cx26 wild type di quelle delle 3 mutazioni considerate. Come era prevedibile, le fluttuazioni delle variabili relative alle fluttuazioni atomiche sono più significative rispetto agli altri casi.
Inoltre, il coefficiente di rugosità del profilo di energia libera della dinamica del canale con Cx32 R220stop è mediamente più alta (Figura 4.8).

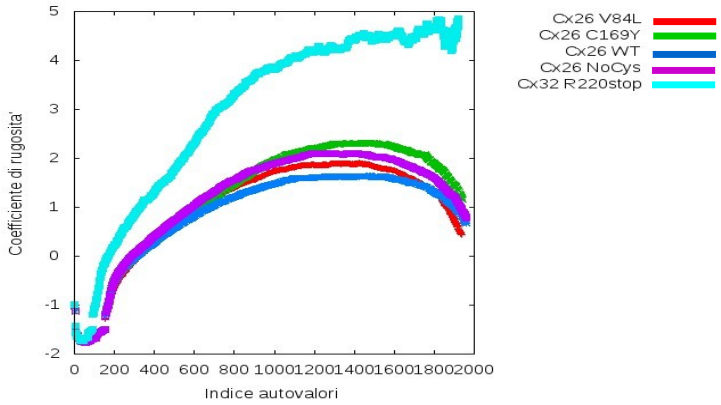

Figura 4.8. *Grafico del coefficiente di rugosità in funzione dell'indice degli autovalori.*

Effettuando nuovamente la PCA delle 34 variabili del dinasoma includendo, oltre ai quattro già analizzati, anche quello con Cx32 R220stop si ottiene che il primo autovalore copre da solo il 95,5% delle fluttuazioni totali, mentre il secondo ne copre solamente il 2,9%. Analizzando gli autovettori relativi a tali autovalori si trova una composizione differente rispetto alla precedente (Figura 4.9a).

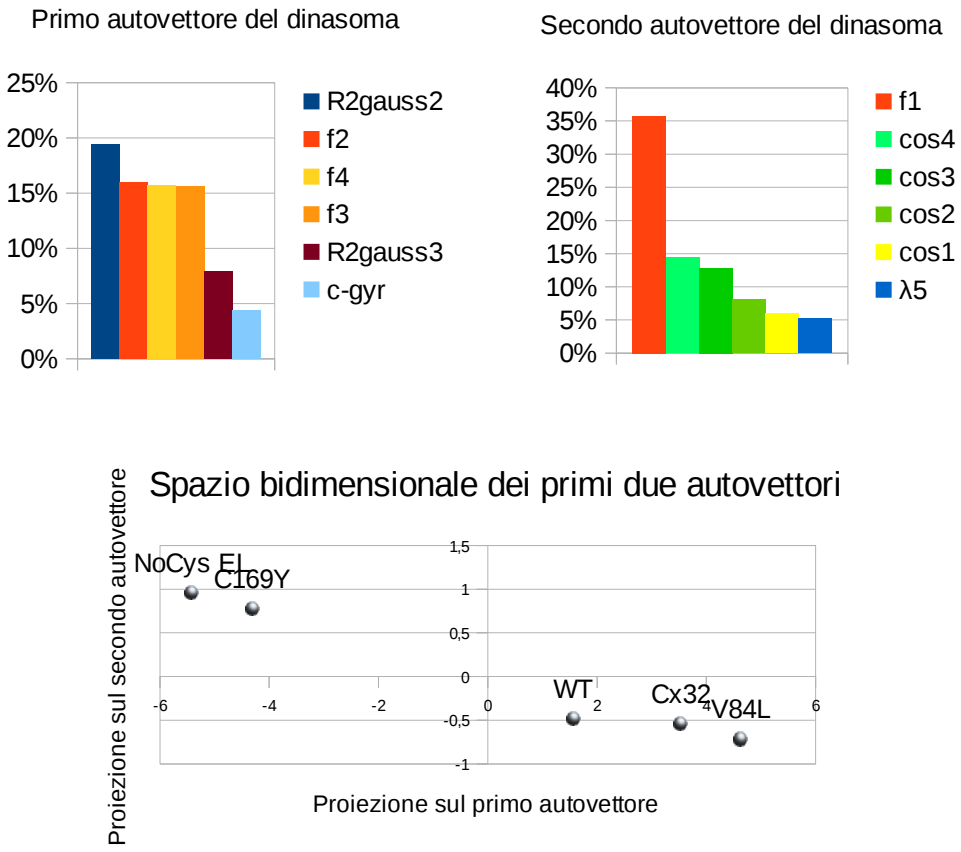

Figura 4.9. a) *Grafico delle percentuali delle componenti dei primi autovettori della matrice di covarianza costruita sulle 34 variabili del dinasoma.* ***b)*** *Proiezione dei dinasomi nello spazio bidimensionale dei primi 2 autovettori.*

