

Metodi di Modellizzazione Multi-scala

1. Organizzazione gerarchica della materia vivente

La materia vivente presenta una complessità enorme se paragonata anche agli esempi più complessi di materia inanimata. Uno dei motivi di questa complessità è la sua intrinseca organizzazione gerarchica multi-scala, schematicamente rappresentata in Fig 1.1. I processi biologici avvengono su diverse scale di lunghezza e tempi. Partendo dal basso, sulla scala degli Å-nm e su tempi dell'ordine di fsec-nsec, normalmente hanno luogo le reazioni chimiche che danno l'avvio ad un processo. Queste coinvolgono di norma siti attivi di proteine o altre biomolecole, che includono la varietà chimicamente attiva e alcuni gruppi funzionali che la circondano. Poiché si tratta di reazioni chimiche, ovvero processi che coinvolgono sostanziali cambiamenti nella struttura elettronica del sistema, è ovvio che sia necessario un tipo di modellizzazione che rappresenti esplicitamente i gradi di libertà elettronici che vanno trattati con la meccanica quantistica. Le diverse metodologie che cadono in questa tipologia di approccio vengono globalmente chiamate Quantum Chemistry (QM)¹.

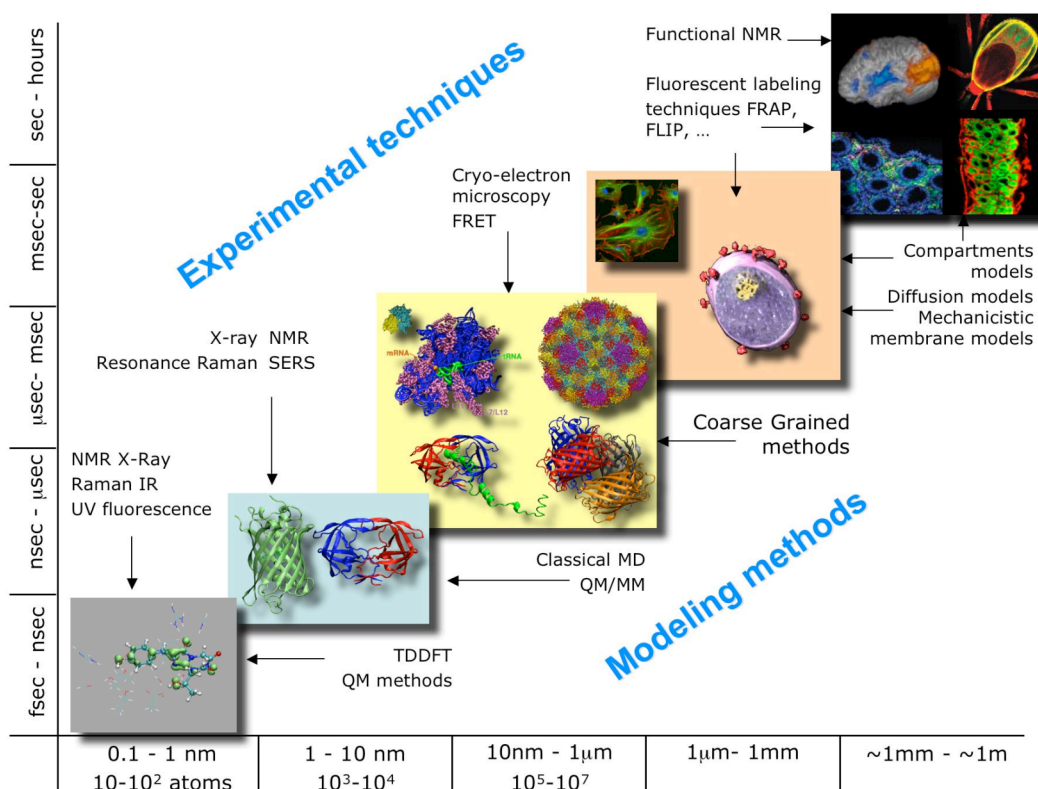


Fig 1.1 Organizzazione gerarchica della materia vivente. Nelle figure, dal basso a sinistra verso alto a destra: cromoforo e sito attivo della GFP; GFP e proteasi di HIV-1; proteasi di HIV-1, tetramero di DsRed, Ribosoma e un piccolo virus; Tomografia di Virus di HIV e cellula con strutture interne visualizzate con tecniche di microscopia a fluorescenza; tessuti, organi e un intero piccolo organismo visualizzati con tecniche di microscopia a fluorescenza e NMR funzionale. Metodi modellistici e sperimentali tipicamente usati sulle varie scale sono riportati. Sigle e acronimi saranno spiegati nel testo nelle sezioni successive.

Al gradino successivo della scala gerarchica, tra 1 e 10 nm e su scale di tempi tra nsec e μsec avvengono transizioni strutturali che coinvolgono intere biomolecole funzionali (proteine, ad esempio). Poiché è necessario simulare l'intera proteina in ambiente acquoso, si ha a che fare con sistemi di 10000-100000 atomi, che sono impossibili da simulare tenendo conto esplicitamente dei gradi di libertà elettronici. Quindi si usano approcci basati sulla descrizione empirica delle interazioni interatomiche nei quali gli effetti degli elettroni sono trattati implicitamente e la dinamica atomica è di tipo classico. Questi metodi si chiamano globalmente *all-atom Force Field based simulations*, e sono spesso indicati anche con le sigle MM (Molecular Mechanics) o FF

based simulations².

Su scale ancora maggiori, quelle tipiche degli aggregati macromolecolari, da sistemi proteina-ligando ($\sim 10\text{nm}$) fino ai capsidi di piccoli virus ($\sim \mu\text{m}$) passando per aggregati proteici, ribosomi, nucleosomi etc, le scale di tempi delle transizioni strutturali si allungano corrispondentemente, fino ad arrivare a tempi macroscopici (msec, ma anche ore o giorni, se si includono ad esempio precessi tipo il ripiegamento delle proteine). Dal momento che il tempo macchina per le simulazioni a risoluzione atomica aumenta polinomialmente con il numero di gradi di libertà, e che per questi sistemi si arriva fino a ~ 10 milioni di atomi, questi sistemi non sono tipicamente affrontabili con metodi basati su campi di forse all-atom, ed è necessario utilizzare rappresentazioni semplificate a bassa risoluzione (*Coarse Grained*, CG) in cui gruppi di atomi (ad esempio interi ammino acidi) vengono rappresentati come un unico centro interattivo (*bead*)³.

