Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale
In
Scienze e Tecnologie dei Bio e Nanomateriali

Ordinamento ex D.M. 270/2004

Tesi di Laurea

Analisi computazionale del rate d'accoppiamento di hairpin di DNA con filamenti singoli complementari

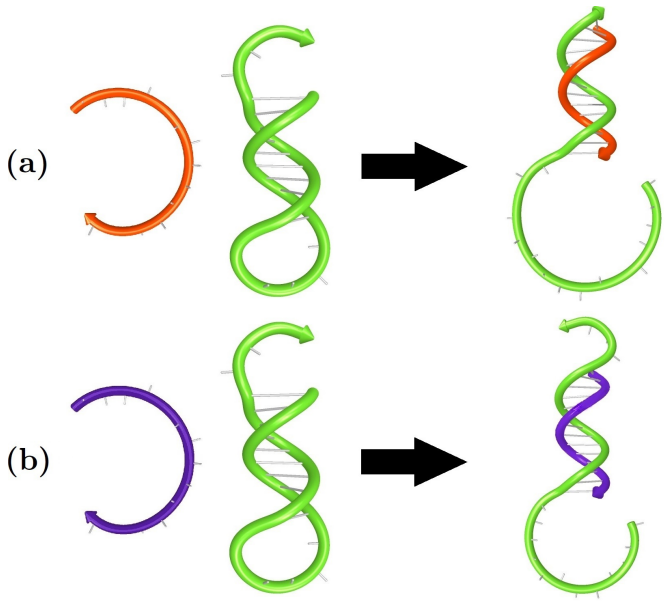
Relatore
Dott. Flavio Romano

Laureando
Daniele Pellicciotta
Matricola 853060

Anno Accademico
2016 / 2017

Sommario

Si è progettato un hairpin di DNA (verde in figura) avente il corpo a doppio filamento lungo 10 paia di basi, un'ansa di 6 basi azotate e una coda di 4; con esso sono stati progettati anche 2 filamenti complementari che potessero ibridarsi con l'hairpin, uno dalla coda(figura a) l'altro dall'interno dell'ansa(figura b). Entrambi i processi sono favorevoli alle temperature considerate, in quanto il numero totale di paia di basi aumenta e l'energia libera totale diminuisce. Lo scopo del lavoro è di valutare i rate d'accoppiamento dell'hairpin con i due diversi filamenti complementari. Per mezzo di un modello coarse-grained per il DNA chiamato *oxDNA* si è provveduto ad eseguire una serie di simulazioni al computer, in più si è fatto uso anche del metodo del *forward flux sampling*(FFS), che permette di suddividere le transizioni di reazione in una serie di interfacce intermedie, facilitando il campionamento di un evento raro. Si è provveduto a ottenere dati a partire da attacchi di 4 basi azotate e poi riducendo questo numero fino a 2. I risultati sono stati valutati confrontando gli errori ottenuti tramite propagazione dell'errore standard e ricampionamento jackknife. Nei risultati si è evidenziato come in tutti i casi i rate di accoppiamento con attacco sulla coda siano più elevati. La tesi è in lingua italiana, ma le immagini presentano didascalie anche in lingua inglese.



Indice

Elenco delle figure

Capitolo 1

Introduzione

Il DNA è uno dei più antichi polimeri a presentarsi in natura, fin dalle origini delle prime forme di vita. Sono passati quasi 60 anni da quando Watson e Crick hanno gettato luce sulla sua struttura a doppia elica che viene a formarsi in presenza di 2 sequenze complementari, grazie ai legami a idrogeno che si formano tra le coppie di basi C-G e A-T e alle interazioni di stacking tra basi adiacenti[41]. Da questa scoperta, la scienza del DNA ha costituito le fondamenta delle attuali ricerche biotecnologiche, come il DNA ricombinante, DNA e RNA anti-senso e i vaccini a DNA, ma l'attenzione sta sempre più spostandosi anche verso un suo uso come nanomateriale nei campi delle nanoscienze e nanotecnologie[21]. Proprio il riconoscimento di sequenze complementari è alla base della costruzione di svariate architetture attraverso l'auto-assemblaggio e di applicazioni che vanno dall'utilizzo in biosensori a sbalzo[26] o che sfruttino la risonanza plasmonica di superficie[46], al suo uso come sagoma nella costituzione di nanocircuiti elettronici[24], all'assemblaggio controllato di liposomi da utilizzare eventualmente come compartimenti separati in reazioni multi-step[23] fino allo studio sull'utilizzo di DNA origami in terapie contro il cancro[22]. Tutti questi sistemi sono basati sull'ibridazione di brevi sequenze di DNA.

Naturalmente, tutte queste applicazioni presuppongono una conoscenza dettagliata della termodinamica e cinetica di ibridazione dei filamenti di DNA utilizzati. La cristallografia ai raggi X e i dati NMR sono stati molto importanti nel fornire informazioni dettagliate sulle strutture tridimensionali di una vasta gamma di oligonucleotidi (anche eliche triple o quadruple) associati o meno a proteine o altre biomolecole. Molte strutture sono catalogate nel Nucleic Acid Database[6, 27], sito[1]. La termodinamica di ibridazione del DNA è già stata ottimamente descritta[32], in quanto può essere approssimata molto bene con un modello a due stati, o si hanno filamenti singoli ben separati o altrimenti delle doppie eliche stabili. La cinetica di ibridazione, invece, è più difficile da comprendere e i risultati non sono ovvi, perché dipendono da stati intermedi a bassa ricorrenza e di difficile accesso sperimentale.

Nel lavoro si vogliono trovare proprietà cinetiche attraverso simulazioni al computer allo scopo di approfondire le conoscenze nelle nanotecnologie a DNA. La modellizzazione al computer può aiutare ad esplorare le complesse vie di transizione che portano a una completa ibridazione di due filamenti di DNA. Sono state già eseguite delle misure sperimentali su queste proprietà cinetiche[45, 11], ma può risultare difficile separate effetti specifici di sequenza da tendenze di base.

Le simulazioni di dinamica molecolare(MD) con potenziali atomici per gli acidi nucleici[10] sono già state utilizzate per comprendere meglio le relazioni tra sequenza e struttura del DNA[43], la sua flessibilità[29], le geometrie di idratazione attorno agli acidi nucleici[44] e le strutture che formano associati a proteine[25]. Poiché questi metodi hanno una risoluzione elevata, però, hanno il limite di richiedere tempi molto lunghi per strutture delle dimensioni rilevanti nel nostro caso e rendono impraticabile lo studio adeguato delle vie di transizione. L'unico modo per rendere praticabili i calcoli necessari è ricorrere a un modello coarse-grained, che però deve essere in grado di riprodurre in modo adeguato le proprietà meccaniche, strutturali e termodinamiche del DNA sia a filamento singolo che doppio.

In questa tesi viene utilizzato un modello coarse-grained di DNA chiamato oxDNA[30, 37], in grado di riprodurre i duplex come delle eliche rigide, mentre il DNA a filamento singolo risulta relativamente più flessibile. È già stato dimostrato come oxDNA riesca a riprodurre la stessa accelerazione del tasso di spiazzamento al variare della lunghezza di un appiglio[36] che era già stata misurata sperimentalmente[45], ma a mia conoscenza non è mai stato valutato nel caso di un hairpin di DNA se e di quanto differiscano i tassi di spiazzamento nei casi in cui l'appiglio si trovi sulla coda al termine del duplex o all'interno dell'ansa. I due casi sono termodinamicamente molto simili e si ipotizza che eventuali differenze dipenderebbero da fattori puramente cinetici. La diversa flessibilità dei filamenti singoli e doppi di DNA riprodotta da oxDNA permette la formazione di hairpin e altre strutture che richiedono questa caratteristica. Il modello si è già dimostrato solido riproducendo adeguatamente fenomeni per i quali non era stato espressamente parametrizzato[36, 16, 13, 31].

Dopo aver approfondito i vari aspetti del modello, delle tecniche di simulazione e dei metodi di calcolo degli errori statistici, si studiano i processi di ibridazione di un hairpin con due diverse sequenze complementari progettate per limitare i complessi indesiderati, una sequenza con attacco sulla coda e l'altra con attacco sull'ansa, ad appigli di lunghezze diverse. I risultati verranno comparati con quelli per il semplice spiazzamento mediato da un appiglio [35], concludendo che i due casi sono effettivamente diversi e c'è un rapporto tra i due rate di spiazzamento di 2 o 3 ordini di grandezza; questo rapporto è il risultato centrale di questa tesi, ed è un numero passibile di verifica sperimentale.

Capitolo 2

Modello

Nell'ambito di una simulazione al computer è necessario avere ben chiare le caratteristiche di ciò che si vuole studiare e che cosa si vuole ottenere, perché ci sono diverse strategie di modellizzazione, ognuna con un suo livello di risoluzione e adeguatezza alla situazione, soprattutto per quanto riguarda i tempi necessari ad ottenere i risultati. In questo capitolo si approfondiranno tali aspetti, le caratteristiche del DNA e il modello che cerca di descriverlo, l'oxDNA.

2.1 Tipi di modellizzazione

Nell'ambito della modellizzazione della materia ci si può imbattere in processi che avvengono su scale di dimensione e tempi estremamente diversi, in particolar modo per processi biologici. Al livello più basso ci sono le reazioni chimiche riguardanti i siti attivi di proteine o altre biomolecole che richiedono una trattazione comprendente persino gli elettroni su un numero relativamente basso di atomi, per arrivare fino alle scale degli aggregati macromolecolari che possono essere anche di milioni di atomi e tempi macroscopici e che richiedono rappresentazioni a più bassa risoluzione. I modelli utilizzati, andando dalle scale più piccole a quelle più grandi, sono:

- Quantomeccanici (QM);
- Meccanica molecolare/Full-atom (MM/FA);
- Coarse Grained (CG).

In figura 2.1 è possibile osservare i diversi tempi e dimensioni coinvolti a diverse scale di risoluzione per una proteina.

L'uso di modelli diversi in base alla dimensione di ciò che deve essere simulato deriva dal concetto di riduzione dei gradi di libertà N_L del sistema; questo perché il costo computazionale di una simulazione scala con un polinomio di N_L che dipende dal tipo di modellizzazione. Ne consegue che ridurre il numero di gradi

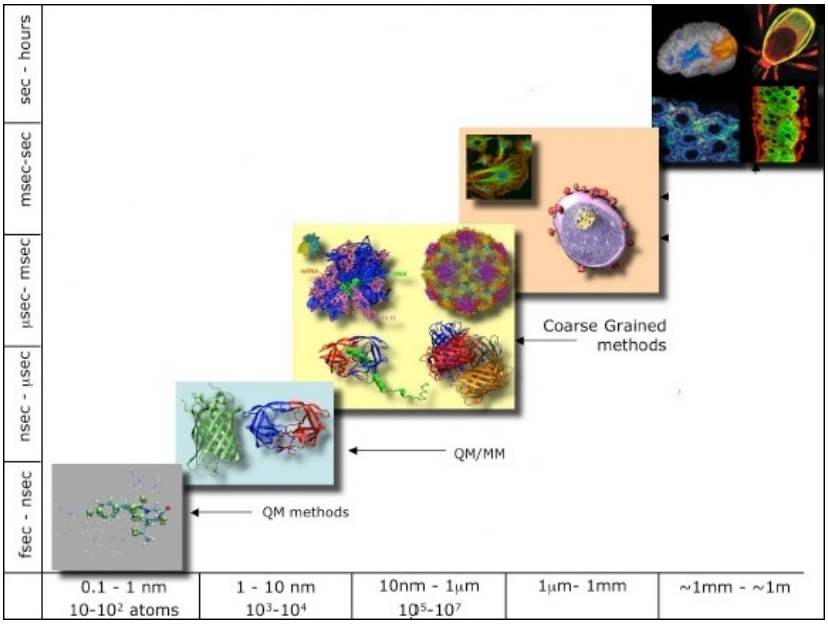


Figura 2.1: Nelle figure, dal basso a sinistra verso l'alto a destra: sito attivo di una proteina; la proteina e proteasi di un virus; la proteasi, un tetramero, un ribosoma e un piccolo virus; una cellula e strutture interne; tessuti, organi e un piccolo organismo. Ai lati sono riportati i tempi dei processi biologici e le dimensioni, anche in numero di atomi, della parte considerata.
Figure 2.1: In the pictures, from the left at the bottom to the right at the top: active site of a protein; the protein and a virus protease, the protease, a tetramer, a ribosome and a small virus; a cell and its internal structures; tissues, organs and a small organism. Times of biological processes and sizes, in number of atoms too, are reported to the sides.

di libertà è il metodo più semplice e diretto per rendere fattibili simulazioni di sistemi che altrimenti richiederebbero tempi troppo lunghi.

Di seguito vengono descritti i vari metodi un po' più in dettaglio, tenendo presente che spesso la descrizione di un processo biologico deve coinvolgere più di una scala di grandezza, rendendo pertanto necessario l'utilizzo anche di approcci multi-scala.

2.1.1 Modello Quantomeccanico (QM)

Questo tipo di modello è adatto per dei processi che comportino cambiamenti importanti nella struttura elettronica del sistema o rotture e formazioni di legami chimici, quindi reazioni chimiche e siti attivi di molecole biologiche. Questo è possibile perché i gradi di libertà elettronici sono trattati in modo esplicito con la meccanica quantistica, attraverso la risoluzione dell'*equazione di Schrödinger*

$$\hat{H}\psi = E\psi \tag{2.1}$$

con $\hat{H}$ l'hamiltoniano, ψ la funzione d'onda e E l'energia totale del sistema con la quale ricavare la superficie di energia potenziale (PES).

Rimane però il problema che la (2.1) è complessa ed irrisolvibile anche per sistemi molto semplici; persino la molecola d'idrogeno, H_2, non è risolvibile in modo esatto a causa della correlazione, ovvero la non indipendenza dei moti delle particelle tra loro. Pertanto nelle simulazioni vengono usate diverse semplificazioni, tra cui una universalmente utilizzata è l'*approssimazione di Born-Oppenheimer* perché semplifica i calcoli ed ha un errore trascurabile sull'energia; deriva dal fatto che essendo la massa dei nuclei atomici molto più grande di quella degli elettroni (il rapporto protone/elettrone è di 1836), questi ultimi hanno moti molto più veloci e pertanto possono essere trattati come in un sistema di nuclei fissi, permettendo di disaccoppiare la funzione d'onda ψ nel prodotto di una nucleare ed una elettronica. Esistono tante altre semplificazioni per l'approssimazione delle funzioni d'onda e dell'operatore, che portano per esempio a metodi come l'*Hartree-Fock* o la *teoria del funzionale densità* (DFT)[8], permettendo calcoli più o meno precisi e veloci per ottenere la PES.

In generale questo tipo di trattazione è adeguato per piccole molecole o parti importanti del sistema in esame compresi tra 10 e 100 atomi, dato che la trattazione degli elettroni comporta un gran numero di gradi di libertà e quindi numeri superiori comporterebbero costi computazionali estremamente alti.

2.1.2 Meccanica Molecolare/Full-atom (MM/FA)

La simulazione di sistemi più grandi nell'ordine di 10^4-10^5 atomi, come un'intera proteina o un acido nucleico in ambiente acquoso, sarebbero troppo pesanti per poter essere eseguite con un approccio di tipo QM, pertanto gli effetti degli elettroni sono trattati implicitamente secondo uno schema dove gli atomi e i legami chimici vengono equiparati a "sfere e molle" con equazioni di tipo classico da cui ricavare i vari contributi all'energia totale del sistema, riassunte nella figura 2.2.

Il MM/FA è altamente parametrizzato, infatti i vari contributi all'energia vengono calcolati attraverso costanti di forza, angoli e distanze di legame all'equilibrio, elettronegatività, ecc. I vari parametri, però, possono variare anche per lo stesso elemento chimico considerato, se per esempio di una ibridazione diversa, perciò si parla di *tipi atomici*. Prendendo ad esempio il carbonio, questo avrà diversi set di parametri, o tipi atomici, a seconda che sia alifatico, aromatico, chetonico, ecc. Anche le cariche parziali possono essere assegnate in funzione dell'atomo stesso e quelli ai quali è legato.

Esistono moltissimi modelli di tipo MM/FA che differiscono per la parametrizzazione, sia come numero di tipi atomici sia come determinazione dei valori dei parametri, per la forma e per il numero di termini energetici; ognuno può

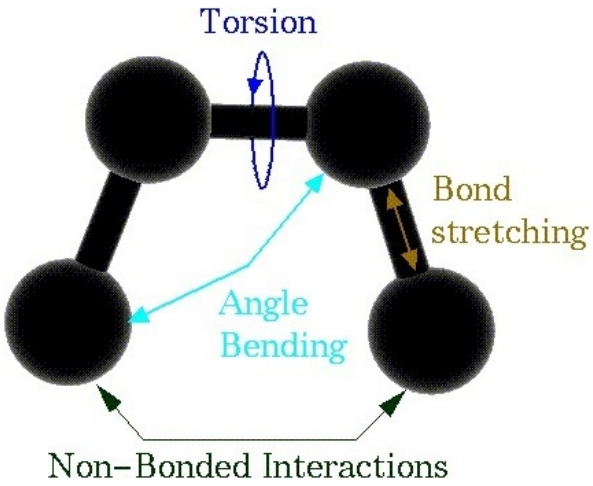


Figura 2.2: Nei modelli MM/FA la PES viene approssimata con una somma di termini empirici analitici con un significato chimico-fisico. Si distinguono contributi di legame e non legame. Quelli di legame comprendono gli stiramenti, i piegamenti e le torsioni e sono descritti attraverso formule di deformazione delle molle; quelli di non legame comprendono ponti a idrogeno, interazioni di Van der Waals e elettrostatiche, più facili da descrivere attraverso atomi considerati come sfere con dei raggi e cariche caratteristici.
Figure 2.2: In MM/FA models the PES is approximated by a sum of analitical empirical terms with physical and chemical meaning. There are bonding and non bonding contributions. The first ones consist of stretching, bending and torsions and they are described by spring deformation formulas; non bonding contributions consist of hydrogen bond, electrostatic and Wan der Waals interactions, easier to describe with atoms considered as sferes with caracteristic radii and partial charges.

adattarsi meglio a un diverso sistema fisico per il quale è stato ottimizzato nel tempo. I sistemi inorganici presentano solitamente un numero di tipi atomici inferiore rispetto a quelli biologici, ma le interazioni di correlazione che presentano sono molto importanti e richiedono funzioni molto complesse; alcuni esempi possono essere i set di parametri MM2, MM3 e CFF. Nei sistemi biologici, al contrario, nonostante vengano utilizzati principalmente pochi elementi, sono presenti centinaia di tipi atomici, ma data la minore complessità delle interazioni è possibile avere funzioni relativamente semplici; per le biomolecole, come proteine e acidi nucleici, si possono utilizzare i set di parametri CHARMM, AMBER o GROMOS. Per la logica di costruzione del modello, naturalmente, non è possibile simulare situazioni in cui si abbia formazione o rottura di legami, polarizzazioni o studiare stati di transizione.

2.1.3 Modello Coarse grained (CG)

Dopo aver eliminato i gradi di libertà quantistici e gli elettroni, rappresentando comunque tutti gli atomi del sistema, il passo successivo per simulare sistemi

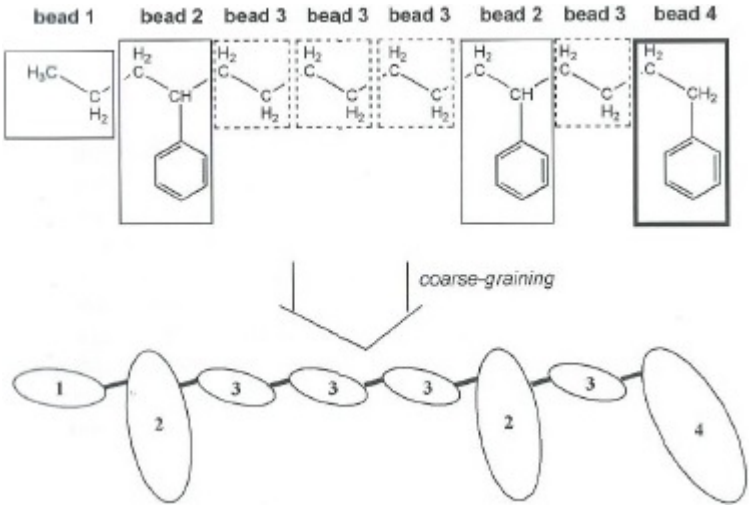


Figura 2.3: Gruppi di atomi che si ripetono, come gli amminoacidi nelle proteine o i nucleotidi negli acidi nucleici, vengono sostituiti dai beads e i legami tra i vari gruppi vengono simulati attraverso interazioni tra i beads con formule opportune. Come avviene per i tipi atomici nei modelli MM/FA, i beads devono essere opportunamente parametrizzati con dei dati sperimentali. Immagine da [8].
Figure 2.3: Repetitive groups of atoms, like amino acids in proteins and nucleotides in nucleic acids, are replaced with beads and the bonds between the various groups are simulated by interactions between beads with appropriate formulas. As is with atomic types in MM/FA models, the beads must be parameterized to reproduce experimental data. Image from [8].

ancora più grandi (fino a milioni di atomi) come gli aggregati macromolecolari, è l'ulteriore riduzione di gradi di libertà classici nei modelli coarse grained, con un guadagno in costi computazionali collegato al numero di gradi eliminati. In questo modello, infatti, interi gruppi di atomi vengono rimpiazzati da dei *beads*(grani)[38] come visibile nella figura 2.3.

Naturalmente non è possibile rimuovere gradi di libertà a caso, ma è necessario che questi abbiano alcune caratteristiche per poter essere eliminati:

- Non devono essere importanti per le proprietà o il processo in esame;
- Sono computazionalmente pesanti o in grande numero, in modo tale che la perdita di accuratezza sia minima o comunque più che compensata dal guadagno nei calcoli.
- Le interazioni che li riguardano devono preferibilmente essere disaccoppiate da quelle degli altri gradi di libertà.