Anche in questo caso, dunque, le variabili con le fluttuazioni maggiori sono quelle che descrivono le componenti principali delle traiettorie. Proiettandosi sullo spazio bidimensionale di questi due autovettori si nota che la dinamica della Cx32 R220stop è distinguibile dalle altre e risulta molto più simile a quella della wild type e della V84L piuttosto che alla dinamica dei canali con modificazioni sulla cisteina delle loops extracellulari (Figura 4.9b).

Conclusioni

Tramite la costruzione di un metodo computazionale per calcolare le variabili del dinasoma, è stato possibile evidenziare delle differenze tra la dinamica di un connessone composto da connesssina 26 nella forma *wild type* e quella di uno stesso connessone in cui la connessina 26 presenta amminoacidi mutati. Calcolare semplicemente una distanza euclidea nello spazio 34-dimensionale tra i dinasomi relativi ai vari connessoni avrebbe dato un'informazione troppo povera, in quanto non si è in possesso di una quantità sufficienti di altre distanze tra dinasomi con le quali poter fare un confronto.
Abbiamo scelto quindi di studiare le variabili dividendole in quattro gruppi che analizzano grandezze fisiche differenti che caratterizzano la dinamica delle proteine. Stimando lo scarto relativo di ciascuna variabile e utilizzando tecniche come la *PCA* si nota che le variabili per le quali si hanno le maggiori fluttuazioni sono quelle che caratterizzano le *componenti principali* della dinamica. Le *componenti principali* sono delle combinazioni lineari dei moti dei singoli atomi che sono costituite prevalentemente dal moto di quegli atomi che presentano le fluttuazioni maggiori. Gli atomi coinvolti in tali moti *collettivi* sono gli stessi nei 4 emicanali studiati, ma con fluttuazioni di poco differenti.
In nessun caso le grandezze che caratterizzano le fluttuazioni dei singoli atomi presentano delle significative differenze.
L'analisi dei dinasomi, inoltre, permette di distinguere nettamente le mutazioni che coinvolgono le loops extracellulari delle connessine. Infatti, nello spazio dei dinasomi la dinamica delle mutazioni C169Y e NoCys_EL si distinguono molto di più dalla forma wild type rispetto alla mutazione V84L, che coinvolge un residuo (V84) della connessina che si trova nella regione transmembrana.

Calcolando il dinasoma di un emicanale costituito da connessina 32 R220stop, di lunghezza paragonabile alla connessina 26, esso risulta più simile a quelli della Cx26V84L e della Cx26 wild type rispetto a quelli della Cx26NoCys_EL e della Cx26C169Y. La maggior parte delle variabili del Cx32R220stop però hanno una variazione maggiore da quelle della Cx26 wild type rispetto alla Cx26V84L. Inoltre, la rugosità del profilo dell'energia libera di Helmoltz ha un andamento differente in funzione degli autovalori della matrice di covarianza delle posizioni degli atomi ed è tendenzialmente maggiore. Il profilo di energia risulta più rugoso, quindi il potenziale armonico cui è soggetto ogni atomo del canale risulta definito da buche di potenziale più frequenti e più profonde che smorzano il moto. Resta da verificare, però, se questo risultato è dovuto effettivamente ad una diversa dinamica del canale o piuttosto al fatto che per tale canale abbiamo analizzato una traiettoria seguita per un tempo più breve (50 ns) rispetto agli altri canali.
Il fatto che nello spazio dei dinasomi la dinamica del Cx32R220stop è simile a quella delle Cx26 wild type e Cx26V84L potrebbe significare che il calcolo del dinasoma non è uno strumento sufficientemente sensibile per analizzare le differenze tra le dinamiche di canali completamente differenti tra loro e canali che variano solo per pochi residui ben localizzati. Viceversa, i risultati ottenuti potrebbero essere interpretati affermando che la dinamica di un canale con Cx26 con una mutazione come la V84L è tanto differente da quella della forma wild type quanto quella di un canale con Cx32 mutata.
Ciò che risulta chiaro dal confronto con la Cx32R220stop è che comunque nello spazio dei

dinasomi le modifiche di un emicanale di membrana che comportano l'assenza di una o più cisteine nelle loop extra-cellulari delle connessine ne variano la dinamica in maniera significativa.