Su scale ancora maggiori, al livello di grandi virus o cellule (fino al mm e al sec) qualsiasi tipo di approccio per rappresentare la dinamica interna di macromolecole diventa praticamente inapplicabile, e si passa a descrizioni molto grossolane, che trattano come singoli oggetti rigidi interi "corpuscoli" e come mezzi continui i sistemi estesi (citoplasma, membrane cellulari). Infine, esistono alcuni sistemi di modellizzazione per trattare alcuni aspetti della meccanica e diffusione di sostanze tra gli organi interni, di cui però non ci occuperemo qui.

In queste lezioni verranno presentate, al livello introduttivo, le tecniche di modellizzazione tipiche adottate per processi che avvengono sulle varie scale di grandezza. È fondamentale notare che normalmente un singolo processo biologico coinvolge più di una di queste scale, dunque necessita, per una sua descrizione completa, l'adozione di un approccio multi-scala. È evidente che la coerenza tra le varie descrizioni è in questo caso un requisito fondamentale.

2. Concetti generali

2.1 Riduzione dei gradi di libertà

Nonostante la varietà di descrizioni e approcci, alcuni concetti generali sono ricorrenti. Uno di questi è la riduzione dei gradi di libertà che si opera nel salire i gradini della scala gerarchica. Il costo computazionale della modellizzazione scala generalmente in maniera polinomiale con il numero di gradi di libertà N , cioè come N^γ , dove γ è generalmente intorno a 2 e dipende dal tipo di modellizzazione. Dunque il metodo più ovvio per rendere possibili simulazioni al calcolatore di sistemi di dimensioni crescenti è ridurre il numero di gradi di libertà. Si è già accennato a due modi tipici in cui ciò viene fatto. Uno è l'eliminazione delle variabili elettroniche nel passaggio dalla descrizione QM alla descrizione atomica. In questo caso non solo viene ridotto N , ma si eliminano variabili quantistiche, che sono più computazionalmente costose. Infatti ad esse sono associate funzioni d'onda che richiedono molte più risorse computazionali di un grado di libertà classico, associato a tre semplici variabili reali. Infatti la funzione d'onda (funzione "continua" nello spazio tridimensionale) viene di fatto discretizzata con diverse metodologie e descritta tramite un insieme di parametri che assumono la funzione di coordinate interne e il cui numero dipende dalla precisione con che si vuole ottenere. Un grado di libertà quantistico può tipicamente equivalere a decine o centinaia di gradi di libertà classici, per cui risulta evidente il vantaggio nell'eliminazione della trattazione esplicita degli elettroni.

Nella salita dei gradini successivi della scala gerarchica si eliminano gradi di libertà classici e quindi il guadagno è direttamente collegato al numero di gradi eliminati. Ad esempio nel *coarse graining* che si opera salendo dal livello di bio-molecola a quello di aggregato molecolare, tipicamente si inglobano interi amminoacidi o nucleotidi in singoli centri interattivi, riducendo di quasi due ordini di grandezza il numero di gradi di libertà. Infine, nel passaggio al continuo, tutti i gradi di libertà discreti vengono eliminati a favore di una descrizione continua del mezzo, tramite un parametro "funzione" delle coordinate spaziali (ad esempio gli spostamenti dal piano della membrana cellulare descritta come un mezzo continuo bidimensionale). Come nel caso delle funzioni d'onda, il parametro viene discretizzato (ad esempio, usando una griglia di punti) e il numero di gradi di libertà necessari dipende dalla precisione con cui si vuole rappresentare il mezzo, comunque, tipicamente, si parla ancora di riduzione di un paio di ordini di grandezza nel numero effettivo di gradi di libertà. In Fig. 2.1 sono mostrati esempi tipici dei diversi tipi possibili di riduzione dei gradi di libertà.

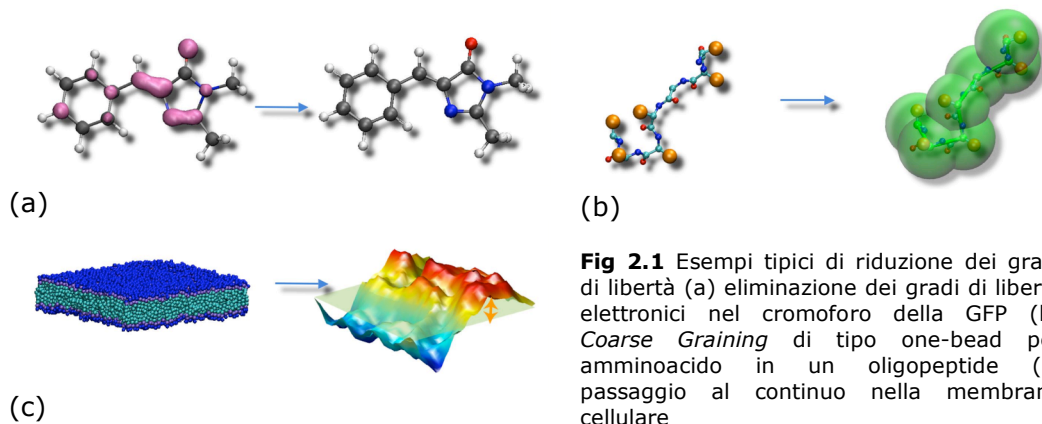


Fig 2.1 Esempi tipici di riduzione dei gradi di libertà (a) eliminazione dei gradi di libertà elettronici nel cromoforo della GFP (b) *Coarse Graining* di tipo one-bead per amminoacido in un oligopeptide (c) passaggio al continuo nella membrana cellulare

2.2 Superfici di energia

Una volta fissato l'insieme di variabili interne è necessario definire come queste interagiscono tra di loro, il che determinerà le proprietà energetiche, dinamiche e termodinamiche del sistema. Qui troviamo un altro concetto ricorrente che è quello di PES (Potential Energy Surface) o della sua estensione a temperatura finita, la FES (Free Energy Surface). In generale, la P(F)ES è una funzione delle coordinate interne del sistema che restituisce l'energia potenziale o l'energia libera. Viene chiamata "superficie" perché viene di solito rappresentata come un profilo o una superficie bidimensionale, che è o la proiezione su uno (o due) assi della superficie multidimensionale oppure un'approssimazione a dimensionalità ridotta dell'intera varietà multidimensionale.