Più i gruppi di atomi sostituiti dai beads sono grandi e inferiore sarà la risoluzione del modello, in cui diventa tra l'altro necessario includere implicitamente l'effetto delle variabili eliminate, come per esempio i legami a idrogeno,

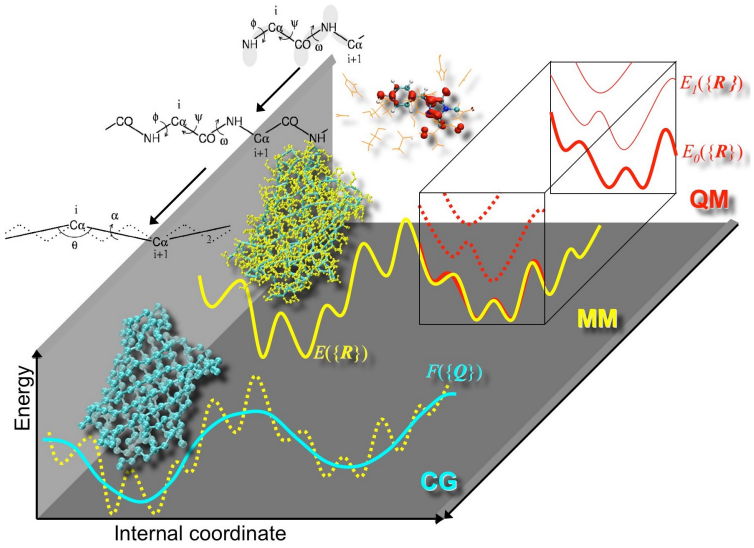


Figura 2.4: Approssimazioni della PES nel passaggio tra i vari livelli descrittivi per una proteina, dal QM (sul retro) al MM/FA (al centro) al CG (di fronte).
Figure 2.4: Approximations of the PES by passing between the various description levels for a protein, from QM (at the back) to MM/FA (center) to CG (front).

rendendo di conseguenza più complesso e difficile parametrizzare i vari termini interattivi. La progettazione dei campi di forze del modello può seguire due approcci differenti in base all'uso che se ne deve fare:

Bottom-up La parametrizzazione avviene seguendo le proprietà meccaniche di modelli più dettagliati. Sono adatti per studiare fluttuazioni intorno alla struttura in questione all'equilibrio, ma possono favorire le strutture dalle quali sono stati ricavati i parametri e possono avere difficoltà nel riprodurre proprietà termodinamiche. Inoltre l'accuratezza del modello di riferimento nel descrivere alcune transizioni può non essere nota.

Top-down Cercano una corrispondenza con quantità termodinamiche globali tramite comparazione con dati sperimentali. Si prestano allo studio in particolare di transizioni strutturali, ma sono limitati dalla disponibilità di dati sperimentali sufficienti alla loro parametrizzazione.

Uno degli effetti del coarse graining è quello di trasformare la superficie di energia potenziale in una superficie di energia libera per via dei fattori entropici e di energia potenziale che si portano dietro le variabili eliminate. Le diverse semplificazioni utilizzate hanno l'ulteriore effetto di "lisciare" la superficie a livello microscopico(figura 2.4), causando una semplificazione della dinamica. Da un lato si eliminano molti minimi locali e si evita che il sistema possa bloccarsi in uno di essi, dall'altro si ha un aumento fittizio della velocità alla quale avvengono i vari processi e bisogna tenerne conto.

Con il coarse graining è possibile descrivere il sistema in esame in modo almeno qualitativo, ottenendo una struttura che mantenga degli aspetti del sistema originario. Non considerando in modo esplicito ogni singolo atomo, non è possibile analizzare proprietà legate specificatamente ad essi.

2.2 DNA

Come noto, il DNA è alla base dei meccanismi di ereditarietà, ma per poterli comprendere è stato necessario conoscerne la sua struttura. Grazie alle regole di Chargaff[9], per cui in una molecola di DNA a filamento doppio globalmente la percentuale di adenina(A) eguaglia quella di timina(T) e ugualmente per citosina(C) e guanina(G) (in breve %A=%T e %C=%G), unitamente a profili di diffrazione ai raggi X di fibre di DNA, Watson e Crick sono riusciti a proporre il modello di DNA[41] che avrebbe enormemente influenzato il corso della biologia e di tutti i suoi rami di ricerca. Dato che molti processi biologici si basano sulle interazioni proteina-acido nucleico, le sequenze di basi hanno un profondo effetto sulle strutture tridimensionali caratteristiche del DNA o del RNA e quindi sulla natura di processi biologici fondamentali.

I progressivi miglioramenti nelle tecnologie non hanno fatto altro che sostenere tutti i campi di ricerca biologici fino al giorno d'oggi, in cui la possibilità di sintetizzare specifiche sequenze di DNA ha aperto le porte a una serie di ricerche sulle nanotecnologie basate sul riconoscimento specifico di sequenze complementari di DNA.

2.2.1 Composizione chimica

Il modello di Watson e Crick descriveva il DNA come una doppia elica destrogira composta da due catene polinucleotidiche tenute insieme al centro da i legami idrogeno instaurati nell'accoppiamento di basi azotate (A-T, G-C) e all'esterno da uno scheletro zucchero-fosfato. Proprio l'accoppiamento delle basi complementari è alla base dei meccanismi di replicazione del DNA e dell'ereditarietà.

I monomeri del DNA sono i nucleotidi. Ognuno di essi è costituito da un gruppo fosfato, uno zucchero a 5 atomi di C, il 2-deossiribosio, e da una base purinica o pirimidinica. I quattro tipi di nucleotidi sono riportati nella figura 2.5, mentre la sola unità contenente zucchero e base azotata viene detta nucleoside. Di fatto gli acidi nucleici sono polimeri di nucleosidi monofosfati e formano una struttura come nella figura 2.6, che a livello tridimensionale risulta in una doppia elica le cui caratteristiche sono approfondite nella sezione successiva. I fosfati, che sono gruppi polari carichi negativamente, sono rivolti verso l'esterno dell'elica e sono disponibili per interazioni fisiche e chimiche con molecole d'acqua, gli ioni metallici e proteine basiche (come gli istoni

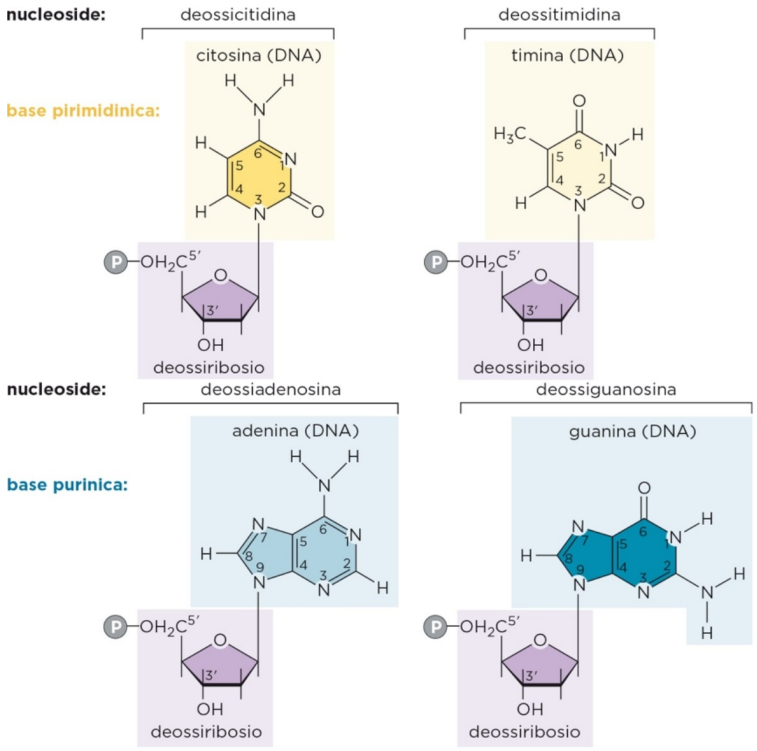


Figura 2.5: Nucleotidi delle quattro basi azotate divise in pirimidine, citosina(C) e timina(T), e purine, adenina(A) e guanina(G). Sono messi in evidenza i carboni 3′ e 5′ del deossiribosio sui quali si può instaurare il legame estere con un fosfato, in questo caso il legame è mostrato sempre sul carbonio 5′. Quando il fosfato si lega da un lato a un carbonio 5′ di un nucleoside e dall'altro a uno 3′ del successivo si instaura un legame fosfodiesterico, che tiene uniti i monomeri. Immagine da [42].
Figure 2.5: Nucleotides of the four nitrogenous bases divided in pyrimidines at the top, citosine(C) and timine(T), and purines at the bottom, adenine(A) and guanine(G). In evidence are the 3′ and 5′ carbons of the deoxyribose where a phophate can form an ester linkage, in this case the linkage is always shown on the 5′ carbon. When a phosphate forms a link between the 5′ carbon of a nucleoside and the 3′ carbon of another one, a phosphodiester bond that connects the monomers is established. Image from [42].

per i superavvolgimenti del DNA nella cromatina delle cellule). Nel RNA il deossiribosio è rimpiazzato dal ribosio e la timina dall'uracile(U).

2.2.2 Forme B, A e Z del DNA

Sebbene i principi essenziali del primo modello di DNA siano rimasti intatti, nel corso del tempo ne sono state approfondite diverse forme e sono state apportate delle correzioni ad alcune misure caratteristiche. Per esempio il modello di Watson e Crick si riveriva a quello che oggi è conosciuto come B-DNA, la forma dominante in condizioni fisiologiche.

Un'altra forma piuttosto comune è il A-DNA, emerso da esperimenti di scattering ai raggi X a umidità inferiori. Si presenta anche come forma prevalente nei duplex di RNA, negli ibridi DNA-RNA e persino nei duplex di DNA in condizioni estreme di solvatazione per certe sequenze. Entrambe le forme A e B di DNA sono destrogire.

Esiste una terza forma particolare di DNA che presenta un'elica levogira ed è stata chiamata Z-DNA per via dell'andamento a zig-zag dello scheletro zucchero-fosfato. Questa forma è stata osservata per polimeri di sequenze di purine-pirimidine alternate (come CGCGCGCG) ad alte concentrazioni di sali, necessarie per schermare le cariche negative dei fosfati che in questa forma vengono a trovarsi più vicine e minimizzando quindi le interazioni repulsive. La funzione biologica del Z-DNA non è ancora chiara. Strutture d'esempio delle varie forme sono illustrate nella figura 2.7, mentre i dati relativi sono riportati nella tabella 2.1.

In tutte le forme si definiscono una scanalatura maggiore e una scanalatura minore. Nella prima si affacciano i gruppi N7 e C6 delle purine e i gruppi C4 e C5 delle pirimidine, che sono siti per la formazione di legami idrogeno. Essendo più ampia e quindi più raggiungibile nel B-DNA, molte proteine per sequenze specifiche interagiscono con la scanalatura maggiore, ma ci sono alcune proteine che interagiscono con la scanalatura minore.

Tabella 2.1: Dati geometrici delle tre forme di DNA descritte. Le misure riportate sono come indicate nelle figure 2.6 e 2.7. $\Omega = 360/N$ è la torsione dell'elica attorno al suo asse tra coppie di basi adiacenti, mentre η rappresenta l'inclinazione dei piani formati dalle basi rispetto alla perpendicolare all'asse dell'elica. Si consideri che l'unità ripetitiva per il Z-DNA è un dinucleotide piuttosto che un mononucleoride.
Table 2.1: Geometric data of the three described DNA forms. The reported measures are as marked on figures 2.6 e 2.7. $\Omega = 360/N$ is the rotation about the helix axis between two neighboring base pairs, while η is the inclination of the planes formed by the bases with respect to the pependicular to the helix axis. The repeating unit for Z-DNA is a dinucleotide rather than a mononucleotide.

Proprietà	A-DNA	B-DNA	Z-DNA
N	11	10.5	12
h	2.6 Å	3.4 Å	3.8 Å
H	28.6 Å	35.7 Å	45.6 Å
D	23 Å	20 Å	18 Å
d	6 Å	7 Å	5.7∼6.1
Ω	∼33°	∼34°	∼60°/dimero
η	∼+19°	∼-1.2°	∼-9°

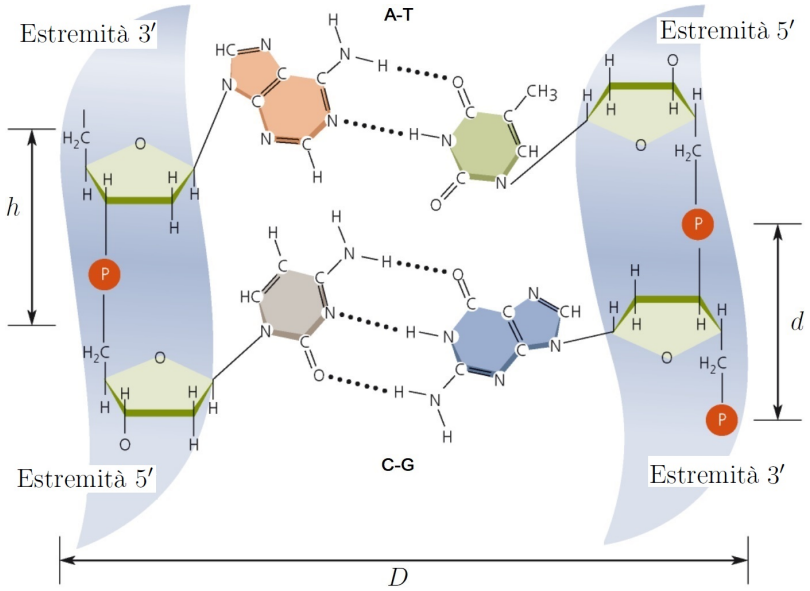


Figura 2.6: Composizione chimica del DNA. I legami fosfodiesterici uniscono i nucleotidi formando lo scheletro zucchero-fosfato esterno, mentre all'interno si ha l'accoppiamento delle basi. I due filamenti vanno in direzioni antiparallele. Convenzionalmente la sequenza di basi in un polinucleotide è specificata nella direzione $5' \to 3'$ e il filamento complementare è automatico per coppie ideali. Notare che la coppia CG interagisce attraverso 3 legami idrogeno(rappresentati da dei punti), mentre la coppia AT ne forma solo 2. La larghezza dei due complessi è quasi identica, ma il numero di legami idrogeno conferisce loro diversa stabilità. Il DNA ricco in CG pertanto risulta più stabile di quello ricco in AT. Il diametro del cilindro elicoidale D rimane approssimativamente costante lungo tutta la struttura proprio grazie all'appaiamento delle basi. Appaiamenti tra due purine o due pirimidine risulterebbero troppo ingombranti nel primo caso o in un restringimento eccessivo nel secondo, destabilizzando la struttura. Un'ulteriore stabilizzazione è data dalla sovrapposizione degli anelli aromatici delle basi azotate nell'elica, con instaurazione di interazioni di Van der Waals. Sono riportati la distanza fosfato-fosfato d e la distanza verticale lungo l'asse della doppia elica tra coppie di basi adiacenti h. Immagine da [5].

Figure 2.6: DNA chemical composition. The phosphodiester bridge links the nucleotides forming the external sugar-phosphate backbone, while in the inside there is the base pairing. The two strands go in antiparallel directions. Conventionally, the base sequence in a polynucleotide is specified for the $5' \to 3'$ direction and the complementary strand is automatic for ideal pairs. Note that CG pairs have 3 hydrogen bonds(represented by dots), while AT pairs only have 2. The widths of the two complexes are nearly identical, but the number of hydrogen bonds gives them different stability. CG rich DNA is therefore more stable than AT rich DNA. The helical cylinder diameter D is approximately constant along all the structure thanks to the pairing of bases. Pairings between two purines or two pyrimidines would be too wide for the first case or too narrow for the second, destabilizing the structure. Another stabilization is given by the stacking of the aromatic rings of the bases in the helix, arising from Van der Waals interactions. The phosphate-phosphate distance d and the vertical distance along the helix axis between adiacent pairs of bases h are marked. Image from [5].

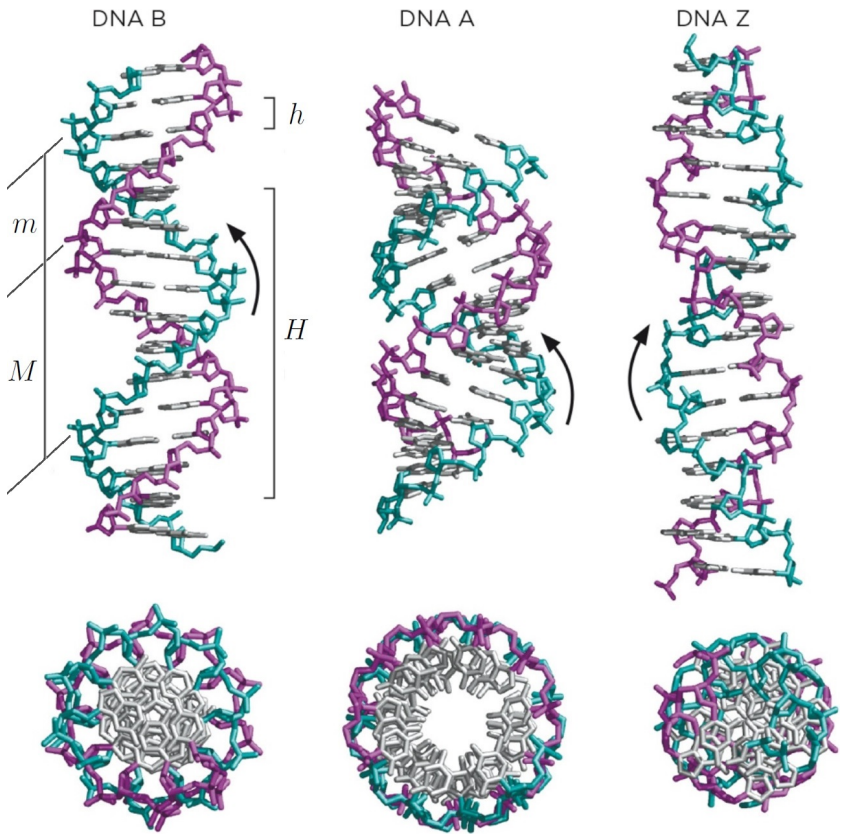


Figura 2.7: Strutture a doppia elica di DNA nelle tre diverse forme B, A e Z, viste dal lato in alto e viste dall'alto in basso. Il senso di rotazione delle eliche è indicato dalle frecce. Il passo dell'elica H è la distanza lungo l'asse per un giro completo, a cui si associa il numero di coppie di basi per giro N. Nel B-DNA la scanalatura maggiore M, definita come il lato più lontano dai due legami glicosidici di una coppia di basi, è più ampia (12 Å) e profonda (8.5 Å), mentre la scanalatura minore m è più stretta (6 Å) e quasi altrettanto profonda (7.5 Å). Le due scanalature sono generate dall'inclinazione delle basi rispetto alla perpendicolare all'asse dell'elica, che nel B-DNA è minima. Nel A-DNA l'inclinazione delle basi è notevole, con stacking delle basi combinato sia intra che interfilamento; m non è profonda come nel B-DNA, ma M è più stretta e profonda di m. Nel Z-DNA le basi sono leggermente inclinate, M sporge all'esterno e m è stretta e profonda. Immagine da [42].

Figure 2.7: Structures of DNA duplexes in the three different conformations B, A and Z, looked at from the side at the top and looked at from the top at the bottom. The arrows indicate the rotation sense of the helixes. The helix pitch H is the distance along the helix axis for one complete turn, at wihch can be associated the number of pairs of bases for every turn N. For B-DNA the major groove M, defined as the side farther from the two glycosyl linkages of a pair of bases, is wider(12 Å) and deeper (8.5 Å), while the minor groove m is narrower(6 Å) and slightly less deep(7.5 Å). The two grooves are generated from the inclination of the bases with respect to the pependicular to the helix axis, which in B-DNA is very small. In A-DNA there is a prominent inclination, with combination of both intrastrand and interstrand base stacking; m is less deep than in B-DNA, but M is narrower and deeper than m. In Z-DNA the bases are slightly inclined, M bulges out and m is narrow and deep. Image from [42].

2.3 oxDNA

L'oxDNA è un modello coarse grained per il DNA [30], la cui documentazione e il codice sono disponibili in download [2]. Nel modello qui descritto non c'è dipendenza da sequenze specifiche e le basi azotate interagiscono solo tra coppie complementari (A-T, G-C), con un utilizzo di "basi medie" che rende di fatto indistinguibili tra loro le forze delle diverse coppie di basi, ma c'è anche una versione con parametrizzazione che tiene conto degli effetti di sequenza [37].

Il modello segue un approccio di tipo top-down, dovuto al fatto di cercare di dare una rappresentazione ragionevole delle proprietà strutturali e meccaniche di DNA a filamento singolo e doppio, così come le termodinamiche coinvolte nelle transizioni di ibridazione. Per quanto riguarda i filamenti doppi, la parametrizzazione è stata fatta in modo da replicare le caratteristiche del B-DNA 2.2.2, il quale è la base della maggior parte delle nanotecnologie a DNA per l'ibridazione di filamenti singoli. In generale le interazioni del modello sono state parametrizzate per riprodurre le temperature di melting di oligonucleotidi ottenute con il modello *nearest-neighbour* di SantaLucia [32], che è stato trattato come un fit empirico qualitativo ai dati sperimentali.