Dall'analisi delle componenti degli autovettori ricavati tramite la PCA abbiamo dedotto quali residui hanno le maggiori fluttuazioni nella dinamica di tali canali e in che misura. Inoltre, analizzando i coefficienti RMSF di ogni atomo, abbiamo verificato che le fluttuazioni maggiori corrispondono agli atomi individuati dalla PCA. Sarà quindi poi possibile concentrare l'analisi su tali fluttuazioni.

Studiare come le differenze nella dinamica inficino la funzionalità del canale e quindi della cellula stessa esula dagli scopi di questa tesi. È comunque significativo aver evidenziato che il calcolo del dinasoma, in precedenza utilizzato per classificare le proteine in base alla loro dinamica e per studiare le relazioni tra la struttura e la dinamica e tra la dinamica e la funzione di una proteina, può essere utile anche nel caso di proteine di membrana cellulare. Questo mezzo, da poco ideato come modello universale e quindi con ancora poche applicazioni, può dare informazioni quantitative sulla dinamica di tali canali e può essere utilizzato per analizzarne le caratteristiche in molti casi differenti, allo scopo di studiarle nel dettaglio successivamente, concentrandosi su diversi aspetti del moto browniano degli atomi delle proteine. L'applicazione della fisica ad un problema di natura biologica quale una mutazione genetica di una proteina è in questo caso un ottimo mezzo per comprenderne le gravi conseguenze funzionali e fisiologiche, che influiscono sulla salute e sulla vita stessa degli individui.

Appendice A

Analisi dello spettro degli autovalori

Si prova che gli autovalori della matrice di covarianza sono inversamente proporzionali ai propri indici [8].

Detta $\vec{x}$ la N-upla delle coordinate unidimensionali delle N particelle, le componenti della matrice di covarianza sono:

$$C_{ij}=\frac{1}{T}\int_0^T (x_i(t)-\langle x_i\rangle)(x_j(t)-\langle x_j\rangle)\,dt$$

e definiamo le *proiezioni degli scarti sul k-esimo autovettore* $\vec{e}_k$ della matrice di covarianza:

$$p_k(t)=\vec{e}_k\cdot(\vec{x}(t)-\langle\vec{x}\rangle)$$

ove $\langle x_i\rangle=\frac{1}{T}\int_0^T x_i(t)\,dt$.

Il primo autovettore $\vec{e}_1$ è per definizione il vettore di norma unitaria tale che, proiettandosi su di esso, la fluttuazione quadratica media è massima, ossia quello che massimizza la quantità:

$$\lambda_1\equiv max_{\|e\|=1}\frac{1}{T}\int_0^T (\vec{e}_1\cdot(\vec{x}(t)-\langle\vec{x}\rangle))^2\,dt$$

che corrisponde quindi al primo autovalore della matrice di covarianza.
E' sempre possibile sviluppare ciascuno scarto $x_i(t)-\langle x_i\rangle$ in una base di funzioni modulo quadro integrabili $f_k(t)$:

$$x_i-\langle x_i\rangle=\sum_{k=0}^{\infty} c_i^k f_k(t)$$

purché valga la condizione di ortonormalizzazione:

$$\int_0^T f_k(t)f_l(t)\,dt=\delta_{kl} \qquad (1)$$

Essendo la famiglia delle $f_k(t)$ una base per lo spazio di Hilbert cui appartengono gli scarti, è possibile dimostrare che è valido lo sviluppo in serie di Fourier [19], dunque:

$$c_i^k = (x_i(t) - \langle x_i \rangle, f_k(t)) = \int_0^T f_k(t)(x_i(t) - \langle x_i \rangle)\, dt$$

ove con la simbologia $(\ ,\ )$ si è inteso il prodotto scalare tra due funzioni modulo quadro integrabili, così definito:

$$(f(t), g(t)) = \int \bar{f}(t)\, g(t)\, dt$$

ove con $\bar{f}(t)$ si intende la complessa coniugata della funzione *f(t)*, in questo caso pari alla funzione stessa essendo questa a valori reali.