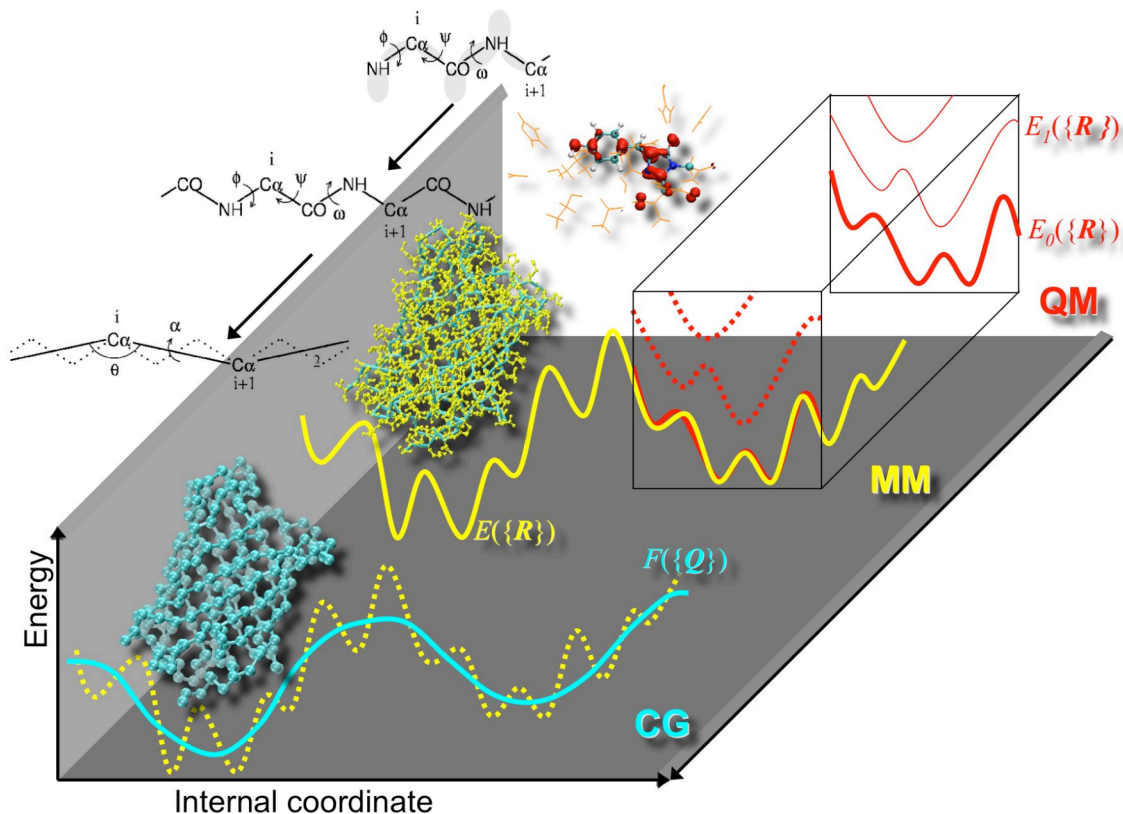


Fig 2.2 Variazione del concetto di P(F)ES nel passaggio tra i livelli gerarchici. In rosso: livello QM, in giallo, livello all-atom MM, in azzurro, livello CG. La proteina GFP è rappresentata ai tre livelli.

In Fig 2.2 è schematicamente rappresentato come varia il concetto di P(F)ES passando da un livello all'altro nella scala gerarchica. Al livello QM la PES è una funzione delle coordinate nucleari $\{\mathbf{R}\}$ e contemporaneamente un funzionale della funzione d'onda elettronica $\{\phi\}$. Tuttavia la dipendenza esplicita da $\{\phi\}$ scompare nell'approssimazione di Born-Hoppenheimer (BO): si

assume che gli elettroni seguano adiabaticamente i nuclei, quindi in ogni istante, per ogni configurazione nucleare, essi occupano uno stato elettronico ben definito che risulta dalla soluzione dell'equazione di Schrödinger a quella fissata configurazione nucleare. Quindi, di fatto, ci si ritrova con un set di energie non più dipendenti dalle funzioni d'onda, ma dipendenti dal numero quantico I , $E_I(\{\mathbf{R}\})$. La soluzione dell'equazione di Schrödinger per calcolare le PES è esattamente ciò che rende pesante il calcolo QM e limita le dimensioni del sistema simulabile con questo approccio.

Per risparmiare tempo di calcolo e consentire la simulazione di sistemi relativamente grandi, negli approcci MM si sceglie di approssimare la PES con una funzione analitica empirica, anziché di calcolarla quantisticamente. Anzitutto è necessario scegliere uno stato elettronico che è tipicamente, ma non necessariamente, lo stato fondamentale. La PES di questo stato viene approssimata con una somma di termini empirici analitici con un significato chimico-fisico: come vedremo in dettaglio si distinguono termini che descrivono il legame chimico, termini di volume escluso (Van der Waals) e termini elettrostatici. Vedremo in dettaglio nelle sezioni seguenti come si parametrizzano questi termini, ma anticipiamo qui che gli attuali campi di forze (FF), ottimizzati per proteine e acidi nucleici negli ultimi trenta anni circa, sono frutto della ricerca di un delicato equilibrio tra accuratezza (capacità di riprodurre strutture, dinamica e proprietà termodinamiche), trasferibilità (capacità di essere accurati in diverse condizioni ambientali esterne, temperatura, pressione etc) e semplicità (non eccessiva proliferazione del numero di parametri, che renderebbe l'approccio troppo computazionalmente costoso e quindi inutile). In ogni caso i FF non sono generalmente in grado di descrivere rotture e formazioni di legami chimici o altre variazioni strutturali che coinvolgono cambiamenti sostanziali di struttura elettronica per i quali un approccio QM è solitamente necessario.

Nel passaggio successivo da MM a CG, una buona parte di variabili atomiche viene eliminata e il sistema viene descritto tramite un insieme più ristretto di variabili interne $\{\mathbf{Q}\}$ (ad esempio, le coordinate dei $C\alpha$). L'effetto delle variabili eliminate deve essere incluso implicitamente nei campi di forze CG. Vedremo nelle sezioni successive come questo sia possibile. Anticipiamo però che le variabili eliminate si portano dietro due termini. Uno è la loro energia potenziale, che va in qualche modo mediata sulle possibili configurazioni a loro accessibili per valori fissati delle variabili CG $\{\mathbf{Q}\}$. L'altro è un fattore entropico: l'ampiezza dello spazio configurabile accessibile alle variabili eliminate dipende in generale dai valori assunti dalle variabili CG, e quindi l'entropia associata ad esse può dipendere da $\{\mathbf{Q}\}$ e può favorire certe configurazioni rispetto ad altre: il termine $-TS$ agisce come un potenziale efficace. Per questo motivo è necessario considerare la superficie di energia libera $F=E-TS$ (FES), piuttosto che la PES.