I beads del modello semplificano interi nucleotidi, che quindi risultano rigidi e presentano diversi centri d'interazione (figura 2.8a) da cui calcolare i vari potenziali che permettono di ottenere le strutture degli acidi nucleici. I nucleotidi interagiscono tra loro in coppie, e l'energia potenziale del sistema può essere scritta come una somma di diversi termini [30]

$$V = \sum_{n.c.}(V_{schel} + V_{stack} + V'_{escl}) \quad + \sum_{altre\,coppie} (V_{LI} + V_{st.inc.} + V_{escl} + V_{st.coa.}) \quad (2.2)$$

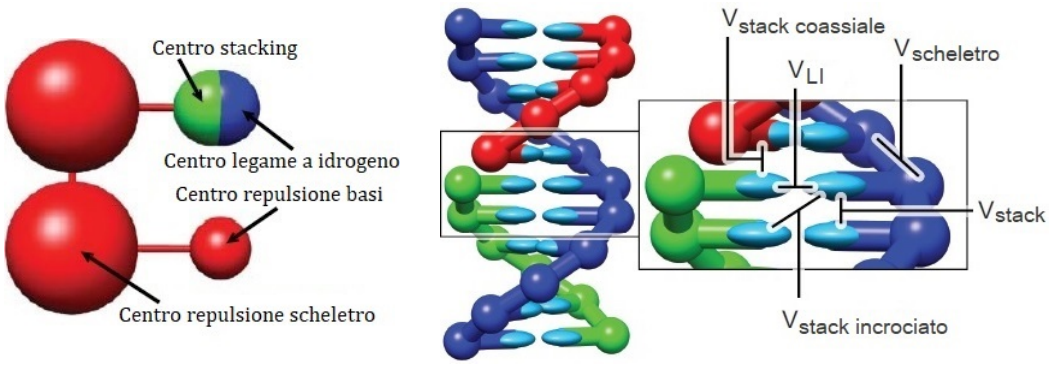
con la prima sommatoria riguardante i nucleotidi contigui su uno stesso filamento e la seconda tutte le altre coppie di nucleotidi. In figura 2.8b è presente una visualizzazione grafica delle interazioni in una doppia elica.

Ognuno dei termini nella (2.2) ha un significato ben preciso che si rifà alle distanze d'interazione in figura 2.9a e viene di seguito spiegato accompagnato da una formula generale, dove i Δ, ϵ, r_0, a, σ rappresentano i parametri.

V_{schel} imita i legami covalenti tra i nucleotidi di un filamento attraverso delle molle non lineari ad estensione limitata(FENE) con distanza di equilibrio pari a 6.4 Å, formando lo scheletro zucchero-fosfato del DNA.

$$V_{FENE}(r, \epsilon, r_0, \Delta) = -\frac{\epsilon}{2} \ln\Big(1 - \frac{(r - r_0)^2}{\Delta^2}\Big) \quad (2.3)$$

V_{stack} imita la tendenza delle basi ad impilarsi in piani paralleli attraverso un potenziale morse troncato gradualmente, in modo da avere il minimo ad una distanza di 3.4 Å tra i piani delle basi. La minor distanza tra le basi



(a) **(b)**

Figura 2.8: (a) Centri d'interazione del modello disposti in linea lungo il nucleotide rigido, con le sfere grandi rappresentanti lo scheletro zucchero-fosfato e quelle più piccole le basi azotate. C'è una leggera separazione tra il centro per lo stacking e quello per il legame a idrogeno. Per chiarezza il centro di repulsione della base azotata è mostrato su un altro nucleotide e la dimensione delle sfere corrisponde alle distanze d'interazione. Immagine da [30]. In (b) è visibile una doppia elica di DNA in cui le basi azotate sono rappresentate come degli ellissoidi nei quali l'asse minore corrisponde alla normale della base. La colorazione degli scheletri identifica diversi filamenti. Nell'ingrandimento un'indicazione schematica delle interazioni. Immagine da [16].
Figure 2.8: (a) The model interaction sites are placed in a line along a rigid nucleotide, the big spheres as the sugar-phosphate backbone and the little ones as the nitrogenous bases. There is a bit of separation between the stacking site and the hydrogen bond one. For clarity the base repulsion site is shown on another nucleotide and the sizes of the spheres correspond to interaction ranges. Image from [30]. A DNA duplex is shown in (b) where the bases are represented as ellipsoids in which the shortest axis corresponds to the base normal. The backbone colorations identify different strands. In the enlargement a schematic representation of the interactions: a backbone potential, the stacking, cross stacking and coaxial stacking potentials and a hydrogen bond potential. Image from [16].

rispetto a quella tra i nucleotidi impostata per V_{schel} lungo lo scheletro causa la formazione dell'elica. Viene modulato attraverso termini angolari calcolati tra le normali delle basi e un vettore che le collega in direzione $3' \rightarrow 5'$. Un'ulteriore modulazione permette di imporre un'elica destrorsa, che si presenta quando determinati angoli sono inferiori a un dato valore, perciò l'interazione diventa progressivamente nulla per angoli più grandi.

$$V_{Morse}(r, \epsilon, r_0, a) = \epsilon(1 - e^{-(r-r_0)a})^2 \tag{2.4}$$

$V_{st.inc.}$ è l'interazione di stacking incrociato che avviene tra una base in un filamento e le basi contigue alla base complementare sul filamento opposto ed è rappresentata attraverso un potenziale armonico troncato gradualmente. Viene modulato tramite l'allineamento dei vettori scheletro-base e delle

normali alle basi con il vettore di separazione.

$$V_{arm}(r, \epsilon, r_0) = \frac{\epsilon}{2}(r - r_0)^2 \tag{2.5}$$

$V_{st.coa.}$ è progettato per essere molto simile allo stacking convenzionale, ma avvenendo su due basi vicine non direttamente connesse attraverso lo scheletro zucchero-fosfato è impossibile definire un asse $3' \to 5'$ e quindi devono esserci delle differenze. Di fatto agisce come uno stacking di basi di nucleotidi non contigui lungo uno stesso filamento[15] ed utilizza un potenziale armonico come in equazione (2.5).

V_{LI} simula i legami a idrogeno che si instaurano nell'accoppiamento delle basi attraverso un potenziale morse troncato gradualmente come in equazione (2.4). È modulato mediante termini angolari ed è impostato come nullo per coppie di basi non complementari. Nel modello medio ha valore equivalente per le coppie di basi, mentre in quello dipendente dalla sequenza la forza dell'interazione, come anche quella di V_{stack}, dipende ulteriormente dall'identità delle basi coinvolte[37].

V'_{escl} **e** V_{escl} sono le interazioni di volume escluso e simulano l'ingombro sterico dei nucleotidi attraverso dei potenziali di Lennard-Jones puramente repulsivi troncati gradualmente tra tutti i siti di repulsione dei due nucleotidi(figura 2.9b). Fanno eccezione le interazioni tra i centri dello scheletro di due nucleotidi contigui perché la loro distanza è già regolata dalla V_{schel}.

$$V_{LJ}(r, \epsilon, \sigma) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{2.6}$$

Come detto, i vari termini di stacking e di legame a idrogeno vengono modulati attraverso termini angolari secondo l'equazione

$$V_{mod}(\theta, a, \theta_0) = 1 - a(\theta - \theta_0)^2 \tag{2.7}$$

dove gli angoli sono di volta in volta specifici per l'interazione da modulare.

Il troncamento di tutti i potenziali di non legame è stato effettuato attraverso combinazione con una funzione di "lisciamento" quadratica

$$V_{lisc}(r, b, x_c) = b(x_c - x)^2 \tag{2.8}$$

dove b rappresenta un parametro e x_c la distanza di cut-off al di fuori della quale il potenziale è nullo. Questo tipo di troncamento rende i vari potenziali continui e differenziabili, permettendo la simulazione non solo con metodi Monte-Carlo, ma anche con altri metodi come la dinamica Langevin, che richiedono delle forze per poter essere utilizzati. Le distanze di cut-off utilizzate nel troncamento

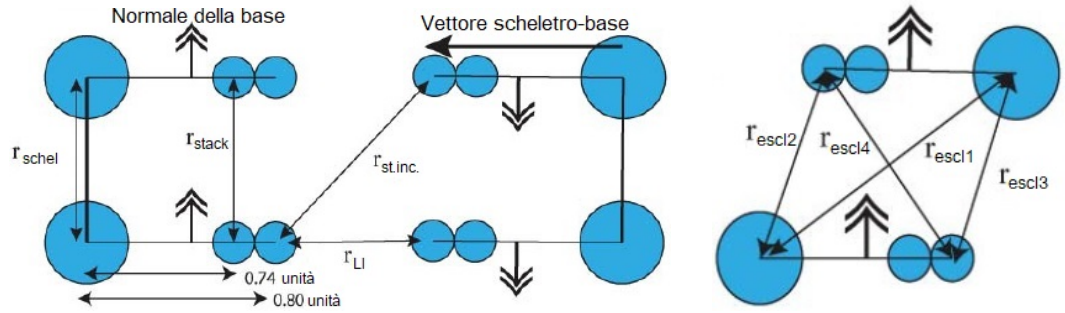


Figura 2.9: (a) Vettori utilizzati nel calcolo dei diversi potenziali dell'equazione (2.2), le distanze riportate sono espresse in unità di simulazione, dove 1 unità = 0.8518 nm. In (b) i vettori per il calcolo di V_{escl} non presenti in (a), dove interagiscono tra loro tutti i centri di repulsione. Immagini da [30].

Figure 2.9: (a) Vectors used for all the potentials in equation (2.2), the lenghts are reported in simulation units, where 1 unit = 0.8518 nm. The normal vector of the base and the backbone-base vector are shown too. In (b) are shown the vectors not present in (a) to calculate V_{escl} for the excluded volume interactions that represent the space occupied by nucleotides, where all repulsion sites interact with each other. Both images from [30].

permettono di non calcolare tutte le interazioni ad ogni passo di simulazione perché avvengono tutte a brevi distanze, rendendo quindi il modello più efficiente dal punto di vista dei costi computazionali. Si rimanda a 3.1.5 per ulteriori dettagli sull'uso dei cut-off.

È importante ripetere che l'elicità dei doppi filamenti non è data da qualche potenziale introdotto ad arte per indurre una torsione, ma deriva direttamente dalla differenza tra la distanza d'equilibrio tra i centri dello scheletro e la separazione delle basi derivante dallo stacking. Le eliche di DNA, però, sono rappresentate in modo semplificato, infatti sono simmetriche e non presentano solco maggiore e solco minore, inoltre tutti i nucleotidi hanno la stessa forma e dimensione; questo significa che il modello non è adatto a studiare fenomeni che dipendono dall'esistenza dei due solchi.

Un'altra semplificazione, nonostante nei filamenti singoli le basi abbiano una carica, è l'assenza di interazioni elettrostatiche esplicite, data da una parametrizzazione eseguita a $[Na^+]$=0.5 M, dove si ha una forte schermatura delle proprietà elettrostatiche; il modello non funziona bene per basse concentrazioni saline(<0.1 M). L'assenza di interazione per legame a idrogeno tra basi non complementari, impostando $V_{LI} = 0$, è anch'essa una semplificazione. In realtà esiste una nuova parametrizzazione del modello, chiamata oxDNA2, in cui sono state corrette entrambe le semplificazioni sui solchi e sulla concentrazione di sali, ora regolabile, ma con tempi di simulazione più lunghi a parità di condizioni.

Infine, ad eccezione delle interazioni di volume escluso, non sono state implementate altre restrizioni sulla conformazione dei nucleotidi contigui. Di fatto le basi hanno la possibilità di ruotare e piegarsi attorno allo scheletro e la mancanza di determinate geometrie specifiche può permettere alcune conformazioni che sarebbero altrimenti escluse in modelli più dettagliati. Questo difetto viene più che compensato dalla possibilità di simulare filamenti singoli che replichino la flessibilità di quelli reali, che era uno degli obbiettivi principali del modello. La possibilità di avere filamenti singoli non impilati permette loro di essere abbastanza flessibili da poter replicare nanostrutture in cui le regioni a filamento singolo sono importanti, come per esempio gli hairpin.

Per concludere, il modello è in grado di restituire una struttura approssimata del B-DNA, avere valide transizioni di stacking e riprodurre molto bene le temperature di melting per le doppie eliche, inoltre riesce a descrivere la minor rigidità di un filamento singolo senza stacking rispetto ad uno che lo presenta, mentre i filamenti doppi risultano i più rigidi. Riesce anche a rappresentare la stabilità di diverse strutture come hairpin(compresa la dipendenza dalle dimensioni di corpo e ansa), mismatch e bolle interne[30]. Lo studio efficace di diversi fenomeni non espressamente parametrizzati, come lo spiazzamento di filamenti legato alla lunghezza d'attacco [36], la formazione di cristalli liquidi ad alta densità[13] o il sovrallungamento della doppia elica sotto una forza applicata [31], ha dato una prova della robustezza del modello, in più è possibile usarlo anche in modo predittivo, se le proprietà sono abbastanza generali da reggere le approssimazioni, come nello studio di un camminatore di DNA [16].

Capitolo 3

Metodi

Una volta scelto il modello sul quale lavorare, per trovare proprietà cinetiche è necessario mettere in movimento le particelle e pertanto si utilizzano le equazioni del moto, ma una loro semplice integrazione non sarebbe sufficiente, perché i computer non sono in grado di risolvere integrazioni di equazioni continue e necessitano che queste vengano tradotte attraverso una loro discretizzazione per differenze finite di tempo, in modo da ottenere determinati algoritmi computabili adatti ad avere simulazioni realistiche. Per ottenere le proprietà cinetiche desiderate è stata usata la *dinamica molecolare* in congiunzione con il *forward flux sampling*, una tecnica per migliorare il campionamento di eventi rari. In questo capitolo vengono spiegate queste due metodologie che sono state applicate nel lavoro svolto.

3.1 Dinamica molecolare (MD)

Le simulazioni MD rappresentano uno strumento per effettuare previsioni statistiche al computer e vengono usate per calcolare proprietà dinamiche e di equilibrio di sistemi classici a molti corpi di cui si conoscono alcuni aspetti. Nella MD si segue solo il moto delle varie particelle attraverso le leggi della meccanica classica, rappresentando una buona approssimazione per molti materiali e sistemi, ma i MD non sono adatti nelle situazioni in cui si debba tener conto del moto degli elettroni e degli effetti quantistici, come per i modelli MM/FA nella sezione 2.1.2.

I moti delle particelle sono importanti per spiegare il collegamento di sequenze con strutture e funzioni e mentre una cristallografia a raggi x di una biomolecola contiene molte informazioni, ne costituisce una visione statica e non è sufficiente per comprendere vari aspetti della sua attività biologica. Le simulazioni MD hanno una corrispondenza agli esperimenti reali, nel senso che in entrambi i casi si ha in un certo senso la preparazione del campione con la scelta di un modello adeguato per le N particelle e la misurazione dopo aver equilibrato il sistema

per mezzo della risoluzione delle equazioni del moto. Tempi di misurazione più lunghi mediano le deviazioni statistiche e permettono di avere misure accurate.

3.1.1 Inizializzazione del sistema

Prima di poter risolvere le equazioni del moto per una simulazione MD, il sistema deve essere inizializzato attraverso la definizione di diversi parametri.

Prima di tutto devono essere definite le *coordinate iniziali* e le *velocità iniziali* delle particelle, avendo cura in ogni caso che non si verifichino sovrapposizioni indesiderate tra le molecole o i nuclei degli atomi. Le posizioni delle particelle possono essere scelte in modi più o meno casuali, ma si deve cercare di assegnare posizioni relative compatibili con le strutture oggetto di studio. Le velocità v_i vengono in seguito assegnate in modo pseudo-casuale ad ogni particella di massa m_i e spesso ad ogni vettore viene sottratta la velocità v_{cm} del centro di massa

$$v_{cm} = \frac{\sum_{i=1}^{N} m_i v_i}{\sum_{i=1}^{N} m_i} \tag{3.1}$$

in modo da portare la quantità di moto iniziale a zero. All'equilibrio termico, per il teorema di equipartizione dell'energia, è inoltre noto che a ogni grado di libertà corrisponde in media un'energia pari a $k_B T/2$, in particolare l'energia cinetica del sistema E_k è legata alla sua temperatura T tramite la relazione $\langle v_\alpha^2 \rangle = k_B T/m$, per cui

$$E_k = \frac{1}{2} \langle \sum_{i=1}^{3N} m_i v_i^2 \rangle = \frac{1}{2} N_L k_B T \tag{3.2}$$

con k_B la costante di Boltzmann e considerando N_L il numero di gradi di libertà del sistema (3N). Questo rende possibile scalare le velocità iniziali delle particelle del sistema in modo da avere la temperatura iniziale desiderata, oltre a poter calcolare la temperatura istantanea nel corso della simulazione.

Impostato tutto, nelle prime fasi di simulazione si assiste a un *tempo di equilibratura* in cui si hanno degli scambi tra energia potenziale e cinetica. Gli scambi proseguono fino a raggiungere una fase di fluttuazione attorno a un valore medio, con numero di particelle, volume e energia totale costanti (NVE) per un insieme microcanonico. Il problema del sistema NVE è che risulta difficile da replicare a livello sperimentale, perciò in genere vengono utilizzati sistemi NVT, in cui si tiene bloccata la temperatura per mezzo di un termostato, per il quale esistono diversi metodi.

3.1.2 Calcolo delle forze

Il calcolo delle forze agenti su ogni singola particella rappresenta il collo di bottiglia della maggior parte delle simulazioni MD. Se si considerano interazioni

additive a coppie, è necessario per ogni particella considerare i contributi alla forza agente su di essa da tutte le particelle vicine, ma questo significa che per N particelle ci sono $N(N-1)/2$ coppie e perciò il tempo necessario a calcolare tutte le distanze, e quindi tutte le forze, scala come N^2. Esistono delle tecniche per rendere più efficienti i calcoli e permettere ai tempi di scalare linearmente con N.

Nel caso del modello oxDNA descritto al paragrafo 2.3, per esempio, una sua descrizione si ha da un hamiltoniano

$$\mathcal{H}(r^{3N}, p^{3N}, q^{3N}, L^{3N}) \tag{3.3}$$

dove r e q sono le coordinate di posizione e orientamento, mentre p e L sono i momenti lineari e angolari. Con esso è possibile calcolare l'energia totale e tutte le proprietà del sistema, che però dipendono esclusivamente dal potenziale. Tenendo conto solo dei termini lineari, si consideri che dalle equazioni hamiltoniane del moto si ha che

$$\frac{\partial \mathcal{H}}{\partial r_i} = -\dot{p_i} \qquad , \qquad \frac{\partial \mathcal{H}}{\partial p_i} = \dot{r_i} \tag{3.4}$$

e che in un sistema chiuso l'energia totale si conserva ed è la somma delle energie potenziali (V) e cinetica (K)

$$H = K + V = \sum_{i=1}^{N} \frac{p_i}{2m_i} + \mathcal{V}(r^{3N}). \tag{3.5}$$

Considerando che le forze F_i agenti sulle particelle sono pari a $F_i = m_i a_i = \dot{p_i}$, si può osservare che questo valore dipende unicamente dal potenziale, unico addendo a dipendere dalle coordinate delle particelle. Conoscendo la formula del potenziale è pertanto possibile calcolare

$$F_i = -\nabla_{\vec{r}_i} \mathcal{V}(r^{3N}) \tag{3.6}$$

e quindi anche il moto attraverso opportuni algoritmi come descritti nei paragrafi 3.1.3 e 3.1.4. Nel paragrafo 3.1.5 si affrontano le tecniche di troncamento dei potenziali allo scopo di ridurre il numero di interazioni da calcolare e abbassare i tempi di calcolo.

3.1.3 L'algoritmo Verlet

Una volta calcolate le forze tra tutte le particelle è necessario integrare le equazioni del moto. È molto importante che le dinamiche siano reversibili nel tempo e uno degli algoritmi di più successo è l'algoritmo Verlet, che risulta spesso uno dei migliori nonostante la sua semplicità. Nella simulazione dei moti

delle biomolecole attraverso campi di forze approssimati, è la stabilità a limitare il passo di tempo utilizzabile, piuttosto che la precisione di simulazione. Il Verlet ha una stabilità molto elevata su tempi lunghi e la sua semplicità permette di applicare delle variazioni al metodo base.

La derivazione dell'algoritmo originale [40] parte da un'espansione di Taylor delle coordinate particellari attorno al tempo t per ottenere

$$r(t+\Delta t) = r(t) + \dot{r}(t)\Delta t + \frac{\ddot{r}(t)}{2}\Delta t^2 + \frac{\dddot{r}(t)}{3!}\Delta t^3 + \mathcal{O}(\Delta t^4), \tag{3.7}$$

ma le derivate nel tempo delle posizioni

$$\dot{r}(t) = v(t) \qquad , \qquad \ddot{r}(t) = a(t) = \frac{F(t)}{m} \tag{3.8}$$

corrispondono alla velocità $v(t)$ e all'accelerazione $a(t)$ della particella. Al posto di quest'ultima si utilizzano la forza $F(t)$ agente su di essa come ricavata dalla (3.6) e la massa m, mentre la derivata terza non è sostituibile con altri valori, quindi l'espansione si esprime come

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 + \frac{\dddot{r}(t)}{6}\Delta t^3 + \mathcal{O}(\Delta r^4). \tag{3.9}$$

Espandendo analogamente per $r(t-\Delta t)$ è poi possibile sommare le due equazioni e ottenere

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \frac{F(t)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4) \tag{3.10}$$

$$r(t+\Delta t) \approx 2r(t) - r(t-\Delta t) + \frac{F(t)}{m}\Delta t^2 \tag{3.11}$$

per la nuova posizione della particella dopo un passo di tempo Δt, con un errore nelle traiettorie al quarto ordine in Δt. Numericamente significa che scegliendo un Δt ragionevolmente piccolo l'errore sulla posizione risulterebbe del tutto trascurabile, conferendo grande stabilità all'algoritmo.

Si noti che la (3.10) presuppone un'iterazione per ogni passo della simulazione, con uno schema di propagazione in due passi, dovuto alla necessità di utilizzare le posizioni ai tempi t e $t-\Delta t$ per ottenere quella al passo successivo. Di fatto il calcolo non comporta l'uso delle velocità, ma per il primo passo di integrazione è necessario utilizzare posizioni precedenti fittizie calcolate mediante le velocità assegnate in fase di inizializzazione. Per il resto della simulazione è sempre possibile ricavare le velocità delle particelle, per esempio per ottenere l'energia cinetica o la temperatura istantanea del sistema, dalla conoscenza delle traiettorie

$$r(t+\Delta t) - r(t-\Delta t) = 2v(t)\Delta t + \mathcal{O}(\Delta t^3) \tag{3.12}$$

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2) \tag{3.13}$$

con un errore del secondo ordine rispetto a Δt. Il metodo risulta quindi complessivamente accurato fino al secondo ordine (non è possibile usare passi particolarmente lunghi), ma esistono algoritmi simili con traiettorie identiche e stime migliori per le velocità.