Si costruisce in tal modo una famiglia di vettori $\vec{c}^k$ composti dai coefficienti c_i^k dello sviluppo in serie.

Per comodità, si sceglie ciascuna $f_k(t)$ in modo tale che sia proporzionale alla proiezione $p_k(t)$. In questo modo, essendo gli autovettori tra loro ortogonali, al calcolo del k-esimo autovalore λ_k contribuisce solamente la k-esima funzione f_k :

$$\lambda_1 = max_{\|e\|=1} \frac{1}{T} \int_0^T (\vec{e}_1 \cdot \sum_0^\infty \vec{c}^k f_k(t))^2\, dt = max_{\|e\|=1}\, max_{f_1} \frac{1}{T} \int_0^T (\vec{e}_1 \cdot \vec{c}^1 f_1(t))^2 dt = max_{\|e\|=1}\, max_{f_1} \frac{1}{T} (\vec{e}_1 \cdot \vec{c}^1)^2$$

Inoltre, risulta evidente dalle definizioni date che:

$$\int_0^T f_k(t)\, dt \approx \int_0^T \vec{e}_k \cdot (\vec{x}(t) - \langle \vec{x} \rangle)\, dt = 0 \qquad (2)$$

Affinché e_1 sia tale da massimizzare λ_1 , risulta $\vec{e}_1 = \frac{\vec{c}_1}{\|\vec{c}_1\|}$.

Mostriamo ora quale forma debba avere $f_1(t)$ affinché λ_1 sia massimizzato.
Per grandi N è possibile approssimare $\vec{c}^1 \cdot \vec{c}^1$ con una sua media di ensemble:

$$\lambda_1 = max_{f_1} \frac{1}{T} \frac{(\vec{c}^1 \cdot \vec{c}^1)^2}{\|c^1\|} = max_{f_1} \frac{1}{T} \vec{c}^1 \cdot \vec{c}^1 \approx max_{f_1} \frac{1}{T} \langle \vec{c}^1 \cdot \vec{c}^1 \rangle = max_{f_1} \frac{1}{T} \langle N\, c_i^1 c_i^1 \rangle = \lambda'_1$$

Analogamente, si approssima $f_1(t)$ con la funzione $f_1'(t)$ che massimizza λ_1' , il cui integrale nell'intero tempo T è ancora nullo. Si calcola dunque il valore di aspettazione di $c_i^1 c_i^1$ [8, appendice A], integrando per parti e sostituendo i valori caratteristici del moto browniano:

$$\langle (c_i^1)^2 \rangle = \langle (\int_0^T f_1'(t)\, x_i(t)\, dt)^2 \rangle =$$

$$= \langle (-\int_0^T (\int_0^t f_1(v)\, dv) \frac{dx_i(t)}{dt}\, dt + [x_i(t) \int_0^t f_1(v)\, dv]_{t=0}^{t=T})^2 \rangle =$$

$$= \langle (-\int_0^T \int_0^t f_1(v)\, dv \frac{dx_i(t)}{dt}\, dt)^2 \rangle =$$

$$= \langle \int_0^T \int_0^t f_1(v)\, dv \frac{dx_i(t)}{dt}\, dt \int_0^T \int_0^u f(w)\, dw \frac{dx_i(u)}{du}\, du \rangle =$$

$$= \int_0^T \int_0^T \int_0^t f(v)\, dv \int_0^u f(w)\, dw \langle \frac{dx_i(t)}{dt} \frac{dx_i(u)}{du} \rangle\, du\, dt =$$

$$= \int_0^T \int_0^T \int_0^t f(v)\, dv \int_0^u f(w)\, dw\, 2D\delta(t-u)\, du\, dt =$$

$$= 2D \int_0^T (\int_0^t f_1'(u)\, du)^2 dt$$

Si pone $g(t) = \int_0^t f_1'(u)\, du$ e, date le condizioni al contorno $g(0) = \int_0^0 f_1'(u)\, du = 0$ e $g(T) = \int_0^T f_1'(u)\, du = 0$, si usa il metodo di massimizzazione dei moltiplicatori di Lagrange μ_1 e μ_2 per le costrizioni date da (1) e (2):

$$\int_0^T (g(t))^2 + \mu_1 g'(t) + \mu_2 ((g'(t))^2 - 1)\, dt = \int_0^T L(t, g, g')\, dt$$