Come per la parametrizzazione dei FF all-atom, anche nel caso della parametrizzazione dei FF CG lo scopo è quello di combinare accuratezza, predittività, trasferibilità e semplicità, compito più arduo in questo caso perché un grande numero di interazioni complesse (legami a idrogeno, elettrostatica, idrofobicità, effetti di forma e variazioni di configurazione interna) devono essere tenute in conto tramite un numero relativamente piccolo di parametri. A differenza che per i FF-all-atom, l'approccio CG è relativamente nuovo, e quindi siamo ancora abbastanza lontani dall'ottimizzazione di questi campi. Inoltre, un importante effetto collaterale del coarse graining è che l'eliminazione di molte variabili provoca un "lisciamento" della FES (Fig 2.2). Questo da una parte è un vantaggio perché semplifica la dinamica, impedendo al sistema di stazionare nei molti minimi locali associati alle variabili eliminate. Ma allo stesso tempo introduce una accelerazione fittizia della dinamica, di cui va tenuto in conto.

Strettamente collegato al concetto di FES è quello di Potenziale di Forza Media (PMF). Molto spesso in un sistema si può identificare una (o un numero ristretto) di variabili collettive particolarmente interessanti perché ad esempio descrivono l'andamento di un processo. È questo il caso delle cosiddette coordinate di reazione o dei parametri d'ordine. Anche le coordinate *coarse grained* sono di questo tipo, anche se di solito non sono poche. In ogni caso, queste variabili si esprimono come una funzione delle variabili interne del sistema (ad esempio le coordinate atomiche $\{\mathbf{R}\}$: $\mathbf{Q} = \mathbf{Q}(\mathbf{R}_i)$). Il potenziale di forza media è correlato all'energia libera associata alla variazione di \mathbf{Q} , $F(\mathbf{Q})$, ricavabile dalla relazione

$$\exp\left(-\frac{F(\mathbf{Q})}{k_B T}\right) = C \int \prod_i d\mathbf{R}_i \exp\left(-\frac{E(\mathbf{R})}{k_B T}\right) \delta(\mathbf{Q} - Q(R_i))$$

dove $C = \exp(-F_0/k_B T)$, dove F_0 è l'energia libera del gas ideale, tiene conto della parte cinetica dell'energia. Nel secondo termine il fattore di Boltzmann $\exp(-E(\mathbf{R})/k_B T)$ viene integrato sulle variabili interne $\{\mathbf{R}\}$ con il vincolo che definisce la variabile collettiva \mathbf{Q} . Per capire meglio questo concetto facciamo un esempio semplice: se, ad esempio, \mathbf{Q} è la coordinata di un certo $C\alpha$ la sua definizione in termini di variabili interne sarà $\mathbf{Q} = R_\alpha = \sum_i R_i \delta(R_i - R_\alpha)$. Sostituendo questa relazione nella formula sopra si ha che $F(\mathbf{Q})$ si ottiene integrando il fattore di Boltzmann su tutte le variabili interne tranne $\mathbf{Q} = R_\alpha$. Anche se spesso il PMF $W(\mathbf{Q})$ e $F(\mathbf{Q})$ vengono considerati la stessa cosa, il PMF ha in realtà una definizione leggermente diversa, tramite la distribuzione di probabilità della variabile da cui dipende

$$W(\mathbf{Q}) = -k_B T \ln \frac{P(\mathbf{Q})}{P_0} \quad P(\mathbf{Q}) = \frac{\exp\left(-\frac{W(\mathbf{Q})}{k_B T}\right)}{Z/C}$$

formula che viene chiamata "inversione di Boltzmann", e Z è la funzione di partizione

$$Z = \exp\left(-\frac{F}{k_B T}\right) = C \int \prod_i d\mathbf{R}_i \exp\left(-\frac{E(\mathbf{R})}{k_B T}\right)$$

e $P_0 = C/Z$. Quindi, a rigore $W(\mathbf{Q})$ differisce dall'energia libera per l'energia libera del gas ideale a pari temperatura

$$W(\mathbf{Q}) = F(\mathbf{Q}) - F_0$$

distinzione del tutto ininfluyente nella maggior parte dei casi, dal momento F_0 è indipendente da \mathbf{Q} . Notiamo anche che $W(\mathbf{Q})$ non è pura energia potenziale: può contenere un termine entropico intrinseco che valuta la degenerazione dello stato $\rho(\mathbf{Q})$. In certi casi, invece, la distinzione tra FES e PMF è invece più sostanziale. Ad esempio, nel caso specifico in cui \mathbf{Q} sia una distanza r tra due particelle non legate anziché una posizione, si preferisce definire il potenziale di forza media come

$$W(r) = -k_B T \ln \frac{P(r)}{4\pi r^2} = -k_B T \ln g(r)$$

che differisce da $F(r)$ per la normalizzazione della probabilità al fattore $4\pi r^2$ anziché alla funzione di partizione. Il vantaggio è che questo $W(r)$ è direttamente legato alla funzione di distribuzione di coppia $g(r)$, tende a zero per grandi r e può essere direttamente utilizzato come una prima approssimazione di potenziale inter-particella. Vedremo nelle prossime sezioni l'uso e le applicazioni del PMF.

Infine un cenno al passaggio al continuo: la FES passa da una funzione di molte variabili a un funzionale di una funzione che descrive lo stato interno del sistema. Ad esempio, nel caso della membrana rappresentata come mezzo bidimensionale elastico, la funzione descrittiva è lo spostamento dal piano della membrana. Daremo nelle sezioni successive alcuni esempi di questo passaggio.

2.3 Equazioni del moto e campionamento dello spazio delle configurazioni

A tutti i livelli di approssimazione simulare la dinamica molecolare significa, di fatto, risolvere numericamente le equazioni dinamiche del moto per ottenere l'evoluzione delle variabili interne in funzione del tempo, $x_i(t)$. Nel caso più semplice il sistema obbedisce alla dinamica Newtoniana

$$m_i \ddot{x}_i = F_i$$

dove x_i sono le variabili interne (ad esempio $\{\mathbf{R}\}$ o $\{\mathbf{Q}\}$), m_i le masse ad esse associate e F_i le forze totali agenti su esse, comprendenti sia i campi esterni che le forze interne che una particella risente da parte di tutte le altre; F_i è ricavabile dalla P(F)ES tramite la relazione

$$F_i = -\frac{\partial E(\{x\})}{\partial x_i}$$

e quindi una volta fissata la P(F)ES e la condizione iniziale, comprendente il set di posizioni e velocità dello stato iniziale, le traiettorie si possono ricavare con una procedura ricorsiva: note posizioni e velocità ad un tempo t , quelle al tempo $t+\Delta t$ si possono ottenere tramite uno sviluppo in serie di Taylor

$$x(t + \Delta t) = x(t) + v(t)\Delta t + \frac{1}{2} \frac{F(t)}{m} \Delta t^2 + O(\Delta t^3)$$