3.1.4 Dinamica Langevin (LD)

Per evitare una rappresentazione esplicita del solvente nella simulazione, si può ricorrere alla dinamica Langevin. Questo metodo permette di lavorare in un sistema canonico (NVT), per cui necessita dell'uso di un termostato. Le forze agenti sulle particelle di fatto non corrispondono più all'equazione (3.6), ma si ha l'aggiunta di forze di frizione e forze casuali alle equazioni del moto. Così facendo si rappresenta un bagno termico approssimando la frizione per trascinamento del soluto e i colpi casuali dovuti ai moti termici delle molecole di solvente, rendendola di fatto un'equazione differenziale stocastica. Questo tipo di calcolo permette facilmente di generalizzare l'algoritmo Verlet per includere i termini sopra citati. Nel caso dell'oxDNA l'hamiltoniano presenta termini sia lineari che angolari, ma qui per semplicità si espone la spiegazione per i soli termini lineari, per lo smorzamento rotazionale la logica è analoga attraverso l'uso di quaternioni, come descritto nell'algoritmo originale [12].

La forma più semplice per l'equazione Langevin è data da

$$F_i(t) = -\nabla_{r_i} V(r_i(t)) - \gamma m_i v_i(t) + bR_i(t) \tag{3.14}$$

dove γ è il coefficiente di smorzamento lineare misurato come il reciproco di un tempo, $R_i(t)$ è un vettore tridimensionale rappresentante una forza casuale con distribuzione gaussiana per ogni componente[33], tale che

$$\langle R(t) \rangle = 0 \qquad , \qquad \langle R(t) \cdot R(t') \rangle = 2\gamma k_B T m \delta(t - t'), \tag{3.15}$$

b è uno scalare che regola la dimensione delle forze casuali e $\delta(t - t')$ è la delta di Dirac. Se necessario b può essere impostata anche come una matrice, con tutte le correzioni che ne conseguono. La 3.15 significa che le $R_i(t)$ non devono influenzare il sistema nel suo complesso e che non sono temporalmente correlate. Per semplicità si assume che b sia uguale per tutte le particelle, ma in caso di sistemi con tipi di particelle aventi differenze rilevanti è possibile assegnare un valore diverso in base alla particella esaminata. In particolare si può stabilire una relazione tra γ e b [12] secondo la

$$\gamma = \frac{\beta b^2}{2m} \qquad , \qquad \beta = (k_B T)^{-1} \tag{3.16}$$

Si noti che ponendo γ e quindi R pari a zero, l'algoritmo si ridurrebbe nuovamente al Verlet, mentre la delta di Dirac nell'equazione (3.15) viene

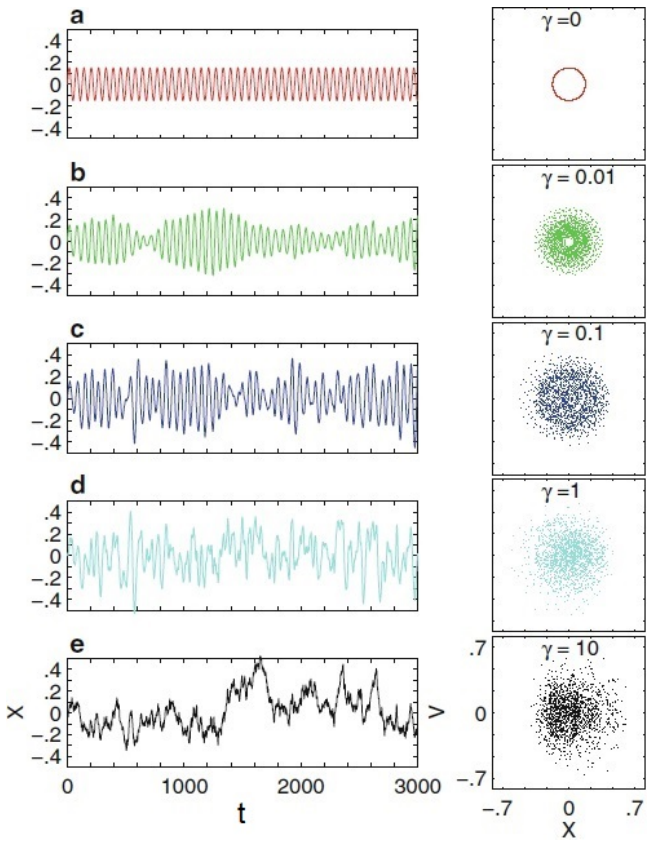


Figura 3.1: Traiettorie Langevin simulate a diversi γ attraverso un'estensione del Verlet con LD. Il sistema rappresenta un oscillatore armonico con frequenza angolare e massa unitari, con passi di 0.1. Il moto armonico ben riconoscibile a viscosità nulla varia man mano che che il contributo relativo alle forze casuali esterne diventa sempre più importante rispetto a quelle interne al sistema. A sinistra i grafici posizione-tempo e a destra i diagrammi dello spazio delle fasi. Immagine da [33].

Figure 3.1: Langevin tajectories simulated at different γ with a Verlet etension to LD. The system represents an harmonic oscillator with unitary angular frequency and mass, with a timestep of 0.1. The harmonic motion well recognizable at zero viscosity changes as the contribution from the external random forces becomes more important relatively to the internal ones of the system. To the left are position-time plots and to the right phase-spase diagrams. Image from [33].

sostituita da $\delta_{ij}/\Delta t$ nella sua forma discreta. Il parametro γ e il suo analogo rotazionale fisicamente rappresentano la forza d'accoppiamento del sistema al bagno termico. Dall'immagine 3.1 si può osservare come con l'aumentare di γ si passi da un regime inerziale ad uno diffusivo per un oscillatore armonico.

Dato che che le collisioni tra le molecole di solvente e quelle del soluto sono simulate attraverso forze stocastiche, si ha la mancanza di interazioni idrodinamiche, per le quali il moto di una molecola di soluto può influenzare quello di altre vicine per mezzo dell'influenza che ha sul soluto, ma poiché la

trattazione di questi effetti consuma molte risorse computazionali e produce risultati simili, viene di solito ignorata. È possibile calcolare un valore fisico di γ per ogni particella dalla legge di Stokes

$$\gamma = 6\pi\eta r_0/m \tag{3.17}$$

dove η è la viscosità del solvente e r_0 rappresenta il raggio idrodinamico delle particelle, che per semplicità possono anche essere considerate tutte come sfere. È anche possibile scegliere γ in modo da cercare di riprodurre costanti di diffusione D osservate sperimentalmente tramite la relazione

$$D = \frac{k_B T}{m\gamma}. \tag{3.18}$$

Dato che le scelte da fare nell'applicazione di un metodo LD risultano essere più di natura qualitativa che quantitativa è possibile comparare cinetiche di sistemi differenti, anche se è sempre necessario interpretare con cautela tutti i risultati. In generale è meglio riuscire a trovare una causa per eventuali differenze e comparare tra loro processi simili.

3.1.5 Tecniche di troncamento

Queste tecniche vengono utilizzate per interazioni a corto raggio, ovvero tutte le interazioni di una determinata particella con tutte le altre particelle ad una distanza inferiore ad un valore di cut-off r_c, che quindi viene definito anche cut-off sferico. Questo nasce dalla necessità di dover simulare sistemi molto grandi e quindi di trovare approssimazioni ragionevoli alle funzioni di energie e forze per ridurre i tempi di calcolo. La scelta della r_c non è scontata, infatti si deve considerare che nonostante il valore dell'energia potenziale diminuisca con l'aumentare della distanza r tra le particelle, il numero dei vicini aumenta molto rapidamente come r^{d-1} [19, capitolo 3], dove d rappresenta il numero di dimensioni del sistema. Questo significa che scegliendo r_c sufficientemente grande, si può rendere l'errore dovuto all'ignorare le interazioni con particelle più lontane arbitrariamente piccolo.

Nello sviluppo di tecniche di troncamento, ci sono alcuni accorgimenti che costituiscono delle linee guida per ottenere dei buoni risultati di simulazione:

- le energie al di sotto di r_c, e di conseguenza le forze, dovrebbero essere alterati il meno possibile;
- si dovrebbero evitare cambiamenti improvvisi dei potenziali, come portarli improvvisamente a 0, in favore di alterazioni graduali onde evitare di generare forze impulsive elevate in corrispondenza della distanza di cut-off, che non vengono gestite bene dagli algoritmi MD. Come conseguenza si comprometterebbe la conservazione dell'energia del sistema;

- il troncamento di un generico potenziale può essere eseguito solo se

$$V(r) \propto \frac{1}{r^{3+n}} \qquad\qquad n > 0$$

  e quindi per esempio il potenziale elettrostatico non può essere troncato, perché non decresce abbastanza rapidamente con la distanza rispetto all'aumento del numero dei vicini.

Poniamo una formula generale in cui l'energia potenziale tra due particelle $V(r)$ viene modificata attraverso una funzione di troncamento $T(r)$ per ottenere il potenziale troncato

$$V_T(r) = T(r)V(r) \tag{3.19}$$

in modo da definire di volta in volta $T(r)$ in base a diversi schemi di troncamento. Si distinguono tre principali trattamenti di cut-off.

Troncamento semplice Come intuibile esclude tutte le interazioni oltre r_c e mantiene inalterati i valori energetici a distanze inferiori, quindi

$$T(r) = \begin{cases} 1 & r < r_c \\ 0 & r \geq r_c \end{cases} \tag{3.20}$$

che però introduce un punto di discontinuità in corrispondenza di $r = r_c$. Come conseguenza il calcolo delle forze per un approccio di dinamica molecolare crea problemi e non si ha conservazione dell'energia. Tuttavia può essere utilizzato con metodi Monte-Carlo, a patto di apportare una correzione per il contributo impulsivo alla pressione, qualora la si volesse misurare, dovuto alla variazione discontinua del potenziale.

Troncamento graduale Il potenziale non viene alterato in un punto, ma in modo più dolce e graduale per mezzo di un polinomio lungo un'intervallo di distanze che termina con r_c. Pur alterando il potenziale in maniera inferiore c'è comunque il rischio di introdurre picchi improvvisi di forze e punti di minimo artificiali. Ponendo che l'intervallo sia $[r_i, r_c]$ con $r_i > 0$, un esempio può essere

$$T(r) = \begin{cases} 1 & r < r_i \\ 1 + y(r)^2[2y(r) - 3] & r_i \leq r \leq r_c \\ 0 & r > r_c \end{cases}$$

$$y(r) = \frac{r^2 - r_i^2}{r_c^2 - r_i^2}.$$

Si noti che il polinomio deve avere un grado abbastanza elevato da assicurare la continuità sia della funzione del potenziale che della sua derivata.

Nel caso dell'esempio si ha una funzione monotona decrescente in cui si passa da $T(r_i) = 1$ a $T(r_c) = 0$. Anche se si considera la sua derivata

$$T'(r) = 12ry(r)\frac{y(r) - 1}{r_c^2 - r_i^2} \quad r_i \leq r \leq r_c \tag{3.21}$$

si può osservare che essa è pari a zero sia da destra che da sinistra per entrambi gli estremi dell'intervallo, perciò si ha la continuità sia della funzione del potenziale che della sua derivata, assicurando dunque che le forze siano continue.

Troncamento con shift Questo metodo altera il potenziale lungo tutto $r < r_c$ allo scopo di evitare i picchi di forze. È una procedura di uso comune in cui si cerca annullare il potenziale in corrispondenza del raggio di cut-off. Si consideri per esempio

$$T(r) = \begin{cases} [1 - (r/r_c)^2]^2 & r \leq r_c \\ 0 & r > r_c \end{cases} \tag{3.22}$$

oppure più semplicemente

$$V_T(r) = \begin{cases} V(r) - V(r_c) & r \leq r_c \\ 0 & r > r_c \end{cases} \quad . \tag{3.23}$$

Così facendo si evitano discontinuità nei potenziali, nonostante si sottovalutino tutte le forze in gioco. Si deve prestare attenzione nel caso di interazioni anisotrope, dove il troncamento deve essere impostato in un punto in cui il potenziale di coppia ha un valore costante, pena non riuscire a shiftare il potenziale a zero in quel punto e avere variazioni di energia, a meno di considerare esplicitamente le forze impulsive generate dal troncamento.

Naturalmente i valori delle proprietà di sistema ottenuti sono diversi in base al metodo di troncamento scelto, ognuno con particolari caratteristiche che possono renderlo più o meno adeguato ad un determinato sistema oggetto di studio.

3.2 Forward Flux Sampling (FFS)

Il FFS è una tecnica che permette di campionare transizioni di eventi rari con dinamiche di tipo stocastico, in cui la semplice simulazione del sistema non permetterebbe di ottenere risultati in tempi accettabili[15, mat. supplementare][4]. Se si definiscono due minimi locali di energia libera A e B, il FFS permette di calcolare il flusso delle traiettorie che vanno da A a B attraverso l'uso di una

serie di interfacce "virtuali" intermedie tra i due stati. Una definizione generale del flusso da A a B è data dalla formula

$$\Phi_{AB} = \frac{N_{AB}}{\tau f_A} \tag{3.24}$$

dove N_{AB} è il numero di volte in cui la simulazione lascia A e raggiunge B, τ è il tempo totale di simulazione e f_A è la frazione di tempo simulato in cui A è stato visitato più di recente rispetto a B.

Definendo una singola interfaccia con il parametro λ si hanno una serie di interfacce $\{\lambda_0, \lambda_1 \dots \lambda_n\}$, per cui lo stato A è dato da $\lambda_A = \lambda_0$ e lo stato B da $\lambda_B = \lambda_n$. Il processo attraverso il quale FFS campiona le traiettorie è illustrato in figura 3.2a. Dalla simulazione iniziale che permette di raggiungere λ_0 e salvarne le configurazioni, si ottiene una stima del flusso di traiettorie che l'attraversano $\phi(\lambda_0)$. Nelle fasi successive si hanno una serie di prove che partono da configurazioni di λ_{i-1} per giungere in caso di successo a λ_i e salvare nuove configurazioni. Ponendo che il numero di prove per quella transizione sia P_i e il numero di successi sia S_i, si può dire che la probabilità di riuscire a raggiungere λ_i prima di tornare ad A sia

$$p(\lambda_i|\lambda_{i-1}) = \frac{S_i}{P_i} \qquad \text{per } i \geq 1 \tag{3.25}$$

con errore calcolato come da sezione 4.1.4 e successiva. Una volta ottenute le probabilità per le transizioni tra tutte le interfacce, attraverso il loro prodotto

$$p(\lambda_n|\lambda_0) = \prod_{i=1}^{n} p(\lambda_i|\lambda_{i-1}) \tag{3.26}$$

si ottiene la probabilità totale che una traiettoria possa raggiungere B prima di tornare ad A. Di conseguenza il flusso totale delle traiettorie tra A e B viene stimato come

$$\Phi_{AB} = \phi(\lambda_0) p(\lambda_n|\lambda_0) \tag{3.27}$$

e il cui errore è calcolabile come da sezione 4.3.3. Il FFS produce delle traiettorie ramificate come visibile in figura 3.2b.

Questo metodo genera ogni volta nuovi percorsi di transizione da zero, perciò ha il vantaggio di generare campioni non correlati tra loro, permettendo un'esplorazione più ampia dei percorsi. Tecniche di campionamento di questo tipo vengono dette statiche. Uno svantaggio che hanno rispetto a quelle dinamiche, in cui i nuovi percorsi di transizione vengono generati modificandone di già esistenti, è rappresentato dal campionamento ripetuto di tante traiettorie che risultano essere dei rami morti, con spreco di tempo.

Un'altra caratteristica di questo metodo è rappresentata dal fatto di far partire le simulazioni di prova dalla fine di una prova precedente, preservando

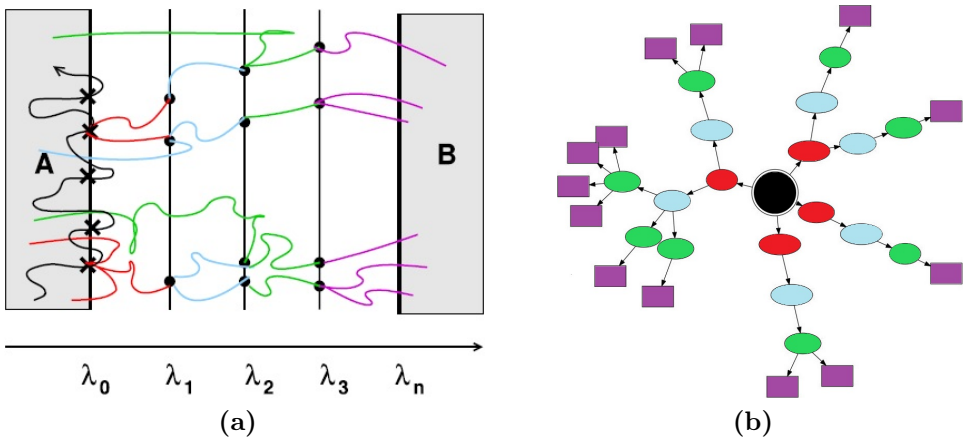


Figura 3.2: (a) Schema di funzionamento del FFS. Si esegue una simulazione in A e quando qualche fluttuazione del sistema fa raggiungere l'interfaccia iniziale λ_0 viene salvata una configurazione fino a raggiungere un numero N_0 scelto e ottenere una stima del flusso per tale interfaccia. In seguito vengono eseguite delle simulazioni di prova su delle configurazioni precedenti scelte casualmente. Esse proseguono fino a quando tornano in λ_0(fallimento), con scelta di un'altra configurazione casuale da λ_0 per un'altra prova, o raggiungono λ_1(successo), con conseguente salvataggio della configurazione, fino a raggiungerne un numero scelto N_1. Il processo viene ripetuto utilizzando ogni volta le nuove configurazioni ottenute per raggiungere l'interfaccia successiva fino al raggiungimento dello stato B con λ_n. Immagine da [4]. (b) Rappresentazione della ramificazione delle traiettorie risultante dal FFS per una sola configurazione di partenza da λ_0, rappresentata dal cerchio centrale. Ogni freccia indica il passaggio a una configurazione dell'interfaccia successiva e le traiettorie che non giungono a λ_n (rami morti) non sono rappresentate.
Figure 3.2: (a) Operating scheme for FFS. A simulation in state A is carried out and when the initial interface λ_0 is reached because of fluctuations in the system, a configuration is saved until there is a chosen number N_0 of them, with a concurrent estimate of the flux. Next, other runs are carried out starting from one of the stored configurations chosen at random. They continue until they return to λ_0(failure), then another trial run starts from another random configuration, or they reach λ_1(success), storing the configuration, until reaching a chosen number N_1 of them. The process is repeated every time using the new configurations to reach the next interface until the B state is reached at λ_n. Image from [4]. (b) Representation of the trajectories ramifications resulting from the FFS for only one starting configuration at λ_0, represented by the central circle. Every arrow indicates the passage to a configuration of the successive interface and trajectories that do not reach λ_n (dead branches) are not represented.

una corretta dinamica di sistema lungo il percorso di transizione. In genere come punti di inizio e fine si scelgono A e B come due stati stabili, ma niente vieta di scegliere l'interfaccia finale λ_n in modo arbitrario.

Il FFS ha una buona efficienza computazionale ed è di semplice applicazione,

ma richiede di conservare molte configurazioni per ogni interfaccia ed è necessario salvare la sequenza storica di connessione tra le configurazioni delle varie interfacce per poter ricavare tutti i percorsi di transizione simulati nelle varie prove. In realtà esistono diverse varianti di FFS, tra le quali quella appena descritta è definita come *FFS diretto*. Esistono anche le varianti a *crescita ramificata* e di *tipo Rosenbluth*[4], non approfondite perché non utilizzate in questa tesi.

Capitolo 4

Errori

Qualsiasi tipo di studio richiede il raccoglimento di dati, in modo da poi poter sviluppare dei ragionamenti adeguati su di essi. Per poter giungere a delle interpretazioni sensate però è necessario analizzare i dati in modo adeguato. L'utilizzo di alcune nozioni di statistica permette di avere in mano un potente strumento per fare chiarezza su ciò che si ha a disposizione. In generale la precisione degli strumenti utilizzati limita l'accuratezza delle misure, oppure possono essere i processi oggetto di studio stessi ad avere natura stocastica.

Tali considerazioni portano a dover esprimere i risultati sotto forma di stime accompagnate ad intervalli di errore, necessari per valutare la precisione di un esperimento ed eventualmente poterla migliorare. L'assegnazione degli errori è importante anche perché permette di comparare le evidenze sperimentali a quanto previsto da teorie quantitative o a risultati di altri esperimenti diversi.

In questo capitolo si descriveranno gli strumenti statistici che sono stati utilizzati nell'analisi dei dati ricavati dal lavoro descritto nel capitolo 5.

4.1 Stima degli errori

4.1.1 Valore medio

Nello studio di una popolazione il primo stimatore usato è la media, per cui si può dire che il valore atteso per tale popolazione, indicato con $\langle x \rangle$, risulta in

$$\langle x \rangle = \mu \tag{4.1}$$

il cui risultato μ può essere definito come *media di popolazione*. Occorre considerare però che nella maggior parte dei casi non è possibile conoscere la popolazione nella sua interezza e μ risulta sconosciuta.

Quel che è possibile fare è cercare di ottenere un campione abbastanza ampio della popolazione e calcolarne la media aritmetica per ricavare la *media*

campione $\overline{x}$, che non coincide generalmente con μ. Questa può essere calcolata con l'equazione:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.2}$$

dove N è la dimensione del campione e x_i sono i singoli valori associati. Questa è una variabile casuale, in quanto lo è il processo di campionamento stesso, e il suo valore atteso è pari ad $\langle x \rangle$

$$\langle \overline{x} \rangle = \frac{1}{N}\sum_{i=1}^{N} \langle x_i \rangle = \langle x \rangle \tag{4.3}$$

4.1.2 Consistenza e distorsione di uno stimatore

Per una statistica finita, in generale si può dire che il valore di uno stimatore $(\hat{\theta})$ differisce da quello del parametro da ottenere (θ) a causa di fluttuazioni statistiche. Aumentando le dimensioni del campione è possibile ridurre le fluttuazioni e, come facilmente intuibile, per un campione arbitrariamente grande, se

$$\lim_{N \to +\infty} (\hat{\theta}) = \theta \tag{4.4}$$

lo stimatore si può considerare consistente.