Ciò è equivalente a risolvere l'equazione di Eulero-Lagrange:

$$\frac{\partial L}{\partial g} - \frac{d}{dt} \frac{\partial L}{\partial g'} = 2g(t) - 2\mu_2 g''(t) = 0 \quad .$$

E' possibile dimostrare che le possibili soluzioni di tale equazione sono tutte le $g_s(t) = C \sin(\frac{\pi s t}{T})$ con s=0, 1, 2, …

Differenziando e utilizzando le proprietà di ortonormalizzazione, ricaviamo che la funzione $f_1'(t)$ cercata può essere pari a una qualsiasi funzione $h_s(t)$ appartenente a un set di funzioni tra loro ortonormali $h_s(t) = \sqrt{\frac{2}{T}} \cos(\frac{\pi s t}{T})$ con s=1 ,2, ...

Sostituendo, si trova che $\langle (c_i^1)^2 \rangle = \frac{T}{\pi^2 s^2}$ e $\lambda_1' = \frac{N}{T} \langle c_i^1 c_i^1 \rangle = \frac{2NDT}{\pi^2 s^2}$.

Tra tutte le possibili funzioni h_s ricavate, si trova che la varianza è massimizzata per s=1, ossia per $f_1' = h_1$.

Ripetendo il calcolo per tutti gli autovalori, con le ulteriori costrizioni date dalla richiesta che le proprietà di ortonormalizzazione in (1) valgano per tutti i $1 \leqslant l < k$, con k ordine dell'autovalore, si ottiene:

$$\lambda_k^{'} = \frac{N}{T} \langle c_i^1 c_i^1 \rangle = \frac{2NDT}{\pi^2 k^2}$$

Riprendendo lo sviluppo dello scarto $x_i(t) - \langle x_i \rangle$:

$$x_i(t) - \langle x_i \rangle \approx \sum_0^\infty c_i^k f_k^{'}(t) = \sqrt{\frac{2}{T}} \sum_{k=1}^\infty c_i^k \cos\left(\frac{\pi k t}{T}\right) \quad \text{con} \quad c_i^k = \sqrt{\frac{2}{T}} \int_0^T x_i(t) \cos\left(\frac{\pi k t}{T}\right) dt$$

si deduce che, se gli atomi della proteina seguono l'andamento di una diffusione random, gli autovalori corrispondenti sono pari a:

$$\lambda_k = max_{\|e_k\|} max_{f_k} \frac{1}{T} (\vec{e}_k \cdot \vec{c^k})^2 \approx \frac{1}{T} \frac{2}{T} \left(\sum_{t=1}^T p_k(t) \cos\left(\frac{\pi k t}{T}\right)\right)^2$$

Appendice B

Teorema di Shannon

Dato un sistema costituito da molti componenti elementari, come particelle microscopiche, è possibile dare una descrizione completa dello stato in cui si trova tramite un insieme di parametri macroscopici. Un medesimo stato macroscopico può però essere ottenuto da più stati microscopici, ossia da più combinazioni di stati di ciascuna componente elementare. L'insieme degli stati microscopici del sistema può essere visto come un insieme finito di alternative A_1 , … A_n mutualmente esclusive, ciascuna delle quali può essere decomposta in più alternative parziali, anch'esse mutuamente esclusive. Definiamo una distribuzione di probabilità discreta $x_1^{rel},\dots x_n^{rel}$ tale che definisce per ogni *i* la probabilità che sia realizzata l'alternativa A_i . È sempre possibile stimare una funzione $S_n(x_1^{rel},...,x_n^{rel})$ che descrive l'incertezza associata a tale distribuzione di probabilità.
Ciò è espresso formalmente dal teorema di Shannon [15].
Esso afferma che, data una funzione $S_n(x_1^{rel},...,x_n^{rel})$ definita sul dominio $x_i^{rel}\geq 0$, $\sum x_i^{rel}=1$, che goda delle seguenti proprietà (ragionevolmente possedute da una funzione che quantifichi l'incertezza associata ad una distribuzione di probabilità):