con un errore numerico al terzo ordine in Δt . A partire dalle condizioni iniziali si va avanti così per passi successivi fino ad ottenere, una traiettoria lunga a piacere, discretizzata nel tempo con passo Δt , il *timestep*, appunto. La procedura si chiama anche integrazione delle equazioni del moto. In realtà si usa più spesso un algoritmo leggermente diverso sempre derivato dall'espansione di Taylor, detto algoritmo di Verlet

$$x(t + \Delta t) = 2x(t) - x(t - \Delta t) + \frac{F(t)}{m} \Delta t^2 + O(\Delta t^4)$$

che ha il vantaggio di introdurre errori più piccoli (al quarto ordine in Δt). Sulla base di queste formule si possono fare alcune importanti considerazioni. Gli errori numerici introdotti a ciascun passo dell'integrazione si accumulano lungo la traiettoria, e pertanto devono essere mantenuti piccoli, che implica scegliere Δt molto piccoli. Un buon criterio è che il timestep sia molto minore (almeno un ordine di grandezza) del periodo associato alla più alta frequenza del sistema

$$\Delta t \ll 2\pi / \omega_{\max}$$

In questo modo si integrano correttamente i moti più veloci del sistema e a maggior ragione quelli più lenti. Tuttavia, il costo computazionale è direttamente proporzionale al numero di timestep, quindi, fissata la lunghezza della simulazione che si vuole ottenere, è inversamente proporzionale all'ampiezza di Δt che deve quindi essere più grande possibile compatibilmente con la corretta integrazione numerica. Nelle sezioni seguenti saranno dati esempi di valori tipici di Δt .

La dinamica Newtoniana descrive l'evoluzione deterministica di un sistema ad energia totale (cinetica + potenziale), volume e numero di particelle costanti, campionando pertanto il cosiddetto insieme statistico microcanonico. Quindi la temperatura del sistema può avere delle fluttuazioni piuttosto ampie così come la pressione, il che non corrisponde alle tipiche condizioni in cui si trovano i sistemi reali, in cui sono piuttosto mantenuti costanti temperatura e pressione, mentre fluttuano il volume e l'energia. Per risolvere questi problemi si utilizzano degli accorgimenti che simulano la presenza di termostati e barostati. Il metodo più semplice per mantenere costante la temperatura è riscaldare ad ogni passo tutte le velocità di un fattore costante $v \rightarrow \lambda v$. Nell'approssimazione più semplice si può porre $\lambda^2 = T_0/T$ per riportare ad ogni passo la temperatura verso il valore di riferimento T_0 . Questo metodo però introduce variazioni troppo brusche delle velocità e può portare ad una dinamica troppo distorta. Nel cosiddetto termostato di Berendsen il sistema viene invece accoppiato ad un bagno termico che rilassa la temperatura verso T_0 con una dinamica del primo ordine con tempo di rilassamento τ . Questo equivale ad applicare un fattore di riscaldamento delle velocità pari a

$$\lambda^2 = 1 + \frac{\Delta t}{\tau} \left(\frac{T_0}{T} - 1 \right)$$

che dipende dal timestep. τ è un tempo efficace di termalizzazione. Se $\tau = \infty$ si riottiene il campionamento microcanonico, mentre per $\Delta t = \tau$ si ritrova il semplice metodo di riscaldamento precedente. Nella pratica valori di τ pari a 10 o 100 volte il timestep assicurano una rapida termalizzazione con fluttuazioni dell'energia sufficientemente realistiche. Tuttavia ancora la termodinamica del sistema non è realistica, nel senso che la distribuzione statistica delle fluttuazioni dell'energia non è quella descritta dall'insieme statistico canonico (distribuzione di Boltzmann delle energie), che sarebbe quello corretto per simulazioni a temperatura costante. Questo risultato si ottiene con il termostato di Nose-Hoover, in cui si accoppia al sistema in maniera Lagrangiana un nuovo grado di libertà che descrive le fluttuazioni del termostato. In pratica questo equivale ad applicare al sistema una nuova equazione del moto con un termine di tipo viscoso (cioè proporzionale alle velocità), ma con coefficiente di attrito γ dipendente dal tempo, la cui variazione è regolata dal discostamento della temperatura da quella di riferimento:

$$m_i \ddot{x}_i = F_i - m_i \gamma \dot{x}_i \quad \dot{\gamma} = \frac{-Nk_B}{Q} T(t) \left\{ \frac{T_0}{T(t)} - 1 \right\} = -\frac{1}{\tau_{NH}^2} \left\{ \frac{T_0}{T(t)} - 1 \right\}$$

Si noti che il termine di attrito, proprio perché ha una sua dinamica e non è costante può sia sottrarre che fornire energia al sistema. Il parametro di accoppiamento tra sistema e termostato è Q , che fa la parte di una "massa" termica del termostato (nel senso che descrive l'inerzia del termostato a scambiare energia), anche se è più intuitivo riferirsi al tempo di rilassamento $\tau_{NH}^2 = Q/Nk_B T$, che assume un ruolo simile a quello del termostato di Berendsen. Il vantaggio del termostato di Nose-Hoover è il corretto campionamento dell'insieme statistico Canonico, che significa che sono accuratamente riprodotte le proprietà termodinamiche. I barostati vengono trattati in maniera analoga ai termostati, ma la variabile considerata in questo caso è la pressione, invece della temperatura.

Le dinamiche che abbiamo visto fino a qui sono di tipo deterministico, cioè la traiettoria è interamente determinata dalle condizioni iniziali. In alcune situazioni può risultare necessario utilizzare dinamiche di tipo stocastico, ovvero che introducono elementi aleatori nella traiettoria. È questo il caso ad esempio quando si vuole simulare la dinamica di particelle che si muovono in un solvente senza dover simulare esplicitamente i gradi di libertà del solvente, computazionalmente dispendiosi. Gli urti delle molecole del solvente sulla particella si possono rendere aggiungendo all'equazione di Newton oltre al solito termine viscoso (eventualmente dipendente dalla natura della particella stessa) un termine di forza casuale (stocastica) F_i^s :

$$m_i \ddot{x}_i = F_i - m_i \gamma \dot{x}_i + F_i^s$$

Questa è l'equazione di Langevin. Il termine viscoso sottrae energia, mentre il termine stocastico la aggiunge. Per avere energia media costante deve valere la seguente relazione tra la funzione di correlazione temporale della forza stocastica e il coefficiente di attrito

$$\langle F_i^s(0) F_i^s(t) \rangle = 2m_i k_B T \gamma_i \delta_{ij}(t)$$

dove la presenza della delta di Dirac al secondo termine indica che gli urti sono temporalmente scorrelati tra loro. In pratica la dinamica di Langevin agisce come una sorta di termostato (in cui il serbatoio di calore è costituito dal solvente stesso) e, dipendentemente dal valore di γ ha anche l'effetto di smorzare i moti veloci, per cui è di solito possibile scegliere timestep più grandi che nella semplice dinamica Newtoniana. Per valori molto grandi di γ è anche possibile trascurare il termine di secondo ordine nell'equazione di Langevin che si riduce a