Supponendo di conoscere θ a priori e di calcolare il valore atteso di uno stimatore da una serie di dati acquisiti, si definisce la *distorsione* (o bias) b come

$$b = \langle \hat{\theta} \rangle - \theta \tag{4.5}$$

e se questo valore è nullo, uno stimatore si definisce senza distorsione o corretto. L'equazione della media campione (4.2) per esempio è uno stimatore consistente e senza distorsione per il valore medio.

4.1.3 Varianza

Per descrivere in modo sufficiente delle quantità stocastiche è necessario un altro stimatore che indichi quanto i campioni siano sparpagliati attorno al loro valore medio. Per esempio, conoscere il reddito medio di un paese non permette di capire se tutti i suoi cittadini abbiano una data situazione economica o se ci sia un profonda divisione tra ricchi e poveri.

Definendo gli scarti come $d_i = x_i - \overline{x}$, si potrebbe pensare di calcolare la media degli scarti, ottenendo un risultato che per costruzione è sempre nullo, infatti:

$$\langle d_i \rangle = \langle x_i - \overline{x} \rangle = \langle x_i \rangle - \langle \overline{x} \rangle = \langle x \rangle - \langle x \rangle = 0 \tag{4.6}$$

perché gli scarti positivi e negativi si annullano tra loro. Per evitare questo fatto si usa lo *scarto quadratico medio* indicato con σ^2, detto anche *varianza*, e definito come

$$\sigma^2 = \langle (x - \mu)^2 \rangle. \tag{4.7}$$

Può essere molto utile anche la seguente derivazione per l'espressione della varianza:

$$\begin{aligned} \sigma^2 &= \langle (x^2 - 2x\mu + \mu^2) \rangle \\ &= \langle x^2 \rangle - 2\langle x \rangle \mu + \mu^2 \\ &= \langle x^2 \rangle - 2\mu^2 + \mu^2 \\ &= \langle x^2 \rangle - \mu^2 \end{aligned} \tag{4.8}$$

che può essere espressa anche nella seguente maniera

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2. \tag{4.9}$$

Così come per la media, anche la varianza viene stimata a partire da un campione, perciò nel caso in cui la media di popolazione sia nota

$$\sigma_\mu^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{4.10}$$

si ottiene uno stimatore in cui $\langle \sigma_\mu^2 \rangle = \sigma^2$, quindi senza distorsione. Il problema è che in genere μ è sconosciuto e viene sostituito dalla media campione $\overline{x}$, ma il calcolo di questo stimatore

$$\sigma_{\overline{x}}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{4.11}$$

restituisce una stima distorta e tende sottostimare il valore σ^2 della popolazione. Una stima migliore e senza distorsione viene ottenuta applicando una correzione che sarà affrontata nella sezione 4.1.6.

4.1.4 Varianza di una proporzione

Se si ha a che fare con un campione di valori che siano 0 o 1, si può affermare che la sua media campione sia una proporzione. Per esempio, nei risultati ottenuti dalle interfacce del FFS spiegato in sezione 3.2 si avranno un insieme di s successi pari a 1 e $N - s$ tentativi falliti pari a 0, perciò dalla (4.2) si ricava che

$$\overline{x} = p = \frac{s}{N} \tag{4.12}$$

$$x_i - \overline{x} = \begin{cases} 1 - \frac{s}{N} & \text{per} \quad s \quad \text{osservazioni} \\ 0 - \frac{s}{N} & \text{per} \quad N - s \quad \text{osservazioni} \end{cases} \tag{4.13}$$

Scorporando la sommatoria dalla (4.11) e combinando con le ultime equazioni si ha

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\overline{x})^2 &= s\Big(1-\frac{s}{N}\Big)^2+(N-s)\Big(0-\frac{s}{N}\Big)^2\\ &= s\Big(1-2\frac{s}{N}+\frac{s^2}{N^2}\Big)+(N-s)\frac{s^2}{N^2}\\ &= s-2\frac{s^2}{N}+\frac{s^3}{N^2}+\frac{s^2}{N}-\frac{s^3}{N^2}\\ &= s-\frac{s^2}{N}=Np-Np^2\\ &= N(p-p^2)\end{aligned} \tag{4.14}$$

che reinserito nell'equazione dello stimatore restituisce

$$\sigma^2_{\overline{x}} = p-p^2 \tag{4.15}$$

4.1.5 Errore standard

Per giungere all'errore standard è necessario prima fare un paio di osservazioni. Ponendo

$$\mathsf{Var}(x)=\sigma^2$$

è noto che in generale, per una costante k si ha

$$\mathsf{Var}(kx)=k^2\mathsf{Var}(x)=k^2\sigma^2 \tag{4.16}$$

Ora si ponga di effettuare diversi campionamenti indipendenti su di una popolazione, rappresentati dalle variabili casuali $\overline{x_i}$, con $1\leq i\leq n$, ognuno con varianza σ^2. Essendo valori indipendenti, la varianza della loro somma

$$\mathsf{Var}\Big(\sum_{i=1}^{n}\overline{x_i}\Big)=\sum_{i=1}^{n}\mathsf{Var}(\overline{x_i})=\sum_{i=1}^{n}\sigma^2=n\sigma^2 \tag{4.17}$$

equivale alla somma delle varianze.

A questo punto usando insieme le (4.16) e (4.17) e ponendo $k=\frac{1}{n}$ si ottiene il risultato

$$\mathsf{Var}\Big(\frac{\sum_{i=1}^{n}\overline{x_i}}{n}\Big)=\frac{1}{n^2}\mathsf{Var}\Big(\sum_{i=1}^{n}\overline{x_i}\Big)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n} \tag{4.18}$$

che corrisponde al calcolo della varianza rispetto alla media delle medie dei vari campionamenti indipendenti effettuati. Naturalmente la varianza della popolazione non è generalmente conosciuta, perciò il valore viene sostituito da uno stimatore opportuno, come (4.11) o (4.15), inoltre non sempre è possibile

effettuare n campionamenti indipendenti per applicare la formula, quindi ci si limita a calcolare la varianza del campione e a dividere per la dimensione dello stesso.

L'errore standard (SE) non è altro che la radice quadrata dell'ultimo risultato ottenuto, perciò

$$SE_{\overline{x}} = \frac{\sigma}{\sqrt{N}} \tag{4.19}$$

può essere vista come una stima della deviazione standard dell'errore della media campione rispetto alla reale media di popolazione. Il denominatore si associa anche alla consistenza (4.4), in quanto all'aumentare della dimensione del campione corrisponde una riduzione dell'errore.

4.1.6 Varianza corretta

Si era detto che per l'equazione (4.11) esiste una correzione che permette di ottenere uno stimatore corretto per la varianza. Dallo sviluppo del suo valore atteso

$$\begin{aligned}\langle \sigma^2_{\overline{x}} \rangle &= \langle (x - \overline{x})^2 \rangle \\ &= \langle (x^2 - 2x\overline{x} + \overline{x}^2) \rangle \\ &= \langle x^2 \rangle - \langle \overline{x}^2 \rangle\end{aligned}$$

che integrata con quanto deriva dai risultati delle (4.9) e (4.18)

$$\langle x^2 \rangle = \mathsf{Var}(x) + \langle x \rangle^2 = \sigma^2 + \langle x \rangle^2$$

$$\langle \overline{x}^2 \rangle = \mathsf{Var}(\overline{x}) + \langle \overline{x} \rangle^2 = \frac{\sigma^2}{N} + \langle x \rangle^2$$

si ottiene

$$\langle \sigma^2_{\overline{x}} \rangle = \sigma^2 - \frac{\sigma^2}{N} = \sigma^2 \Big(\frac{N-1}{N} \Big)$$

per cui si evidenzia che $\sigma^2_{\overline{x}}$ restituisce una sottostima del valore σ^2 per un fattore $\frac{N-1}{N}$. Dalla (4.5) si evidenzia anche che la distorsione per questo stimatore corrisponde a

$$b = -\frac{\sigma^2}{N}$$

L'uso di questo fattore correttivo viene detto correzione di Bessel e per distinguere la (4.11) dalla sua variante corretta in genere si incontra un altro tipo di notazione, S^2_N e S^2_{N-1} rispettivamente,dal denominatore delle due

$$\frac{N}{N-1} S^2_N = \frac{N}{N-1} \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 = S^2_{N-1} \tag{4.20}$$

È bene precisare che sebbene quest'ultimo stimatore sia senza distorsione, lo stesso non si può dire per la sua radice, perciò la deviazione standard da esso calcolata risulta comunque essere non corretta.

4.2 Ricampionamento

Con l'uso sempre più esteso dei computer i metodi di ricampionamento rappresentano una strada sempre più percorribile per la stima di parametri ed errori, verifica di ipotesi, stima di distorsioni e convalida di modelli. In quanto utilizzato in questo lavoro di tesi, verrà posta attenzione soprattutto sul ricampionamento jackknife, per poi esporre brevemente un'altra tecnica correlata, il bootstrap.

4.2.1 Jackknife

Il jackknife era stato originariamente proposto da Quenouille per stimare e correggere la distorsione di uno stimatore, poi Tukey gli diede il nome attuale e lo espanse per permettere la costruzione di intervalli di confidenza. Il nome nasceva dal fatto che come un coltello a serramanico, dalla traduzione di jackknife, questa tecnica potesse essere utilizzata come uno strumento "veloce e sporco" in sostituzione di altri strumenti più specifici e sofisticati.

Si ponga di aver ottenuto un campione casuale di dati $x = \{x_1, \ldots, x_i, \ldots, x_N\}$ da N osservazioni e di ottenere la stima per un parametro della popolazione per mezzo di uno stimatore s tale che, in generale

$$s = f(x_1, \ldots, x_N) \tag{4.21}$$

in questo caso si definisce come *replica jackknife* il valore dello stimatore s calcolato sul *campione jackknife*, costruito sul set di dati originale ad esclusione di una osservazione. In tal modo si ottengono N campioni jackknife X_i di dimensioni $N - 1$

$$X_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\} \tag{4.22}$$

e N repliche jackknife s_i

$$s_i = s(X_i). \tag{4.23}$$

Queste considerazioni rendono il jackknife un metodo definito *senza re-immissione*. Da queste basi si definisce la stima dello *pseudo-valore*, calcolato come

$$T_i = Ns - (N-1)s_i \tag{4.24}$$

per cui la stima jackknife x_{jack} per il parametro da trovare è ottenuta come media degli pseudo-valori

$$x_{\mathsf{jack}} = \overline{T} = \frac{1}{N}\sum_{i=1}^{N} T_i \tag{4.25}$$

e quindi la *varianza jackknife* σ^2_{jack} si calcola come da formula senza distorsione (4.20) per gli pseudo-valori

$$\sigma^2_{\mathsf{jack}} = \frac{1}{N-1}\sum_{i=1}^{N}(T_i - \overline{T})^2. \tag{4.26}$$

È interessante notare come nel caso in cui lo stimatore iniziale corrisponda alla media campione, gli pseudo-valori diventano equivalenti ai valori del campione iniziale

$$s = \overline{x} \Rightarrow s_i = \frac{1}{N-1}(N\overline{x} - x_i)$$
$$T_i = N\overline{x} - (N\overline{x} - x_i) = x_i$$

e quindi la (4.26) si riduce alla formula usuale (4.20).

L'idea dietro gli pseudo-valori T_i è quella di trattarli, con buona approssimazione secondo Tukey[39], come N dati indipendenti e perciò si può ottenere l'incertezza sul parametro stimato per mezzo del calcolo dell'errore standard utilizzando la (4.26) nella formula dell'errore standard della media (4.19), ottenendo

$$SE_{\mathsf{jack}} = \frac{\sigma_{\mathsf{jack}}}{\sqrt{N}} = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(T_i - \overline{T})^2}. \tag{4.27}$$

Gli pseudo-valori hanno il problema di non essere efficienti dal punto di vista computazionale, ma questo può essere superato dal fatto che la (4.27) si può esprimere in funzione delle repliche jackknife e della loro media, rendendo il calcolo più diretto ed intuitivo. Se si considerano i termini all'interno delle parentesi e si sviluppano con la (4.24)

$$\begin{aligned} T_i - \overline{T} &= Ns - (N-1)s_i - \frac{1}{N}\sum_{i=1}^{N} T_i \\ &= Ns - (N-1)s_i - \frac{1}{N}\sum_{i=1}^{N}(Ns - (N-1)s_i) \\ &= Ns - (N-1)s_i - Ns + \frac{N-1}{N}\sum_{i=1}^{N} s_i \\ &= (N-1)(\overline{s} - s_i) \end{aligned} \tag{4.28}$$

si porta il risultato al quadrato e si ottiene

$$SE_{\mathsf{jack}} = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(s_i - \overline{s})^2}. \tag{4.29}$$

Quanto descritto fino ad ora si può chiamare jackknife *elimina-1*, ma si potrebbe generalizzare in un approccio *elimina-d*. La dimensione dei campioni jackknife diverrebbe $(N - d)$, mentre il loro numero equivarrebbe al coefficiente

binomiale $\binom{N}{d}$, perciò a un incremento anche solo unitario di d possono corrispondere diversi ordini di grandezza di differenza. In ogni caso la formula per la stima jackknife elimina-d per l'errore standard è simile a quella elimina-1

$$SE_{d-\mathsf{jack}} = \sqrt{\frac{N-d}{d\binom{N}{d}} \sum_i (s_{di} - \overline{s_d})^2} \tag{4.30}$$

dove s_{di} è la stima applicata al campione a cui è stato rimosso un numero d di osservazioni e

$$\overline{s_d} = \frac{1}{\binom{N}{d}} \sum_i s_{di}$$

potendo facilmente verificare che se $d = 1$ ci si riconduce alla (4.27).

Un'assunzione di questo strumento è che le osservazioni siano indipendenti e identicamente distribuite, perciò non è in generale appropriato per le serie temporali, è stato inoltre dimostrato che la stima jackknife della varianza ha una leggera distorsione verso l'alto [18].

Dato che il jackknife elimina le distorsioni per sottrazione, funziona correttamente solo per statistiche che siano funzioni lineari dei parametri e le quali distribuzioni siano continue, perciò la versione elimina-1 tende a restituire stime consistenti per stimatori continui o "sufficientemente lisci" da poter essere considerati tali, ma per stimatori non lisci, come per esempio la stima della mediana o dei quantili in genere, si possono ottenere delle stime inconsistenti. In alcuni casi è possibile linearizzare le statistiche tramite trasformazioni, ma per stimatori non lisci d deve dipendere da una misura di quanto siano lisci, in particolare nel caso dei quantili si è dimostrato che nel caso in cui $\sqrt{N}/d \to 0$ e $N - d \to \infty$ si ottiene una stima consistente e asintoticamente senza distorsione[34]. In breve per stimatori sufficientemente lisci si raccomanda il jackknife elimina-1, perché comunque consistente e più veloce da calcolare, mentre nel caso non siano lisci è preferibile scegliere d tale che $\sqrt{N} < d < N$.

4.2.2 Jackknife come stima della distorsione

Ricordando la definizione generale di distorsione (4.5), la distorsione jackknife si definisce come

$$b_{\mathsf{jack}} = (N-1)(\overline{s} - \hat{\theta}) \tag{4.31}$$

dove $\hat{\theta}$ rappresenta lo stimatore valutato sull'intero campione. Il senso del fattore $N - 1$ deriva alcune considerazioni che partono dello stimatore non corretto per la varianza del campione S_N^2 (4.11) con distorsione $-\sigma^2/N$, mentre l'uso della media campione sarebbe stato inutile, in quanto la sua distorsione è nulla. Se come stima per la distorsione dello stimatore S_N^2 si usa la (4.31), si ottiene che il risultato è $-S_{N-1}^2/N$, al cui numeratore c'è lo stimatore corretto per la

varianza, perciò la stima jackknife della distorsione è costruita con l'intento di emulare la distorsione di S^2.

Il passo successivo è quello di definire la stima jackknife del parametro d'interesse come

$$\hat{\theta}_{\mathsf{jack}} = \hat{\theta} - b_{\mathsf{jack}} = N\hat{\theta} - (N-1)\overline{s} \tag{4.32}$$

che è del tutto equivalente alla media degli pseudo-valori (4.24). In teoria per definizione la distorsione del nuovo stimatore $\hat{\theta}_{\mathsf{jack}}$ dovrebbe essere nulla, ma in pratica non è sempre esattamente così. Il trattamento appena esposto può essere considerato un'approssimazione al primo ordine di un'espansione di Taylor, perciò non elimina del tutto la distorsione di uno stimatore, ma riesce comunque a ridurla.

Si ponga di avere N elementi il cui valore atteso di uno stimatore equivale al parametro da stimare più una serie infinita di termini di distorsione

$$\langle\hat{\theta}\rangle = \theta + \frac{b_1(\theta)}{N} + \frac{b_2(\theta)}{N^2} + \frac{b_3(\theta)}{N^3} + \dots \tag{4.33}$$

considerando che le repliche jackknife hanno $N-1$ termini, il valore atteso per la loro media risulta

$$\langle\overline{s}\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle s_i\rangle = \theta + \frac{b_1(\theta)}{N-1} + \frac{b_2(\theta)}{(N-1)^2} + \frac{b_3(\theta)}{(N-1)^3} + \dots \tag{4.34}$$

per cui, approssimando al primo ordine, si ottiene che il valore atteso per la differenza $\overline{s} - \hat{\theta}$ risulta

$$\langle\overline{s} - \hat{\theta}\rangle = \theta + \frac{b_1(\theta)}{N-1} - \theta - \frac{b_1(\theta)}{N} = \frac{b_1(\theta)}{(N-1)N} \tag{4.35}$$

e quindi moltiplicando questa differenza per $N-1$ si ottiene la (4.31), equivalente alla distorsione dello stimatore originale in questa approssimazione.

Se invece si volessero considerare i termini oltre al primo ordine, il valore atteso della distorsione jackknife diventa

$$\langle b_{\mathsf{jack}}\rangle = (N-1)\langle\overline{s} - \hat{\theta}\rangle = \frac{b_1(\theta)}{N} + \frac{(2N-1)b_2(\theta)}{N^2(N-1)} + \frac{(3N^2-3N+1)b_3(\theta)}{N^3(N-1)^2} + \dots$$

e di conseguenza il valore atteso dello stimatore jackknife è

$$\begin{aligned}\langle\hat{\theta}_{\mathsf{jack}}\rangle = \langle\hat{\theta} - b_{\mathsf{jack}}\rangle &= \theta - \frac{b_2(\theta)}{N(N-1)} + \frac{(2N-1)b_3(\theta)}{N^2(N-1)^2} + \dots \\ &\approx \theta - \frac{b_2(\theta)}{N^2} - \frac{2b_3(\theta)}{N^3} + \dots\end{aligned} \tag{4.36}$$

Quest'ultimo passaggio mostra come, non essendoci il termine di primo ordine perché eliminato, la distorsione per $\hat{\theta}_{\mathsf{jack}}$ sia asintoticamente più piccola rispetto a un qualunque stimatore incorretto.

4.2.3 Bootstrap

Il bootstrap è un metodo successivo al jackknife, il cui successo è in effetti stato l'ispirazione per la sua creazione da parte di Efron[17]. Al contrario del jackknife il bootstrap ricampiona *con re-immissione*. Si supponga di avere un campione di dimensione N con osservazioni indipendenti e identicamente distribuite da una distribuzione sconosciuta, quindi σ^2 non è disponibile. Dalla (4.19) è noto che con l'aumentare delle dimensioni del campione il valore stimato si avvicina a quello di popolazione. Il bootstrap è utile quando in presenza di statistiche a cui il discorso precedente non si applica, oppure di campioni di dimensioni relativamente piccole.

Si ponga di avere il campione $x = \{x_1, x_2, \ldots, x_N\}$, il bootstrap genera B campioni di dimensione N dai dati, ricampionando con re-immissione a partire dal campione x originario. È possibile ottenere un set di N^N campioni diversi in questo modo, e più B è grande più si avvicina a questo numero che viene chiamato il *campione bootstrap ideale*. È facile verificare che per $N = 2$ si ha un numero massimo di campioni bootstrap pari a 4, ma in generale questo numero cresce in modo estremamente veloce e le richieste computazionali diventano incredibilmente pesanti, a meno di avere campioni piuttosto piccoli.

Nel ricampionamento casuale, oltretutto, il risultato più probabile è il campione originario stesso, perciò è quasi inevitabile che tra i campioni bootstrap ci siano delle sue repliche, portando alla conclusione che ottenere il campione ideale è praticamente impossibile. Il numero B sufficiente per ottenere delle stime affidabili è oggetto di discussione e può variare in base alla statistica che si vuole calcolare.

Ponendo di applicare lo stimatore desiderato s al set di campioni, si ottiene un insieme di repliche $\{s_{b1}, s_{b2}, \ldots, s_{bB}\}$, quindi si calcola

$$\overline{s_b} = \frac{1}{B}\sum_{i=1}^{B} s_{bi}$$

e si ottiene la stima bootstrap dell'errore standard come deviazione standard delle repliche bootstrap

$$SE_{\mathsf{boot}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(s_{bi} - \overline{s_b})^2} \tag{4.37}$$

Si dimostra che la funzione di distribuzione empirica usata per generare i campioni bootstrap è uno stimatore consistente e senza distorsione per la funzione di distribuzione cumulativa dalla quale i campioni vengono estratti. Questo metodo funziona bene perché invece di andare a convergenza nell'ordine di $1/\sqrt{N}$, lo fa con un tasso più veloce di $1/N$.