1) $S_n(x_1^{rel},\dots,x_n^{rel})$ continua nelle x_i per ogni n;
2) $S_n(x_1^{rel},\dots,x_n^{rel})$ simmetrica nei suoi argomenti (indipendenza dall'indice);
3) $S_{n+1}(x_1^{rel},\dots,x_n^{rel},0)=S_n(x_1^{rel},\dots,x_n^{rel})$ (considerare un'alternativa in più irrealizzabile non fa aumentare l'incertezza legata al sistema);
4) $S_n(x_1^{rel},\dots,x_n^{rel})\leq S_n(\frac{1}{x_1^{rel}},\dots\frac{,1}{x_n^{rel}})$ (l'incertezza è massima quando tutte le alternative sono ugualmente probabili) ;
5) posto $x^{relk}=x_1^{rel}+...+x_m^{rel}$ con $\sum_{k=1}^{n}\sum_{l=1}^{m} x^{relk}{}_l=1$ si ha che

$$S_{nm}(x_1^{rel(1)},\dots x_m^{rel(n)})=S_n(x_1^{rel},\dots x_n^{rel})+\sum_{k=1}^{n} x_k^{rel}\,S_m(\frac{x^{relk}{}_1}{x_k}\dots\frac{x^{relk}{}_m}{x_k})$$

(se un'alternativa è decomponibile in m alternative mutuamente esclusive bisogna aggiungere all'incertezza quella dovuta al dover scegliere tra le sotto alternative alle quali si assegna probabilità condizionata);

allora

$$S_n(x_1^{rel}\dots x_n^{rel})=-\sum_{k=1}^{n} x_k^{rel}\ln x_k^{rel} \quad .$$

Bibliografia

1. Martin R.B. Free energy and equilibria of peptide bond hydrolysis and formation. Biopolymers, 45(5): 351–353, 1998.
2. Michel Daune. Molecular biophysics. Structeres in motion. Oxford University Press, 1999.
3. Anfinsen C, Haber E (1961) Studies on the reduction and re-formation of protein disulfide bonds. J Biol Chem 236: 1361–1363.
4. Pascual-Garcia A, Abia D, Mendez R, Nido GS, Bastolla U (2010) Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins* 78: 181–196.
5. Hensen U., Meyer T., Haas J., Rex R., Vriend G., Grubmuller H. (2012) Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. PLoS One 7.
6. A. Einstein (1956) , Investigations on the theory of the Brownian movement.
7. D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, GROMACS User Manual version 4.6.5, www.gromacs.org (2013)
8. Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev.* E 62:8438–8448, 2000.
9. N. J. D. Nagelkerke, *Biometrika*, Vol. 78, No. 3. (Sep., 1991), pp. 691-692.
10. M. Loreti, Teoria degli errori e fondamenti di statistica (2006), pp: 283-291.
11. Uhlenbeck GE, Ornstein LS (1930) On the theory of the Brownian motion. Phys Rev 36: 823–841.
12. Zwanzig R (1988) Diffusion in a rough potential. P Natl Acad Sci Usa 85: 2029–2030.
13. Ansari A, Berendzen J, Bowne S, Frauenfelder H, Iben I, et al. (1985) Protein states and proteinquakes. Proceedings of the National Academy of Sciences 82: 5000–5004.
14. Kabsch W, Sander C (1983) Dictionary of protein secondary structure – patternrecognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
15. F. Cardin, M. Favretti (2013). Modelli fisico matematici, Cleup, cap. 6.
16. S. Maeda, S. Nakagawa, M. Suga, E. Yamashita, A. Oshima, Y. Fujiyoshi, and T. Tsukihara. Nature 458, 597-602 (2009).
17. Zonta, F., G. Polles, G. Zanotti, F. Mammano (2012). Permeation pathway of homomeric connexin 26 and connexin 30 channels investigated by molecular dynamics. J. Biomol. Struct. Dyn. 29:985–998. doi:10.1080/073911012010525027
18. M. Beltramello, V. Piazza, F. F. Bukauskas, T. Pozzan, and F. Mammano. Nature Cell Biology 7, 63-69 (2005).
19. W. Rudin (1976). Principles of mathematical analysis, Third Edition, pp. 185-188.

Riconoscimenti

Per l'uso di AmberTools14:

D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman (2014), AMBER 14, University of California, San Francisco.

Per l'uso di cpptraj:

Daniel R. Roe and Thomas E. Cheatham, III, "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data". J. Chem. Theory Comput., 2013, 9 (7), pp 3084-3095.

Per l'uso di VMD:

This work was supported by the Theoretical and Computational Biophysics group, NIH Center for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign.

Per l'uso di GROMACS:

This work makes use of results produced by the ScalaLife project which is co-funded by the European Commission (under contract number INFSO-RI-261523). More information is available at http://www.scalalife.eu.