$$dx_i = \frac{D_i}{k_B T} F_i dt + f_i^s$$

ovvero l'equazione della dinamica Browniana. D è il coefficiente di diffusione, legato al coefficiente d'attrito da

$$\gamma = k_B T / mD$$

e si può stimare su base empirica conoscendo il coefficiente di viscosità del solvente η e il raggio della particella r_i

$$D_i = \frac{k_B T}{6\pi\eta(r_i + 1.4 \text{ \AA})}$$

La dinamica Browniana ha l'ulteriore vantaggio che, essendo una dinamica del primo ordine, l'integrazione dell'equazione del moto è più efficiente. Senza entrare in dettagli, l'ampiezza del timestep può di norma essere aumentata di circa un ordine di grandezza rispetto alla dinamica Newtoniana.

Le equazioni di dinamica stocastica e in particolare l'equazione di Langevin hanno un'applicabilità generale a tutti quei casi in cui si voglia simulare realisticamente la dinamica di sistemi in cui alcune variabili siano state eliminate. Le variabili eliminate possono essere di qualsiasi natura, non necessariamente solo gradi di libertà del solvente, ma anche ad esempio, le variabili interne alla "bead" nei modelli Coarse Grained. Abbiamo già accennato al fatto che in questi sistemi, poiché la FES viene "lisciata" si ha un'accelerazione fittizia. Tuttavia la dinamica dei gradi di libertà

eliminata può essere reintrodotta proprio tramite l'equazione di Langevin, meglio se nella sua versione generalizzata

$$m\ddot{x}_i(t) = F_i(t) - \int_0^t dt' M_i(t-t')\dot{x}_i(t') + F_i^s(t) \quad \langle F_i^s(0)F_j^s(t) \rangle = 2m_i k_B T M_i(t)\delta_{ij}$$

dove si è introdotta la possibilità che la correlazione temporale tra le forze stocastiche si estenda su intervalli di tempo non infinitesimi (processi con memoria, la $M_i(t)$ viene anche detta funzione di memoria). La dinamica dei gradi delle variabili interne alla bead è nascosta nella funzione di memoria che si può stimare in certi casi con semplici approssimazioni. Ad esempio, la più semplice approssimazione per le variabili interne ad una bead è approssimarle con oscillatori armonici. La forza stocastica sarà la somma delle forze che questi oscillatori esercitano sulla bead stessa. Con un po' di algebra e usando la seconda delle equazioni sopra si arriva a

$$M_i(t) = \sum_I a_i(\omega_I) \cos(\omega_I t)$$

dove I scorre sulle variabili eliminate. A seconda della dipendenza dei coefficienti a dalle frequenze, a sua volta dipendente dal tipo di dinamica interna, la funzione di memoria può assumere diverse forme, anche quella di una delta di Dirac, che riporta all'equazione di Langevin semplice. Questa trattazione ci mostra che la reintroduzione della dinamica dei gradi di libertà nascosti di fatto porta un termine di smorzamento e un termine di forza stocastica, che come abbiamo già detto hanno l'effetto globale di "rallentare" la dinamica. In pratica l'uso della dinamica di Langevin nei modelli CG contrasta l'accelerazione fittizia e ci restituisce una dinamica realistica del sistema, a patto che la funzione di memoria e la forza stocastica assumano valori opportuni, valutabili come abbiamo visto tramite ipotesi sulla dinamica delle variabili eliminate. Di fatto, la valutazione è piuttosto complessa e porta a risultati comunque approssimati, quindi in generale risulta più conveniente usare l'equazione di Langevin semplice e usare γ come parametro di fit fino a che siano riprodotte nella dinamica simulata costanti di tempo note (ad esempio, tempi di rilassamento).

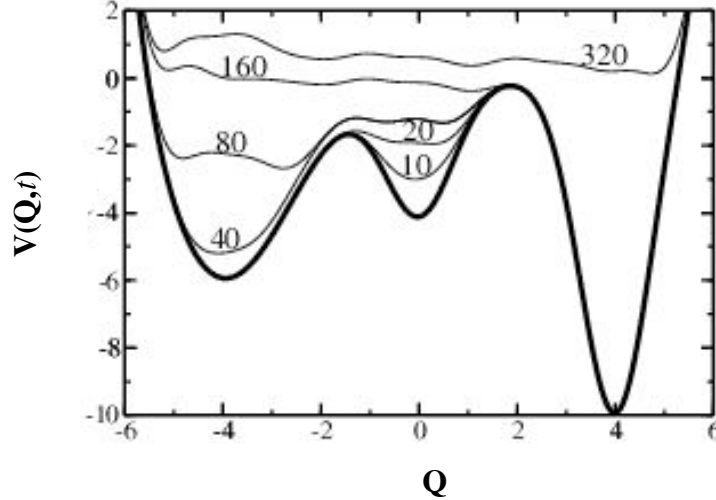


Fig 2.3 Illustrazione della metadinamica. La linea spessa è il PMF della variabile collettiva Q . A intervalli di tempo regolari (i numeri sulle linee fini indicano il valore del timestep) vengono aggiunte al potenziale delle "colline" (gaussiane) centrate sul valore di Q in quell'istante. Quindi il sistema viene spinto fuori dalle buche di potenziale con un potenziale artificiale dipendente dal tempo, $\Delta V(Q, t)$. Per tempi sufficientemente lunghi lo spazio delle configurazioni sarà accuratamente campionato, il potenziale totale $V(Q, t)$ sarà circa piatto e $-\Delta V(Q, t) \sim F(Q)$.

In altri casi l'interesse non sta nel riprodurre accuratamente la dinamica temporale, ma nel campionare più esaustivamente possibile lo spazio delle configurazioni del sistema. In questo caso l'accelerazione della dinamica è un effetto collaterale utile della procedura di coarse graining. Questo ragionamento ci porta ad un altro parametro che determina la scelta del tipo di dinamica da usare, e cioè l'efficienza di campionamento dello spazio delle configurazioni (o delle fasi, se ci interessa anche la distribuzione degli impulsi associati alle variabili interne). Un adeguato

campionamento è infatti necessario per riprodurre correttamente la distribuzione delle variabili all'equilibrio, in altre parole, per riprodurre la corretta termodinamica del sistema. Alcuni metodi risultano più efficaci di altri nel superare barriere di potenziale e permettere un'esplorazione più esaustiva del sistema. Abbiamo già visto che mentre la dinamica Newtoniana non permette fluttuazioni dell'energia totale e quindi di fatto impedisce il superamento di barriere di energia maggiori dell'energia totale, i termostati (specialmente quello di Nose-Hoover) risolvono almeno in parte il problema permettendo fluttuazioni dell'energia. Un effetto analogo lo hanno le dinamiche stocastiche. Esistono poi una serie di metodi di dinamica deterministica appositamente studiati per superare le barriere di potenziale. Ne descriviamo di seguito uno – la metadinamica – che ha il pregio di avere una chiara interpretazione in termini fisici e di illustrare alcuni concetti generali.