4.3 Propagazione degli errori

Quanto già visto permette di stimare l'errore nei dati ottenuti direttamente, ma nel caso in cui sia necessario manipolare tali dati è indispensabile sapere come calcolare l'incertezza sul risultato finale. In questo paragrafo si descriveranno le basi per il calcolo per la funzione in una variabile, per poi procedere al calcolo multivariabile attraverso una spiegazione della covarianza.

4.3.1 Propagazione come funzione di una variabile

Si ponga di conoscere il valore di una variabile e il suo errore associato sotto la forma $\overline{x} \pm \delta x$ e di voler determinare il valore di $y_c = f(x)$ e del suo errore δy. In modo da non dover necessariamente conoscere la funzione densità di probabilità di $f(x)$, si esegue un'espansione di Taylor di y attorno al valore $\overline{x}$

$$y = y(\overline{x}) + y'(\overline{x})\Delta x + \frac{1}{2!}y''(\overline{x})\Delta x^2 + \mathcal{O}\big(\Delta x^3\big) \tag{4.38}$$

e si ignorano i termini quadratici e di ordine superiore, supponendo che l'errore dx sia abbastanza piccolo da giustificare la linearizzazione della funzione.

A questo punto si effettuano le seguenti uguaglianze:

$$\begin{aligned} y_c &= \langle y(x) \rangle \\ (\delta y)^2 &= \langle (y - y_c)^2 \rangle \\ \Delta x &= x - \overline{x} \end{aligned}$$

Si ricorda che, come visto nella (4.6), il valore atteso di Δx è pari a zero, mentre quello di $(\Delta x)^2$ corrisponde a $(\delta x)^2$(a meno di un fattore nel caso sia stato stimato con la (4.19)). Applicando l'approssimazione di primo ordine si ottiene

$$\langle y(x) \rangle \approx \langle y(\overline{x}) \rangle + \langle y'(\overline{x})\Delta x \rangle = y(\overline{x}) = y_c$$

ovvero che il valore da determinare corrisponde al calcolo della funzione di $\overline{x}$, e poi

$$\begin{aligned} \langle (y - y_c)^2 \rangle &\approx \langle (y(\overline{x}) + y'(\overline{x})\Delta x - y_c)^2 \rangle \\ &= y'^2(\overline{x}) \langle (\Delta x)^2 \rangle \\ &= y'^2(\overline{x})(\delta x)^2 = (\delta y)^2 \\ \delta y &= |y'(\overline{x})| \delta x \end{aligned}$$

che restituisce un risultato già deducibile dalla figura 4.1. In tabella 4.1 son riportati degli esempi per alcune funzioni.

Riassumendo, nell'utilizzo di questo approccio si deve fare attenzione ad alcune cose:

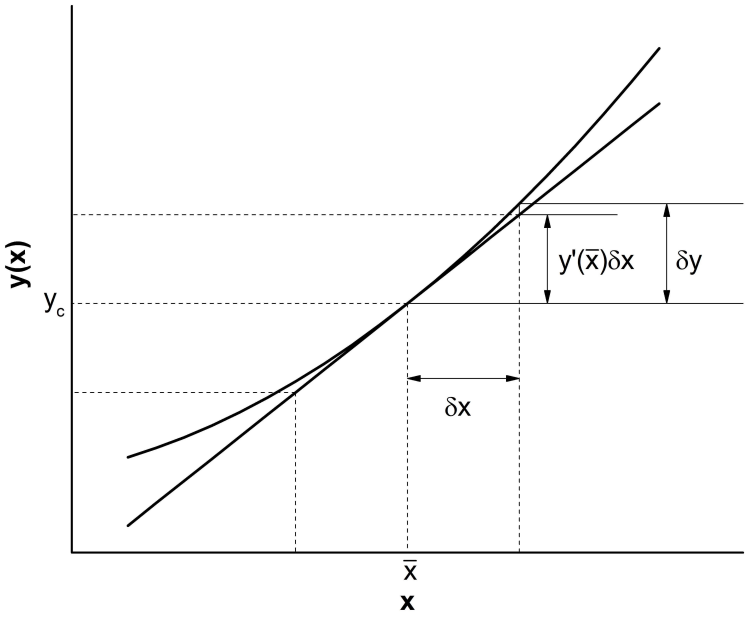


Figura 4.1: La $y(x)$ è assunta non lineare per mantenere la generalità dell'argomento e la tangente rappresenta l'approssimazione di primo ordine. Le linee tratteggiate rappresentano graficamente la conversione dell'errore su x nell'errore su $y(x)$. È evidente come al crescere dell'errore δx la discrepanza tra l'errore nella funzione e quello nell'approssimazione diventi sempre meno trascurabile, ma per valori abbastanza piccoli l'approssimazione risulti adeguata. Anche il grado di non linearità del modello influenza il risultato, infatti meno esso è lineare e meno l'approssimazione è una misura adeguata dell'errore. Per modelli estremamente poco lineari l'approssimazione al primo ordine potrebbe essere del tutto inadeguata.
Figure 4.1: The function $y(x)$ is taken as nonlinear to keep the argument general and the tangent represents the first order approximation. The dashed lines are a graphic representation of the conversion of the error in x into the error in $y(x)$. It is obvious that for increasing values of δx the discrepancy between the error in the function and the error in the approximation becomes less and less negligible, but for sufficiently small values the approximation is adequate. The result is also influenced by the non-linearity degree of the model, the error measurement with such approximation becoming less adequate the less linear is the model. For extremely non-linear models, the first order approximation could be completely unsuitable.

- nella propagazione dell'errore si assume che l'incertezza relativa dei valori sia piccola;

- non si consiglia la propagazione dell'errore se l'incertezza può essere misurata direttamente(cioè come variazione tra esperimenti ripetuti);

- l'incertezza non si riduce mai con i calcoli, solo con dei dati migliori.

Tabella 4.1: Esempi di propagazione lineare dell'errore per funzioni in una sola variabile x, gli altri termini sono delle costanti.

Funzione	Errore propagato
$y = ax^b$	$\delta y = \lvert\frac{by}{x}\rvert\delta x$
$y = a\ln(bx)$	$\delta y = \lvert\frac{a}{x}\rvert\delta x$
$y = ae^{bx}$	$\delta y = \lvert by\rvert\delta x$
$y = \tan x$	$\delta y = \lvert\frac{1}{\cos^2 x}\rvert\delta x$

4.3.2 Covarianza e coefficiente di correlazione

Nel caso in cui si abbia a che fare con più variabili è importante avere una misura della dipendenza tra di esse, in quanto il variare di una può influenzare l'andamento dell'altra. Ponendo di avere dati per due diverse variabili casuali x e y, una prima misura può essere ottenuta con la *covarianza*, definita come

$$\mathsf{Cov(x, y)} = \sigma_{x,y} = \langle (x - \mu_x)(y - \mu_y) \rangle \tag{4.39}$$

ma che solitamente viene stimata, in mancanza delle medie di popolazione, sulla base delle medie campione nel seguente modo

$$\sigma_{x,y} = \frac{1}{N} \sum_{i=1}^{N} (x - \overline{x})(y - \overline{y}) \tag{4.40}$$

Il valore ricavato dalla (4.40) dovrebbe esprimere la correlazione tra le due variabili, ma questa espressione è sensibile all'unità di misura utilizzata, perciò è chiaro che se per esempio si usano Km al posto dei mm, si otterrà un valore molto piccolo, portando potenzialmente a pensare che non ci sia correlazione tra due variabili che invece lo sono. Per evitarlo si utilizza un valore adimensionale, chiamato *coefficiente di correlazione*

$$\begin{aligned} \rho_{x,y} &= \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \\ -1 \leq \rho_{x,y} &\leq +1 \end{aligned} \tag{4.41}$$

per cui si evidenziano tre casi:

$\rho > 0$ Avviene se al variare in un senso di una variabile, si ha lo stesso effetto anche per l'altra e la correlazione di dice *diretta* o *positiva*.

$\rho = 0$ Si ha questo risultato se le due variabili sono indipendenti tra loro, ma non è vero il contrario, una correlazione nulla non implica necessariamente l'indipendenza delle due variabili.

$\rho < 0$ Avviene se al variare di una variabile l'altra varia in senso inverso e la correlazione di dice *inversa* o *negativa*.

Più $|\rho|$ si avvicina a 1 e più la correlazione si dice forte, fino a diventare perfetta. Il coefficiente come calcolato nella (4.41) e viene chiamato *coefficiente di correlazione di Pearson* o *coefficiente di correlazione lineare* perché esprime un'eventuale relazione di linearità tra di esse, che può essere presente anche se una variabile non è funzione lineare dell'altra. È importante evidenziare una proprietà che deriva direttamente dalle definizioni (4.39) e (4.41), cioè

$$\sigma_{x,y} = \sigma_{y,x} \Rightarrow \rho_{x,y} = \rho_{y,x} \tag{4.42}$$

Esistono anche altri coefficienti più sensibili a relazioni di tipo non lineare e che sono quindi più robusti di quello di Pearson[14], ma non verranno trattati.

4.3.3 Errore di una funzione di più variabili

Si consideri per semplicità una funzione in due variabili $y = f(x_1, x_2)$ alle quali sono associate le incertezze δx_1 e δx_2, in seguito si generalizzerà per una funzione generica in più variabili. Un esempio può essere il calcolo dell'area di un rettangolo, dopo aver misurato i due lati e aver assegnato le incertezze alle relative misure, in cui la funzione corrisponde al prodotto delle variabili.

Ponendo i valori misurati per le 2 variabili come x_{1m} e x_{2m} e il valore calcolato per y come y_c, eseguendo l'espansione di Taylor approssimata al primo ordine secondo lo stesso principio della sezione 4.3.1, si ottiene

$$y = y(x_{1m}, x_{2m}) + \frac{\partial y}{\partial x_1}(x_{1m}, x_{2m})\Delta x_1 + \frac{\partial y}{\partial x_2}(x_{1m}, x_{2m})\Delta x_2 + O\big((\Delta x_1)^2, (\Delta x_2)^2\big)$$
$$\langle \Delta x_1 \rangle = \langle \Delta x_2 \rangle = 0$$

da cui

$$y_c = \langle y(x_1, x_2) \rangle = y(x_{1m}, x_{2m}) \tag{4.43}$$

e, evitando di trascrivere ogni volta che le derivate sono calcolate in x_{1m} e x_{2m}

$$\begin{aligned}(\delta y)^2 &= \langle (\Delta y)^2 \rangle \\ &\approx \langle (\frac{\partial y}{\partial x_1}\Delta x_1 + \frac{\partial y}{\partial x_2}\Delta x_2)^2 \rangle \\ &= \Big(\frac{\partial y}{\partial x_1}\Big)^2 (\delta x_1)^2 + \Big(\frac{\partial y}{\partial x_2}\Big)^2 (\delta x_2)^2 + 2\Big(\frac{\partial y}{\partial x_1}\Big)\Big(\frac{\partial y}{\partial x_2}\Big)\langle \Delta x_1 \Delta x_2 \rangle\end{aligned} \tag{4.44}$$

si evidenzia nell'ultimo addendo il termine $\langle \Delta x_1 \Delta x_2 \rangle$, che dalla definizione (4.39) corrisponde alla covarianza per le due variabili. Quest'ultimo termine viene preferibilmente espresso mediante l'uso del coefficiente di correlazione ρ secondo la (4.41) per restituire

$$(\delta y)^2 = \Big(\frac{\partial y}{\partial x_1}\Big)^2 (\delta x_1)^2 + \Big(\frac{\partial y}{\partial x_2}\Big)^2 (\delta x_2)^2 + 2\Big(\frac{\partial y}{\partial x_1}\Big)\Big(\frac{\partial y}{\partial x_2}\Big)\rho_{12}\delta x_1 \delta x_2 \tag{4.45}$$

Come detto questa è una formula approssimata, ma tornando all'esempio iniziale sull'area del rettangolo, o in generale $f(x_1, x_2) = x_1 x_2$, la nostra formula approssimata restituirebbe

$$(\delta y)^2 = x_{2m}^2(\delta x_1)^2 + x_{1m}^2(\delta x_2)^2 + 2\rho_{12}x_{1m}x_{2m}\delta x_1 \delta x_2 \tag{4.46}$$

mentre la formula esatta per un prodotto, ricavata da Goodman nel 1960[20], è la seguente

$$\begin{aligned}\sigma^2_{(\overline{x}\,\overline{y})} = \frac{1}{N}\Big[&\mu_x^2\sigma_y^2 + \mu_y^2\sigma_x^2 + 2\mu_x\mu_y E_{11} \\ &+ 2\mu_x\frac{E_{12}}{N} + 2\mu_y\frac{E_{21}}{N} + \frac{\sigma_x^2\sigma_y^2}{N} + \frac{\mathsf{Cov}[(\Delta x)^2, (\Delta y)^2] - E_{11}^2}{N^2}\Big]\end{aligned} \tag{4.47}$$

dove, avendo mantenuto la notazione originale:

- x e y rappresentano le due variabili di cui si deve eseguire il prodotto;
- $E_{ij} = \langle(\Delta x)^i(\Delta y)^j\rangle$, quindi E_{11} corrisponde alla covarianza $\sigma_{x,y}$;
- $\mathsf{Cov}[(\Delta x)^2, (\Delta y)^2] = E_{22} - \sigma_x^2\sigma_y^2$.

È evidente come la prima riga della (4.47) sia del tutto analoga alla (4.46), infatti se gli errori sulle variabili da moltiplicare sono sufficientemente piccoli, i termini della seconda riga risultano ancora più piccoli e possono essere ignorati.

Generalizzare la (4.46), per una funzione di N variabili $y(x_1, ..., x_N)$ è piuttosto semplice e si ottiene

$$\begin{aligned}(\delta y)^2 &= \sum_{i,j=1}^{N}\Big(\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\rho_{ij}\delta x_i \delta x_j\Big) \\ &= \sum_{i=1}^{N}\Big((\frac{\partial y}{\partial x_i})^2(\delta x_i)^2\Big) + 2\sum_{i=1<j}^{N}\Big(\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\rho_{ij}\delta x_i \delta x_j\Big)\end{aligned} \tag{4.48}$$

perché per definizione (4.42) e $\rho_{ii} = 1$. La covarianza però, e con essa il coefficiente di correlazione, può essere difficile da stimare se le variabili non sono misurate in coppia e a volte i termini della seconda sommatoria nella (4.48) vengono omessi. Una buona pratica di questo trattamento prevede che:

- se le variabili misurate sono tra loro indipendenti, la covarianza associata è nulla;
- se una variabile è una somma di due grandezze(come la massa di due pesi, ecc.), in generale i valori riportati per gli oggetti di prova nelle calibrazioni hanno covarianze non nulle di cui si deve tenere conto;

- nella pratica le covarianze dovrebbero essere incluse solo se sono state stimate a partire da un numero di dati sufficiente.

Nel caso di un prodotto di variabili indipendenti, o in cui si sceglie di omettere i termini di correlazione, può risultare comodo utilizzare gli errori relativi

$$y = \prod_{i=1}^{N} x_i^{a_i}$$
$$\left(\frac{\delta y}{y}\right)^2 = \sum_{i=1}^{N} \left(a_i \frac{\delta x_i}{x_i}\right)^2$$

per cui è sufficiente calcolare la radice quadrata dei valori ottenuti qui sopra per ottenere l'errore propagato. Esiste anche un'altra stima più pessimistica per l'errore, che è la seguente

$$\frac{\delta y}{|y|} = \sum_{i=1}^{N} |a_i| \frac{\delta x_i}{|x_i|}$$

Idealmente la stima per l'errore propagato come da (4.46) è un valore sostanzialmente identico a quello ottenuto direttamente da delle misurazioni, ma talvolta i due valori possono essere diversi. Questo può succedere a causa di covarianze impreviste o più semplicemente di sbagli nel calcolo della propagazione o nelle misurazioni. Sul sito NIST/SEMATECH[28] sono presenti diverse formule per funzioni specifiche in una, due o più variabili.

Capitolo 5

Lavoro di tesi

5.1 Impostazione del lavoro

Nel lavoro si sono dovuti ottenere al computer i rate di ibridazione di 2 diversi filamenti singoli di DNA con un hairpin in base a diverse lunghezze di attacco. Di seguito viene spiegato come sono stati progettati i diversi filamenti, come sono stati impostati il modello e i metodi utilizzati al computer e come sono stati trattati i dati ottenuti.

5.1.1 Progettazione dei filamenti di DNA

I diversi filamenti di DNA sono stati progettati utilizzando nupack [3] per ottenere sequenze che potessero dare un basso numero di complessi secondari. L'hairpin in figura 5.1b ha un fusto di 10 paia di basi, l'ansa è lunga 6 nucleotidi e la coda 4 nucleotidi. I filamenti complementari hanno uno l'attacco di 4 basi sulla coda (figura 5.1a) e l'altro un attacco di 4 basi all'interno dell'ansa(figura 5.1c). Le sequenze utilizzate, con le basi aggiunte per creare l'attacco tra parentesi quadre, sono:

Hairpin (H) 5' CGTACCAGGC TTCGCT GCCTGGTACG ACGC 3';

Coda (C) 5' [GCGT] CGTACCAGGC 3';

Ansa (A) 5' CGTACCAGGC [AGCG] 3'.

Il modello utilizzato per la progettazione dei filamenti ha anche permesso di calcolare eventuali strutture secondarie, in cui considerando una concentrazione iniziale di tutti i filamenti pari a 25 µM si hanno solo tracce di complessi indesiderati, a parte un complesso tra i 3 filamenti che risulta avere comunque una concentrazione almeno tre volte più bassa rispetto ai complessi desiderati all'equilibrio. In ogni caso nella valutazione dei complessi utilizzando solo due filamenti, quelli indesiderati risultano essere quasi assenti, permettendo di poter eventualmente controllare sperimentalmente i dati ottenuti in questa tesi.

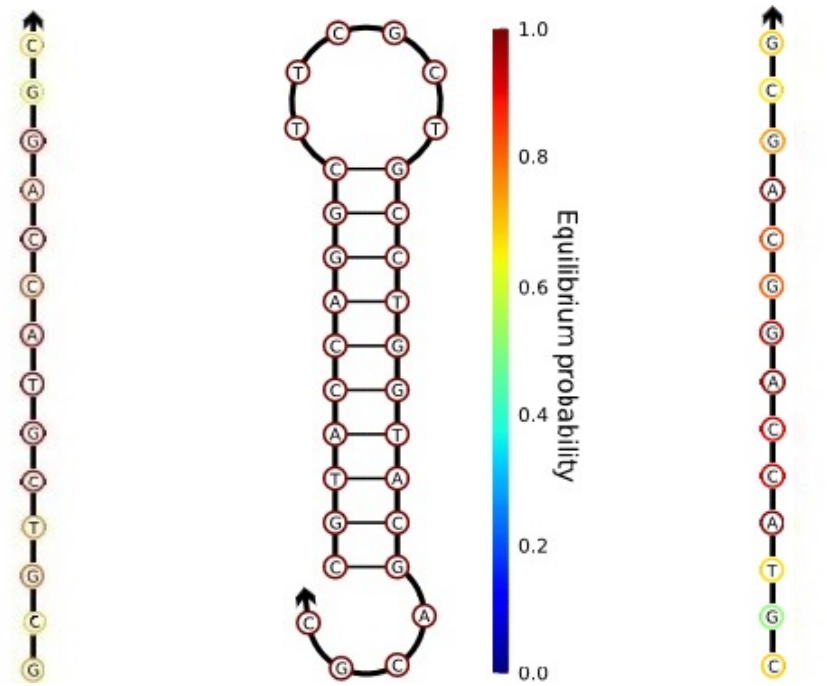


Figura 5.1: Illustrazioni delle sequenze di DNA progettate. L'hairpin (filamento H)(b) è un filamento di 30 nucleotidi con un fusto di 10 paia di basi che può essere attaccato da (a) sui 4 nucleotidi della coda (filamento C) e da (c) a partire da 4 dei 6 nucleotidi dell'ansa(filamento A). La direzione delle frecce indica la direzione $5' \rightarrow 3'$ dei filamenti. Si prevede che le sequenze scelte permettano di evitare una quantità consistente di strutture secondarie anche in un eventuale test di laboratorio utilizzando tutti i tre filamenti contemporaneamente. I test con numero di basi d'attacco inferiore sono stati eseguiti eliminando i nucleotidi più esterni dell'attacco alla coda di (a) e dell'attacco all'ansa di (c).
Figure 5.1: Designed DNA sequences illustrations. The hairpin(strand H) (b) is a strand of 30 nucleotides with a stem of 10 pairs of bases that can be attacked by (a) on the 4 nucleotide of the tail(strand C) and by (c) on 4 of the 6 nucleotides of the loop(strand A). The arrows point in the direction $5' \rightarrow 3'$ of the strands. The sequences should prevent the formation of a substantial quantity of secondary structures even in a potential lab test with all 3 strands used together. Tests with a reduced number of bases for the toehold were executed by elimination of the more exeternal nucleotides of attack on the tail for (a) and of the attack on the loop for (c).

5.1.2 Impostazioni di simulazione

Le sequenze di interesse sono state inserite in oxDNA[2] per avviare due diverse simulazioni LD, una con i filamenti H e C ed un'altra con i filamenti H e A. Le condizioni di simulazione sono state scelte in modo da riprodurre tipiche condizioni di laboratorio a temperatura $T = 300K$ e concentrazione di sali

di sodio alta [Na^+]= 1 M, in modo mascherare l'approssimazione del modello riguardo la mancanza di cariche elettrostatiche esplicite. Il coefficiente di diffusione è stato impostato come $D_{sim} = 5.99 \cdot 10^{-7}\ m^2/s$, più elevato rispetto al DNA reale, ma le proprietà dinamiche in un modello coarse grained risultano già di per sé accelerate dal lisciamento della superficie di energia potenziale e un ulteriore loro accelerazione può rappresentare un vantaggio, perché permette lo studio di sistemi più complessi, che altrimenti risulterebbero eccessivamente lenti. È stato dimostrato che questa differenza non influisce sul risultato qualitativo ottenuto[15], perciò, considerate le diverse semplificazioni già presenti nel modello, questo tipo di approccio può essere giustificato in nome di un campionamento efficiente. Essendo i processi in esame qualitativamente simili, è possibile compararli tra loro avendo cura di concentrarsi più sul rapporto tra i tassi trovati piuttosto che sui loro valori assoluti.