La metadinamica viene di solito applicata non a tutte le variabili interne, ma ad un numero ristretto di variabili collettive $\mathbf{Q} = Q(R_i)$ ritenute rilevanti per descrivere l'evoluzione del sistema (coordinate di reazione). Allo scopo di far fuggire il sistema eventualmente intrappolato in minimi locali, che appaiono come buche profonde nella $F(Q)$, si aggiunge all'energia potenziale del sistema e ad intervalli di tempo discreti t_i il valore

$$\Delta V(Q(R_i), t) = A \sum_{t=t_i} \exp\left(-\frac{(Q(R_i) - Q(R_i(t)))^2}{2\sigma^2}\right)$$

cioè una gaussiana di opportuna ampiezza e larghezza, localizzata sul valore di \mathbf{Q} che il sistema in quel momento assume. Poiché le gaussiane vengono aggiunte ad intervalli di tempo regolari e il sistema occupa più a lungo buche di potenziale più profonde, l'effetto di $\Delta V(Q(R_i), t)$ è quello di spingere il sistema fuori dalle buche e permettergli di campionare altre zone dello spazio delle configurazioni. Come mostrato in Fig 2.3 $\Delta V(Q(R_i), t)$ con il passare del tempo "riempie" le buche del PMF fino a "livellarle". Quindi per tempi sufficientemente lunghi l'energia potenziale totale $V + \Delta V(Q(R_i), t)$ sarà all'incirca piatta, e $\Delta V(Q(R_i), t)$ corrisponderà al negativo di $F(Q)$, fornendo quindi anche un modo pratico per calcolare $F(Q)$.

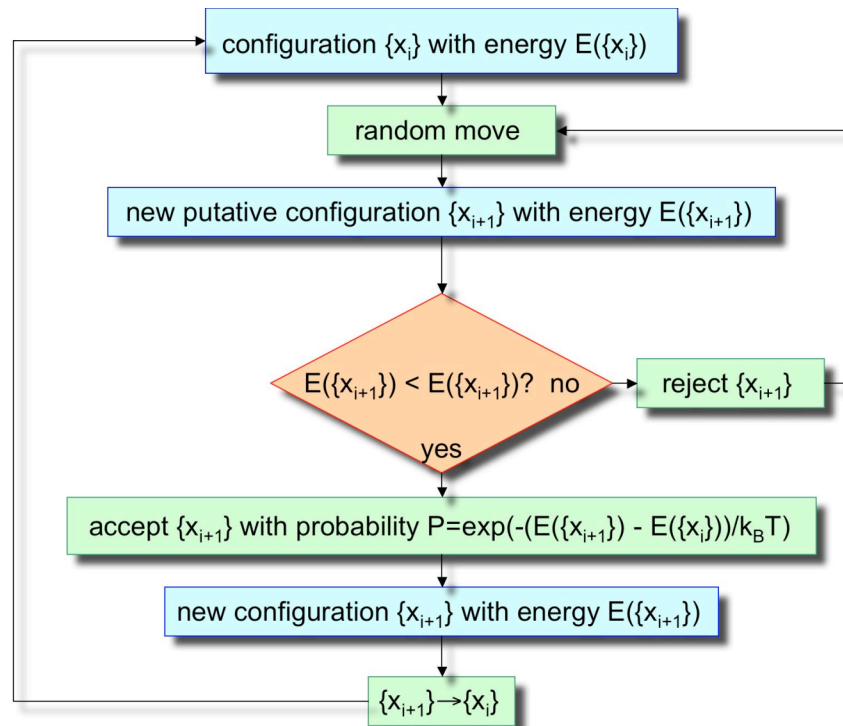
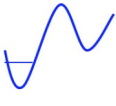
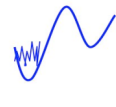
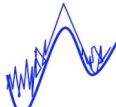

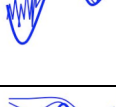
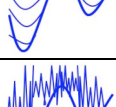


Fig 2.4 Diagramma di flusso dell'algoritmo Metropolis

Se da una parte la metadinamica è un valido sistema per ottenere un adeguato campionamento dello spazio delle configurazioni, dall'altra l'introduzione di un potenziale esterno dipendente dal tempo modifica la dinamica del sistema che quindi non può più essere considerata realistica. Questo è un problema generale: i metodi che consentono una dinamica realistica solitamente non

danno un adeguato campionamento, quelli che consentono un buon campionamento (e quindi restituiscono una accurata termodinamica) alterano la dinamica del sistema. Un esempio estremo di questo principio sono i metodi completamente stocastici come il MonteCarlo. In questo caso l'unico scopo è il corretto campionamento dello spazio delle configurazioni, in modo che vengano riprodotte le distribuzioni di determinati insiemi statistici e quindi la corretta termodinamica. L'algoritmo Metropolis, creato specificamente per il campionamento dell'insieme Canonico, è illustrato nella Fig. 2.4. Ogni configurazione viene generata a partire dalla precedente con una mossa "casuale" nello spazio delle configurazioni, e accettata con probabilità proporzionale al fattore di Boltzmann derivato dalla differenza di energia delle due configurazioni. Il campionamento è tanto più accurato quanto più la "traiettoria" è lunga. Si noti che il termine "traiettoria" per l'insieme delle configurazioni successive x_i non è del tutto appropriato, perché due mosse successive non sono collegate tra loro dall'evoluzione guidata da forze. Quindi non è possibile estrarre tempi caratteristici da una traiettoria di questo tipo.

Le principali caratteristiche dei metodi di campionamento dello spazio delle configurazioni descritti sono riassunte in Tab 2.1.