Per ogni test è stata prima eseguita una singola simulazione di equilibrazione di 10^8 passi con Δt =15.15 fs per permettere al filamento H, inizialmente lineare, di formare l'hairpin. La configurazione ottenuta da questa prima simulazione è stata utilizzata come punto di partenza per avviare una seconda simulazione secondo il metodo FFS per ottenere i tassi di ibridazione. Le varie interfacce sono state impostate in modo che il salvataggio delle configurazioni avvenisse a determinate condizioni:

λ_0 all'avvicinamento tra due nucleotidi dei due diversi filamenti.

λ_1 formazione di 1 legame a idrogeno tra le basi.

λ_2 formazione di 3 legami a idrogeno, cosa che in pratica equivale sempre a partire dai nucleotidi dell'ansa o della coda.

λ_3 formazione di 6 legami a idrogeno, portando necessariamente ad una parziale apertura della doppia elica.

λ_4 completa ibridazione tra i due filamenti in esame.

Al termine delle simulazioni si ottengono i flussi alla prima interfaccia $\phi(\lambda_0)$ e le varie probabilità di passaggio da un'interfaccia all'altra $p(\lambda_i|\lambda_{i-1})$, oltre a degli schemi ramificati delle diverse traiettorie di ibridazione. Si ottengono anche le probabilità p_j^i per ogni singola configurazione j di riuscire a raggiungere l'interfaccia successiva λ_{i+1}, calcolate come il numero di successi in rapporto al numero di tentativi a partire da quella particolare configurazione j. Il medesimo trattamento è stato applicato ai vari filamenti C e A ad attacco ridotto.

5.1.3 Trattamento statistico dei dati

Il calcolo dei tassi di ibridazione tra i filamenti segue quanto visto in sezione 3.2, mentre l'errore complessivo viene calcolato per mezzo della propagazione(sezione

4.3.3) degli errori standard delle singole interfacce (sezione 4.1.4). L'errore sul flusso viene calcolato come errore standard su tutti i tempi di simulazione per ottenere le configurazioni dell'interfaccia λ_0(sezione 4.1.6), calcolandone poi il valore per il reciproco rispetto alla media dei tempi per ottenere il numero di eventi al secondo, come da sezione 4.3.1. Per motivi pratici la covarianza tra le varie interfacce non è stata calcolata e il valore del coefficiente di correlazione è stato impostato pari a 1. Assumendo un valore di correlazione tanto elevato, è ragionevole supporre che l'errore tra le interfacce risulti più grande del necessario, ottenendo quindi delle stime pessimistiche per l'errore sul tasso di ibridazione complessivo.

Per il calcolo della probabilità di passare dalla prima all'ultima interfaccia $p(\lambda_n|\lambda_0)$ è stato utilizzato anche un metodo alternativo che in teoria dovrebbe restituire risultati uguali per un numero infinito, o comunque molto grande, di campioni. Si è calcolato il valore $p(\lambda_n|\lambda_0)_i$ per ogni singola configurazione i dell'interfaccia λ_0 secondo la logica in figura 5.2, quindi nel caso specifico di questo lavoro la probabilità per ogni singola configurazione iniziale $p(\lambda_4|\lambda_0)_i$ è data dalla

$$p(\lambda_4|\lambda_0)_i = p_i^0 \langle\, p_{i\to j}^1 \langle\, p_{j\to k}^2 \langle\, p_{k\to l}^3 \,\rangle \,\rangle \,\rangle, \tag{5.1}$$

permettendo di ottenere un campione con un numero di valori di probabilità pari al numero delle configurazioni presenti all'interfaccia λ_0. Utilizzando la media come stimatore della probabilità complessiva

$$p(\lambda_4|\lambda_0) = \langle p(\lambda_4|\lambda_0)_i \rangle \tag{5.2}$$

a tale campione è stato applicato il ricampionamento jackknife come in sezione 4.2.1, per ottenere l'errore jackknife sui dati. Per il funzionamento del FFS, al termine della simulazione alcune configurazioni possono rimanere del tutto inutilizzate, in questi casi si è scelto di escluderle del tutto dai calcoli, in quanto non sarebbe possibile stabilire se siano in grado di raggiungere l'interfaccia successiva e qualunque valore sostituito loro comporterebbe un'alterazione del risultato.

L'utilizzo degli schemi delle traiettorie ha consentito di poter osservare il numero di traiettorie indipendenti e la presenza o meno di ramificazioni anomale, consentendo di calibrare meglio il numero di configurazioni da campionare ad ogni interfaccia per ridurre l'errore sul tasso d'ibridazione finale o di cogliere eventuali errori d'impostazione. L'utilizzo del jackknife ha consentito di avere un confronto numerico che ha permesso anch'esso di cogliere eventuali difetti del campione.

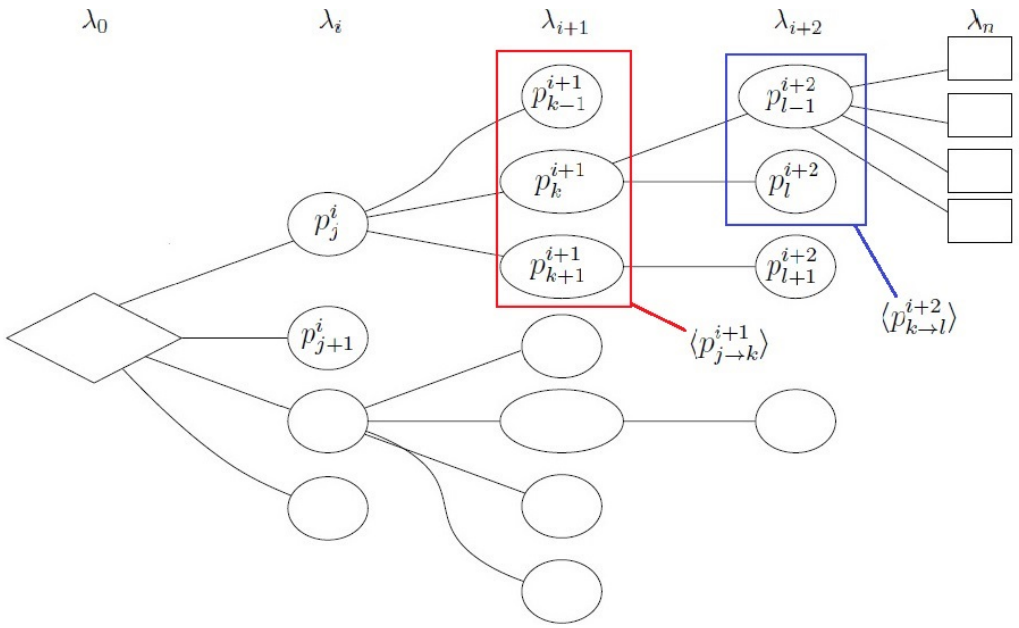


Figura 5.2: Schema ramificato di traiettorie risultanti dal FFS. Sono indicate l'interfaccia iniziale λ_0, quella finale λ_n e delle interfacce intermedie λ_i, con le probabilità delle singole configurazioni di passare all'interfaccia successiva p^i_j. Si consideri $\langle p^{i+1}_{j\to k}\rangle$ come la media delle probabilità di raggiungere l'interfaccia successiva da parte di tutte le configurazioni k dell'interfaccia $i+1$ discendenti dalla configurazione j. La probabilità che la configurazione j possa raggiungere l'interfaccia $i+2$ è $p(\lambda_{i+2}|\lambda_i)_j = p^i_j \langle p^{i+1}_{j\to k}\rangle$, ma a sua volta si ha $p(\lambda_{i+3}|\lambda_{i+1})_k = p^{i+1}_k \langle p^{i+2}_{k\to l}\rangle$. Il calcolo è iterativo, inizia da tutte le configurazioni all'interfaccia $i = n-2$ e procedendo a ritroso nell'albero fino ad arrivare a $i = 0$, come nel caso pratico dell'equazione (5.1). I rami morti hanno valore nullo, per esempio $p^i_{j+1} = p^{i+2}_{l+1} = 0$.

Figure 5.2: Branched diagram of FFS' trajectories. The starting interface λ_0, the final one λ_n and some intermediate interfaces λ_i are specified, with the probabilities of the single configurations to reach the next interface p^i_j. Consider $\langle p^{i+1}_{j\to k}\rangle$ as the mean of the probabilities to reach the next interface for all the configurations k of interface $i+1$ coming from configuration j. The probability that configuration j could reach interface $i+2$ is $p(\lambda_{i+2}|\lambda_i)_j = p^i_j \langle p^{i+1}_{j\to k}\rangle$, but on the other hand $p(\lambda_{i+3}|\lambda_{i+1})_k = p^{i+1}_k \langle p^{i+2}_{k\to l}\rangle$. The calculation is iterative, starting from all configurations at interface $i = n-2$ and going backward over the tree until $i = 0$, like in the practical case in equation (5.1). Dead branches have zero value, for example $p^i_{j+1} = p^{i+2}_{l+1} = 0$.

5.2 Risultati

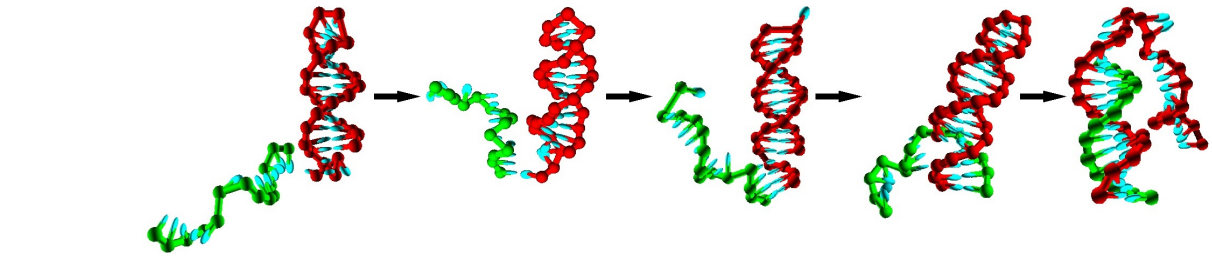
Nelle figure 5.3a e 5.4a sono rappresentate delle tipiche traiettorie di associazione per H-C e H-A rispettivamente. Si può osservare come dopo un iniziale contatto per diffusione si abbia l'accoppiamento di un paio di basi e il successivo "zippering" a riempire il sito di attacco sulla coda o sull'ansa. Naturalmente se le coppie di basi non si formano abbastanza velocemente, il contatto iniziale

può dissociarsi o il primo legame a idrogeno può rompersi. In seguito si ha la progressiva apertura della doppia elica dell'hairpin fino alla totale associazione dei due filamenti, salvo la regressione del processo. Lo studio delle traiettorie evidenzia che l'associazione dei due filamenti può fallire anche alla penultima interfaccia, quando l'apertura della doppia elica dell'hairpin è già in corso. Le immagini b-d delle figure 5.3 e 5.4 rappresentano le traiettorie che per le varie simulazioni hanno raggiunto λ_4.

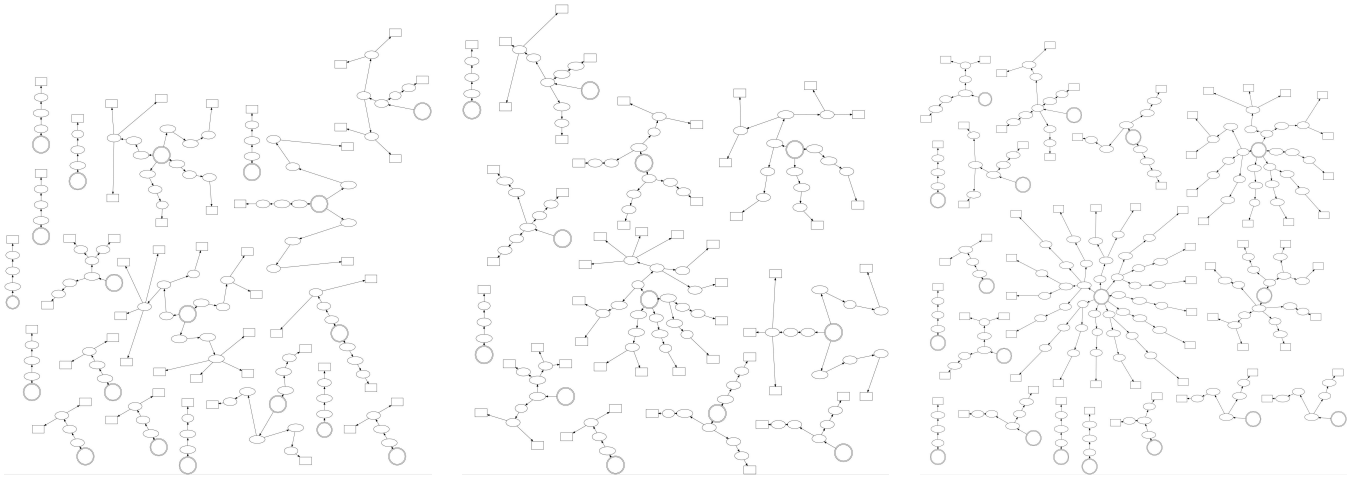
Nella tabella 5.1 e nella figura 5.5 sono riportati i dati sui tassi d'ibridazione HC in funzione della lunghezza d'attacco. È evidente che il tasso d'associazione aumenti esponenzialmente all'aumentare della lunghezza d'attacco, così come per l'ibridazione H-A i cui dati sono riportati in tabella 5.2 e figura 5.6. Entrambi i grafici mostrano che i risultati ottenuti con i due diversi metodi di calcolo sono praticamente sovrapponibili. La discrepanza maggiore si evidenzia per l'ibridazione H-C con attacco di 4 basi, dovuto probabilmente al maggior numero di configurazioni inutilizzate.

Entrambi i processi possono essere fittati linearmente rispetto al logaritmo dei tassi d'associazione ricavati, ma differiscono sia per l'ordine di grandezza a parità di basi d'attacco sia per la pendenza della retta trovata. Entrambe le rette sono riportate in figura 5.7b per poterle confrontare meglio tra loro. Data la natura di queste simulazioni, si è scelto di confrontarle con i risultati in figura 5.7a, ottenuti in un altro lavoro per la cinetica di spiazzamento di filamenti di DNA in rapporto alla lunghezza d'attacco [35], ritenuto un processo simile a quello qui in esame. La presenza di di dati sperimentali [45] rende questi dati un punto di riferimento piuttosto affidabile. Qualitativamente viene confermato il regime lineare nel range 2-4 basi d'attacco, mentre per attacchi più lunghi è preventivabile il raggiungimento di un valore massimo per il rate d'associazione. Di certo sia per l'ibridazione H-C che per quella H-A un attacco più lungo permette di avere una base più forte e stabile per portare a compimento l'associazione dei due filamenti prima che il contatto possa venir meno, come tra l'altro già dimostrato.

È interessante osservare proprio come l'attacco sull'ansa sia generalmente più lento di 2-3 ordini di grandezza. Si può ipotizzare che l'ansa dell'hairpin rappresenti un ostacolo che impedisca la formazione di un punto d'appoggio stabile e limiti i riarrangiamenti necessari al filamento che deve inserirsi, pertanto quando l'attacco è di sole 2 basi questo effetto si risente di più. Con l'aumentare della lunghezza d'attacco sembra che questo effetto limitante vada riducendosi e si potrebbe ipotizzare che per attacchi sufficientemente lunghi i due tipi di ibridazione potrebbero presentare differenze sempre minori. Il discorso però dipende sicuramente anche dalla dimensione dell'ansa che rappresenta non solo un limite alla possibile lunghezza d'attacco massima, ma è ragionevole supporre che per anse sufficientemente grandi le ibridazioni H-C e H-A debbano essere equivalenti.



(a)



(b) **(c)** **(d)**

Figura 5.3: (a) Un esempio di traiettoria di ibridazione H-C tipica con attacco di 4 basi andando da λ_0 verso λ_4. Si può vedere come l'attacco iniziale avvenga sulla coda dell'hairpin per poi aprire la doppia elica man mano che l'ibridazione si completa. In (b), (c) e (d) sono presenti gli schemi ramificati delle traiettorie per l'ibridazione a 2 basi, 3 basi e 4 basi d'attacco rispettivamente, privi dei rami morti. I doppi cerchi sono le configurazioni iniziali e i rettangoli quelle finali. Le ramificazioni sono ben presenti a tutti i livelli per l'attacco a 4 basi e si riducono per attacchi più corti.
Figure 5.3: (a) An example of H-C hybridization trajectory for a toehold of 4 bases going from λ_0 to λ_4. It can be viewed how the initial attack takes place on the tail of the hairpin and then opens the duplex as the hybridization becomes complete. The branched diagrams in (b), (c) and (d) represent the trajectories for hybridization with 2 bases, 3 bases and 4 bases of attack respectively, without dead branches. Double circles are initial configurations and rectangles are final configurations. The branching is well present at all levels for the 4 bases attack and becomes less present for shorter toeholds.

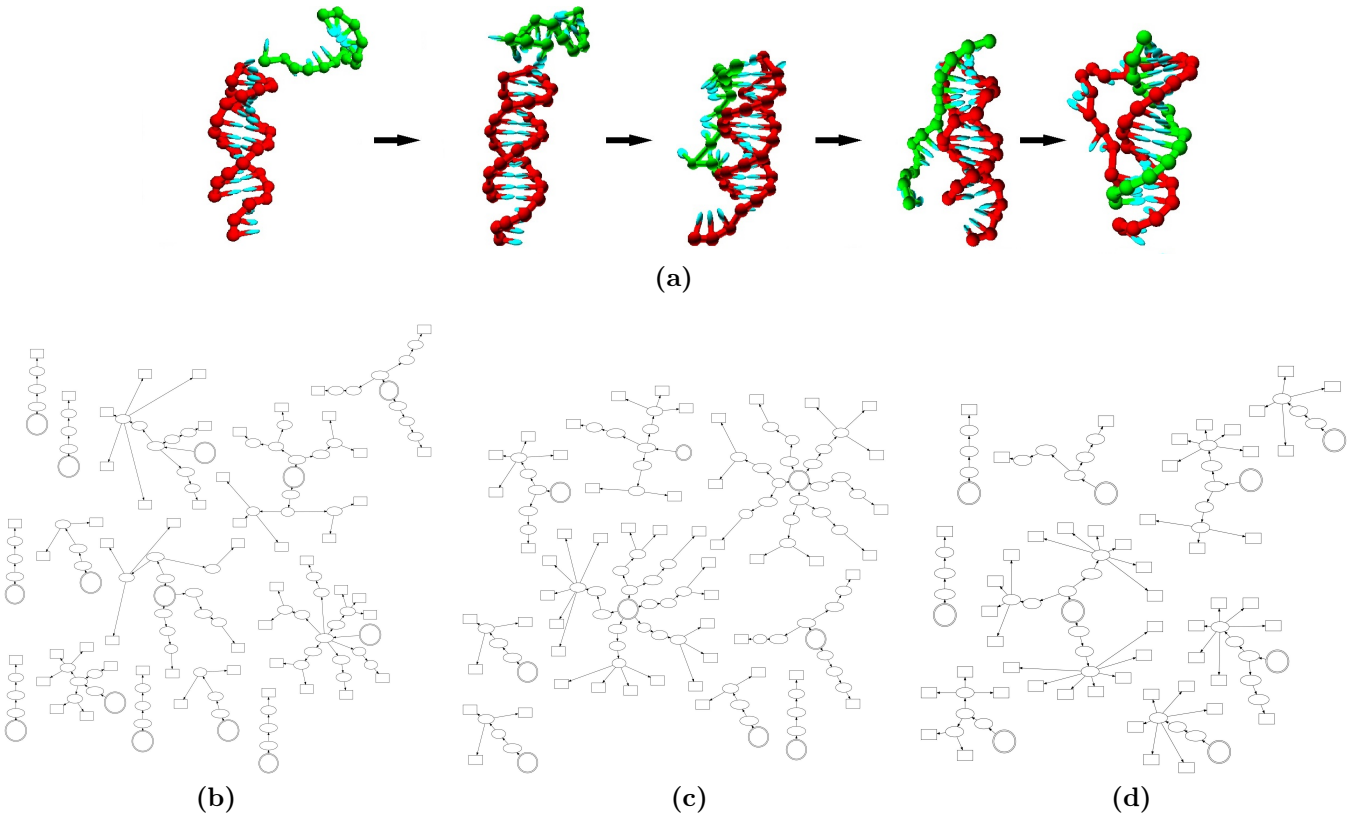


Figura 5.4: (a) Un esempio di traiettoria di ibridazione H-A tipica con attacco di 4 basi andando da λ_0 verso λ_4. Si può vedere come l'attacco iniziale avvenga sull'ansa dell'hairpin per poi aprire la doppia elica man mano che l'ibridazione si completa. In (b), (c) e (d) sono presenti gli schemi ramificati delle traiettorie per l'ibridazione a 2 basi, 3 basi e 4 basi d'attacco rispettivamente, privi dei rami morti. I doppi cerchi sono le configurazioni iniziali e i rettangoli quelle finali. Rispetto all'ibridazione H-C le ramificazioni tendono a concentrarsi soprattutto alle ultime interfacce, forse perché l'ansa presenta un ingombro che limita il numero di conformazioni spaziali favorevoli all'associazione completa dei due filamenti.

Figure 5.4: (a) An example of H-A hybridization trajectory for a toehold of 4 bases going from λ_0 to λ_4. It can be viewed how the initial attack takes place on the loop of the hairpin and then opens the duplex as the hybridization becomes complete. The branched diagrams in (b), (c) and (d) represent the trajectories for hybridization with 2 bases, 3 bases and 4 bases of attack respectively, without dead branches. Double circles are initial configurations and rectangles are final configurations. Compared to H-C hybridization the branching tends to concentrate particularly in the last interfaces, maybe because the loop sterically limits the number of spatial conformations favourable for the full association of the two strands.

Tabella 5.1: Risultati di simulazione per l'ibridazione H-C a 300 K. Nella seconda colonna è riportato il numero di traiettorie indipendenti rispetto alle configurazioni totali di λ_4.
Table 5.1: Results for ibridization rates and related errors for various toehold lenghts for ibridization H-C at 300 K. In second column there is the number of independent trajectories in relation to the total number of configurations for λ_4.

Ibridazione H-C N° basi	Traiettorie indipendenti	Rate	Errore	Errore relativo
2 basi	19/50	1.22515E+01	4.63273E+00	0.3781363
jackknife		1.13781E+01	4.23987E+00	0.3726356
3 basi	12/52	1.80717E+02	6.66682E+01	0.3689089
jackknife		1.84314E+02	7.33071E+01	0.3977289
4 basi	18/70	1.42020E+03	4.10965E+02	0.2893719
jackknife		1.19339E+03	4.32984E+02	0.3628193

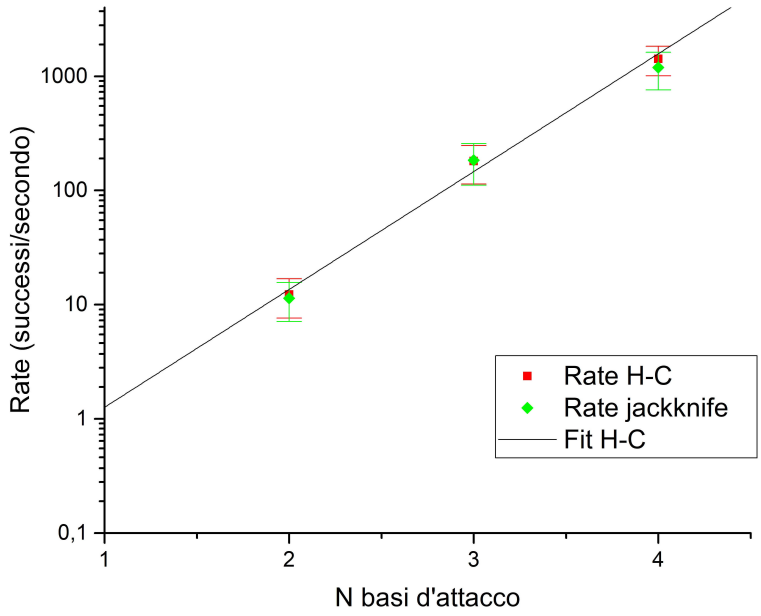


Figura 5.5: Grafico dei rates contro il numero di basi d'attacco dai dati in tabella 5.1. Si può notare una sostanziale coincidenza dei valori per entrambi i metodi di calcolo. Per ogni base d'attacco in meno si ha un abbassamento del rate d'ibridazione di circa un ordine di grandezza, con fit lineare rispetto al logaritmo base 10 dei rates. La retta ha pendenza 1.03208, ma è ipotizzabile che ad un numero maggiore di basi si raggiunga un rate massimo.
Figure 5.5: Graph of rates versus toehold lenght from data in table 5.1. The values for both calculation methods are almost corresponding. For every base less of attack there is about an order of magnitude between the corresponding rates, with linear fit against the logarithm base 10 of the rates. The straight line has slope 1.03208, but it can be supposed to be a maximum rate for longer toeholds.

Tabella 5.2: Risultati di simulazione per l'ibridazione H-A a 300 K. Nella seconda colonna è riportato il numero di traiettorie indipendenti rispetto alle configurazioni totali di λ_4.
Table 5.2: Results for ibridization rates and related errors for various toehold lenghts for ibridization H-A at 300 K. In second column there is the number of independent trajectories in relation to the total number of configurations for λ_4.

Ibridazione H-A N° basi	Traiettorie indipendenti	Rate	Errore	Errore relativo
2 basi	14/51	1.14681E-02	3.94195E-03	0.3437329
jackknife		1.00623E-02	3.82859E-03	0.3804871
3 basi	9/52	1.41303E-01	5.57091E-02	0.3942515
jackknife		1.46298E-01	7.12328E-02	0.4869004
4 basi	9/51	4.65863E+00	1.83813E+00	0.3945640
jackknife		4.68211E+00	2.50291E+00	0.5345686

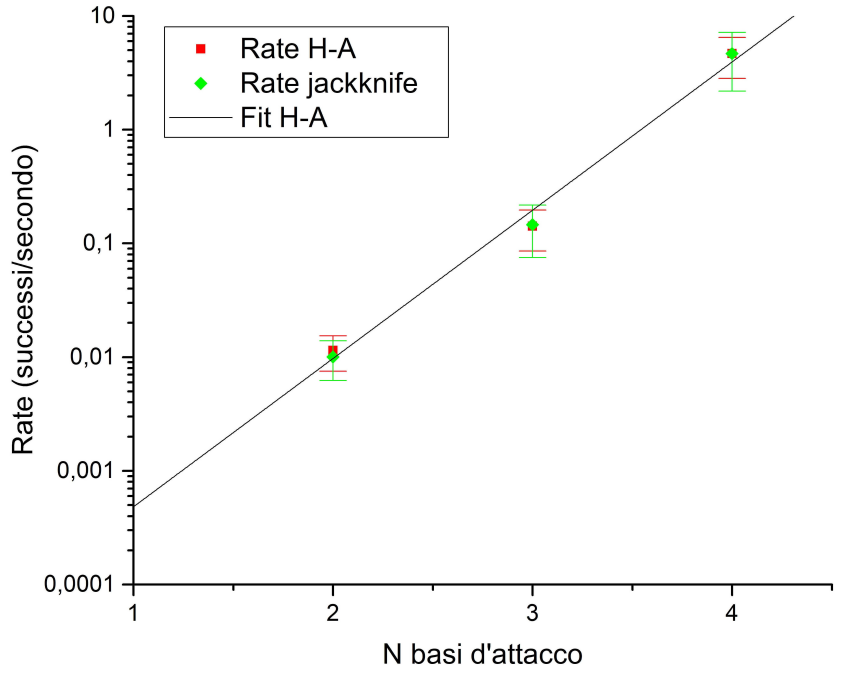


Figura 5.6: Grafico dei rates contro il numero di basi d'attacco dai dati in tabella 5.2. Si può notare una sostanziale coincidenza dei valori per entrambi i metodi di calcolo. Si nota un andamento simile a quello in grafico 5.5, anche se per valori 2-3 ordini di grandezza più bassi, con fit lineare rispetto al logaritmo base 10 dei rates. La retta ha pendenza 1.304383.
Figure 5.6: Graph of rates versus toehold lenght from data in table 5.2. The values for both calculation methods are almost corresponding. It can be noted a similar trend with respect to figure 5.5, even if for values smaller of 2-3 orders of magnitude, with linear fit against the logarithm base 10 of the rates. The straight line has slope 1.304383.

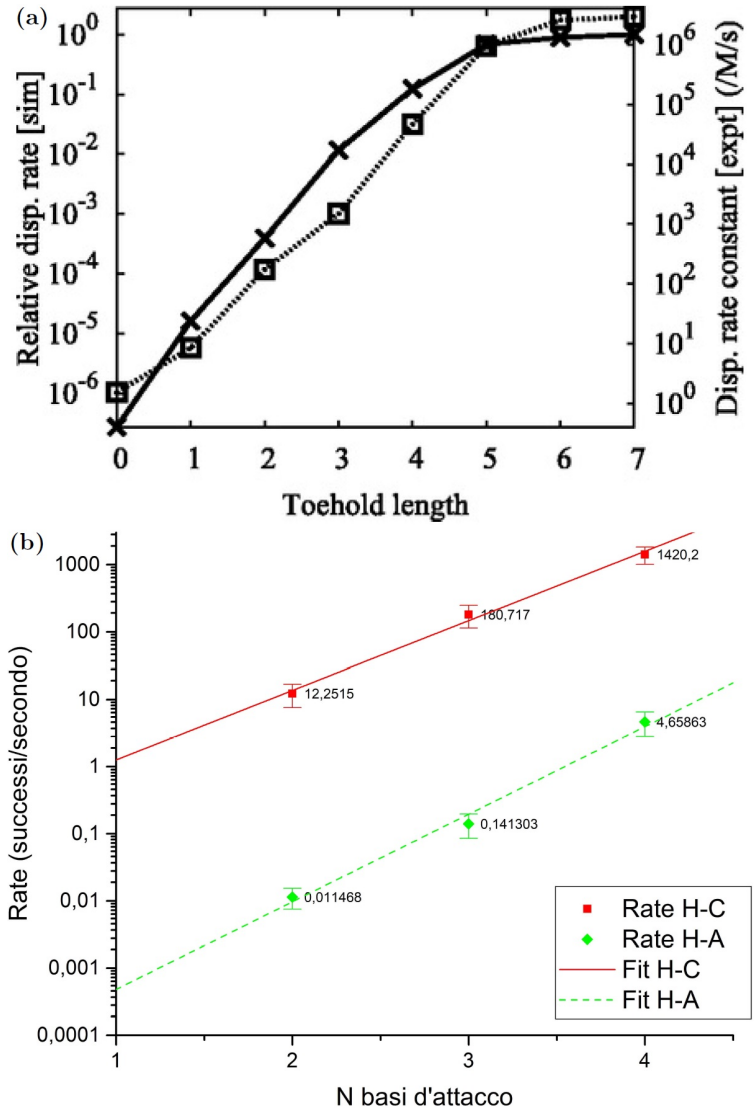


Figura 5.7: (a) Grafico dei rates relativi di spiazzamento di filamenti da duplex in rapporto al numero di basi d'attacco a 25°C. Le croci indicano dati da simulazioni con oxDNA e i quadrati dati sperimentali ottenuti da [45]. I due andamenti sono qualitativamente simili e hanno la stessa dipendenza esponenziale per lunghezze d'attacco brevi fino a 4-5 basi, per poi tendere ad un valore massimo per lunghezze d'attacco più grandi. L'accelerazione complessiva tra 2 e 4 basi d'attacco è di $10^{2.498}$ per la simulazione. Immagine da [35]. (b) Confronto dei dati di questa tesi per l'attacco sulla coda (quadrato) e l'attacco sull'ansa (rombo). Le due simulazioni restituiscono accelerazioni complessive di $10^{2.064}$ e $10^{2.609}$ rispettivamente.

Figure 5.7: (a) Rate of strand displacement from a duplex as a function of toehold length. The crosses are data from simulations with oxDNA and the squares are experimental data from [45]. The two trends are qualitatively similar and and have the same exponential dependence for short toeholds until 4-5 bases, then it reaches a plateau for longher toeholds. The overall acceleration from 2 to 4 bases toehold is $10^{2.498}$ for the simulation. Figure from [35]. (b) Confrontation of data from this thesis for the attack on the tail (squares) and the attack on the loop(diamond). The two simulations give overall accelerations of $10^{2.064}$ and $10^{2.609}$ respectively.

5.3 Conclusioni

In questo studio per l'ibridazione con hairpin di DNA è stato utilizzato un modello coarse grained, oxDNA, che è stato ottimizzato per rappresentare in modo adeguato il DNA sia a filamento singolo che doppio, permettendo quindi una descrizione realistica sia della rigidità delle doppie eliche che della flessibilità dei filamenti singoli. Queste caratteristiche permettono di descrivere in modo adeguato anche gli hairpin.

Altri studi sullo spiazzamento di filamenti da un duplex di DNA[35] avevano già messo in luce come i tassi d'associazione dipendessero esponenzialmente dalla lunghezza di brevi punti di appoggio, per poi giungere ad un plateau per attacchi più lunghi. In questo caso si è voluto verificare la presenza di eventuali differenze nei tassi d'associazione in funzione della lunghezza del punto d'appoggio in due diversi casi, se l'attacco avviene sulla coda dell'hairpin o alternativamente sulla sua ansa.

Si è verificato che i contatti iniziali avvengono effettivamente sui punti di appoggio, anche se spesso si dissociano prima di condurre a un'ibridazione completa perché le configurazioni non sono favorevoli. Il fallimento può avvenire anche in stadi più avanzati per via della richiusura della doppia elica dell'hairpin.

Le limitate rirorse computazionali a disposizione hanno permesso di valutare un numero limitato di casi, nello specifico con punto d'appoggio variabile tra le 2 e le 4 basi. Trattandosi di attacchi brevi si è riscontrato lo stesso tipo di dipendenza esponenziale in funzione del numero di basi d'attacco per entrambi i casi d'ibridazione, ma si sono rilevate anche delle marcate differenze. In questo range l'attacco sull'ansa ha restituito tassi d'associazione generalmente 2 o 3 ordini di grandezza più grandi rispetto al caso dell'attacco sull'ansa, dovuto probabilmente allo spazio limitato che l'ansa rappresenta per eventuali riarrangiamenti del filamento A in entrata, rendendo più improbabile avere una configurazione favorevole al raggiungimento di un'associazione completa.

La diversa pendenza delle rette sembra suggerire che per punti d'appoggio più lunghi i tassi d'associazione dei due casi potrebbero giungere ad un plateau allo stesso valore, ma sarebbe da valutare anche l'ingombro sterico che rappresenterebbe l'ansa per un numero di basi di attacco maggiore.

Nelle condizioni approfondite in questo lavoro ci si può limitare a dire che un punto d'appoggio sull'ansa non equivale ad un punto d'appoggio sulla coda di un hairpin e restituisce tassi d'associazione inferiori. Approfondendo come l'associazione di filamenti complementari ad un hairpin vari in base a diverse condizioni è possibile suggerire modi per modulare i tassi d'ibridazione e quindi riuscire a dirigere nel modo desiderato le nanotecnologie a DNA che prevedano l'uso di hairpin. Derivando i risultati da proprietà generiche del DNA ci si aspetta che questi possano applicarsi anche a sistemi simili basati su acidi nucleici, come l'RNA.

Lavori futuri potrebbero mettere in luce questi aspetti esplorando un range più ampio di punti d'appoggio o in aggiunta valutare l'effetto di anse di dimensioni diverse. Le sequenze di DNA progettate dovrebbero ipoteticamente permettere la formazione di quantità minime di complessi secondari indesiderati, perciò sarebbe auspicabile anche ottenere un confronto con dei dati sperimentali. Un ulteriore lavoro che potrebbe essere svolto sarebbe la simulazione di queste stesse sequenze in condizioni biologiche o in generale per un range di temperature.

Bibliografia

[1] URL: `http://ndbserver.rutgers.edu/` (visitato il 03/06/2017).

[2] URL: `https://dna.physics.ox.ac.uk` (visitato il 27/04/2017).

[3] URL: `http://www.nupack.org` (visitato il 18/05/2017).

[4] Rosalind J. Allen, Chantal Valeriani e Pieter Rein ten Wolde. «Forward flux sampling for rare event simulations». In: *Journal of Physics: Condensed Matter* 21.463102 (2009).

[5] Lizabeth A. Allison. *Fondamenti di biologia molecolare*. Zanichelli, 2008.

[6] Helen M. Berman, Wilma K. Olson, David L. Beveridge et al. «The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids». In: *Biophysical Journal* 63 (1992), pp. 751–759.

[7] G. Bohm e G. Zech. *Introduction to statistics and data analysis for physicists*. DESY, 2010.

[8] Marco Bortoluzzi. *Approccio qualitativo alla Chimica computazionale*. Aracne, 2009.

[9] E. Chargaff. «Chemical specificity of nucleic acids and mechanism of their enzymatic degradation». In: *Experimentia* 6 (1950), pp. 201–209.

[10] T. E. Cheatham III. «Molecular modeling and atomistic simulation of nucleic acids». In: *Annual Reports in Computational Chemistry* 1 (2005), pp. 75–89.

[11] Chunlai Chen, Wenjuan Wang, Zhang Wang et al. «Influence of secondary structure on kinetics and reaction mechanism of DNA hybridization». In: *Nucleic Acid Research* 35 (2007), pp. 2875–2884.

[12] Ruslan L. Davidchack, Richard Handel e M. V. Tretyakov. «Langevin thermostat for rigid body dynamics». In: *The Journal of Chemical Physics* 130.234101 (2009).

[13] C. De Michele et al. «Self-Assembly of Short DNA Duplexes: From a Coarse-Grained Model to Experiments through a Theoretical Link». In: *Soft Matter* 8 (2012), pp. 8388–8398.

[14] C. F. Dietrich. *Uncertainty, Calibration and Probability. The Statistics of Scientific and Industrial Measurement.* Taylor & Francis Group, 1991.

[15] Thomas E. Ouldridge, Petr Šulc, Flavio Romano et al. «DNA hybridisation kinetics: zippering, internal displacement and sequence dependence». In: *Nucleic Acid Research* 41.19 (2013), pp. 8886–8895.

[16] Thomas E. Ouldridge, Rollo L. Hoare, Ard A. Louis et al. «Optimizing DNA Nanotechnology through Coarse-Grained Modeling: A Two-Footed DNA Walker». In: *American Chemical Society* 7.3 (2013), pp. 2479–2490.

[17] B. Efron. «Bootstrap Methods: Another Look at the Jackknife». In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[18] B. Efron e C. Stein. «The jackknife estimate of variance». In: *The Annals of Statistics* 9.3 (1981), pp. 586–596.

[19] D. Frenkel e B. Smit. *Understanding Molecular Simulations.* Academic Press, 1996.

[20] Leo A. Goodman. «On the Exact Variance of Products». In: *Journal of the American Statistical Association* 55.292 (1960), pp. 708–713.

[21] Yoshihiro Ito e Eiichiro Fukusaki. «DNA as a 'Nanomaterial'». In: *Journal of Molecular Catalysis B: Enzymatic* 28 (2004), pp. 155–166.

[22] V. Kumar, S. Palazzolo, S. Bayda et al. «DNA Nanotechnology for Cancer Therapy». In: *Theranostics* 6 (2016), pp. 710–725.

[23] Young-Wan Kwon et al. «General Method for Modification of Liposomes for Encoded Assembly on Supported Bilayers». In: *Journal of the American Chemical Society* 127 (2005), pp. 1356–1357.

[24] Young-Wan Kwon et al. «Materials science of DNA». In: *Journal of Materials Chemistry* 19 (2009), pp. 1353–1380.

[25] A.D. MacKerell Jr. e L. Nilsson. «Molecular dynamics simulations of nucleic acid-protein complexes». In: *Current Opinion in Structural Biology* 18 (2008), pp. 194–199.

[26] Rachel McKendry, Jiayun Zhang, Youri Arntz et al. «Multiple label-free biodetection and quantitative DNA-binding assays on a nanomechanical cantilever array». In: *Proceeding of the National Academy of Sciences* 99.15 (2002), pp. 9783–9788.

[27] Buvaneswari Coimbatore Narayanan, John Westbrook, Saheli Ghosh et al. «The Nucleic Acid Database: new features and capabilities». In: *Nucleic Acid Research* 42 (2013), pp. 114–122.

[28] NIST/SEMATECH. *e-Handbook of Statistical Methods.* URL: `http://www.itl.nist.gov/div898/handbook/mpc/section5/mpc55.htm` (visitato il 29/03/2017).

[29] Modesto Orozco, Agnes Noy e Alberto Pérez. «Recent advances in the study of nucleic acid flexibility by molecular dynamics». In: *Current Opinion in Structural Biology* 18 (2008), pp. 185–193.

[30] Thomas E. Ouldridge, Ard A. Louis e Jonathan P. K. Doye. «Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model». In: *The Journal of Chemical Physics* 134.8 (2011).

[31] F. Romano, D. Chakraborty, J. P. K. Doye et al. «Coarse-Grained Simulations of DNA Overstretching». In: *The Journal of Chemical Physics* 138.085101 (2013).

[32] John SantaLucia e Donald Hicks. «The Thermodynamics of DNA Structural Motifs». In: *Annual Review of Biophysics and Biomolecular Structure* 33 (2004), pp. 415–440.

[33] Tamar Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer, 2010.

[34] Jun Shao e C.F.J. Wu. «A general theory for jackknife variance estimation». In: *The Annals of Statistics* 17.3 (1989), pp. 1176–1197.

[35] Niranjan Srinivas et al. «On the biophysics and kinetics of toehold-mediated DNA strand displacement». In: *Nucleic Acids Research* 41.22 (2013), p. 10641.

[36] N. Srinivas, T. E. Ouldridge, P. Šulc et al. «On the biophysics and kinetics of toehold-mediated DNA strand displacement». In: *Nucleic Acids Research* 41 (2013), pp. 10641–10658.

[37] Petr Šulc, Flavio Romano, Thomas E. Ouldridge et al. «Sequence-dependent thermodynamics of a coarse-grained DNA model». In: *The Journal of Chemical Physics* 137.13 (2012).

[38] V. Tozzini. «Coarse Grained Models for Proteins». In: *Current Opinion in Structural Biology* 15 (2005), pp. 144–150.

[39] John W. Tukey. «Bias and confidence in not-quite large samples (abstract)». In: *Annals of Mathematical Statistics* 29.2 (giu. 1958), pp. 614–623.

[40] L. Verlet. «Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules». In: *Physical Review* 159 (1967), pp. 98–103.

[41] J. D. Watson e F. H. C. Crick. «The structure of DNA». In: *Cold Spr. Harb. Symp. Quant. Biol.* 18 (1953), pp. 123–131.

[42] James D. Watson et al. *Biologia molecolare del gene*. Zanichelli, 2015.

[43] Qian. Xiaoliang, Daniel Strahs e Tamar Schlick. «Dynamic simulations of 13 TATA variants refine kinetic hypotheses of sequence/activity relationships». In: *Journal of Molecular Biology* 308 (2001), pp. 681–703.

[44] M. A. Young, G. Ravishanker e D. L. Beveridge. «A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation». In: *Biophysical Journal* 73 (1997), pp. 2313–2336.

[45] David Yu Zhang e Erik Winfree. «Control of DNA Strand Displacement Kinetics Using Toehold Exchange». In: *Journal of American Chemical Society* 131 (2009), pp. 17303–17314.

[46] Decai Zhanga, Yurong Yana, Qing Li et al. «Label-free and high-sensitive detection of Salmonella using a surface plasmon resonance DNA-based biosensor». In: *Journal of Biotechnology* 160 (2012), pp. 123–128.