Method	Features		Typical applications
Newton dyn	<ul style="list-style-type: none"> ➤ Deterministic ➤ Realistic time dynamics ➤ Microcanonical ensemble ➤ Inefficient conf space sampling 		All-atom descriptions Only explicit degrees of freedom
Thermostats and Barostats	<ul style="list-style-type: none"> ➤ Deterministic ➤ Realistic time dynamics ➤ Canonical ensemble ➤ More efficient sampling 		All-atoms + implicit reservoir (thermal or baric bath)
Langevin dyn	<ul style="list-style-type: none"> ➤ Partially stochastic ➤ Realistic time dynamics (with proper parameters) ➤ Solvent effects ➤ More efficient sampling ➤ Realistic statistical ensemble 		Implicit solvent and/or CG descriptions with implicit internal degrees of freedom
Brownian dyn	<ul style="list-style-type: none"> ➤ Stochastic ➤ Realistic time dynamics (with proper parameters) ➤ Solvent effects ➤ More efficient sampling 		Implicit solvent and/or CG descriptions with implicit internal degrees of freedom
Meta dynamics	<ul style="list-style-type: none"> ➤ Deterministic ➤ More efficient sampling ➤ Unrealistic time dynamics 		All cases
Monte Carlo	<ul style="list-style-type: none"> ➤ Stochastic ➤ Very efficient sampling ➤ Unrealistic time dynamics 		All cases

Tab 2.1 Riassunto delle principali caratteristiche dei metodi di campionamento dello spazio delle configurazioni. I metodi sono ordinati per crescente efficienza di campionamento.

2.4 Approcci multi-scala, filosofie di parametrizzazione e coerenza tra i livelli

A conclusione di questa prima sezione alcune considerazioni generali sui metodi multi-scala. Gli approcci multi-scala si possono applicare sia "in parallelo" che "in serie"⁴. "In parallelo" significa usare due (o più) descrizioni a diversa risoluzione contemporaneamente sullo stesso sistema. Questo approccio è utile ogni volta che sia possibile isolare una parte del sistema che necessita una descrizione più accurata dal resto, che può essere trattato ad un livello di approssimazione maggiore. Un esempio tipico sono le proteine, che di solito hanno un sito attivo in cui avviene una reazione chimica e che va quindi trattato con metodi QM, mentre il resto della proteina assume spesso solo ruoli strutturali e quindi si può trattare con metodi MM. Analogamente in alcuni casi

sono stati considerati approcci ibridi⁵, MM/CG o MM/CG/Continuo. Nelle prossime sezioni verranno presentati esempi illustrativi. Qui ribadiamo che in questo caso un requisito fondamentale è che le varie descrizioni siano coerenti tra loro. Più specificamente è necessaria una coerenza meccanica tra le forze calcolate con i due metodi diversi per le interazioni che attraversano le interfacce. Un metodo per ottenere questo tipo di coerenza è il *Force Matching* che consiste nel fittare i parametri dei FF del metodo a bassa risoluzione sulle forze calcolate su traiettorie di dinamica ottenuta con il metodo più accurato. Questo metodo si usa sia per ottimizzare i campi MM su calcoli QM, sia per fittare i campi CG su simulazioni MM, ed equivale evidentemente al fit delle P(F)ES di cui si è già parlato. Il metodo ha due limitazioni: la prima è che si opera un fit su dati che provengono essi stessi da una simulazione, e quindi contengono approssimazioni. La seconda è che la trasferibilità del FF ottenuto dipende da quanto accurato è il campionamento del metodo ad alta risoluzione (lunghezza della traiettoria). Poiché esso è anche più computazionalmente dispendioso, in molti casi bisogna suddividere il sistema in pezzi piccoli per adottare questo metodo. Osserviamo infine che mentre i metodi multi-scala in parallelo portano un guadagno sostanziale nel costo e quindi consentono di trattare sistemi delle dimensioni tipiche del metodo "meno accurato", non si ha invece un analogo vantaggio nel calcolo delle proprietà dinamiche. Come abbiamo visto, infatti, il timestep è limitato dalle frequenze più alte del sistema, che sono in questo caso quelle pertinenti alla descrizione più accurata. Quindi il timestep, e conseguentemente la lunghezza totale della traiettoria, sarà quello tipico del metodo "meno accurato". Si simulano sistemi grandi dunque, ma su brevi scale temporali.

Nelle metodologie multi-scala "in serie" si applicano, una dopo l'altra, descrizioni a diversa risoluzione all'intero sistema. Col metodo a bassa risoluzione si possono simulare sistemi molto grandi su scale temporali lunghe. Dalle traiettorie si possono poi estrarre campioni di configurazioni da usare come punto di partenza per dinamiche (più corte) operate con metodi più accurati. Anche in questo caso ovviamente la coerenza tra le descrizioni è un requisito fondamentale, ma qui il modo più naturale per ottenerla è il matching "termodinamico" dei due sistemi, espresso dall'equazione

$$Z = \int dQ \exp\left(-\frac{F(Q)}{k_B T}\right) = \int dq \exp\left(-\frac{E(q)}{k_B T}\right)$$

dove $F(Q)$ ed $E(q)$ sono le P(F)ES del sistema a bassa e ad alta risoluzione, rispettivamente, e $\{Q\}$ e $\{q\}$ le loro variabili interne (è sottinteso che si tratta di integrali multipli sui due set di variabili e sui corrispondenti impulsi). Questa è la stessa equazione che definisce anche il potenziale di forza media (vedere sopra). Se fosse possibile scegliere le variabili interne di un sistema in modo che la FES fosse una somma di singoli termini ciascuno dipendente da una sola di esse (gradi di libertà completamente scorrelati), allora si avrebbe

$$F(Q) = \sum_i W(Q_i) + F_0$$

e i PMF $W(Q_i)$ di singola variabile ottenuti dall'inversione di Boltzmann (vedi sopra) corrisponderebbero esattamente ai termini additivi del potenziale totale. In generale però le variabili interne non sono mai disaccoppiate, e quindi il PMF e il termine di potenziale associato alla variabile $V(Q_i)$ differiscono per termini che contengono le correlazioni tra le variabili. $V(Q_i)$ viene calcolato a partire da $W(Q_i)$ con approssimazioni successive. Nelle prossime sezioni verranno dati alcuni esempi. La procedura è piuttosto complessa, ma si può dimostrare che, se portata avanti esattamente, è equivalente al force matching. In altre parole consistenza termodinamica e consistenza meccanica in linea di principio sono equivalenti, e danno vita a campi di forze uguali, se entrambe applicate esattamente e utilizzando lo stesso set di dati come input (ad esempio dati da simulazioni ad alta risoluzione). Di fatto però in entrambi i casi si usano approssimazioni, che sono diverse, e molto spesso si usano set di dati differenti. Ad esempio il PMF si calcola di solito a partire da distribuzioni di probabilità generate da set di dati sperimentali, che hanno il vantaggio, rispetto ai dati generati da simulazioni ad alta risoluzione, di non contenere essi stessi modellizzazioni e di poter esplorare, in linea di principio, uno spazio configurazionale più ampio.

3. Esempi: “zoomata” sulle tecniche di modellizzazione

In questa sezione sono riportati, livello per livello, approfondimenti dei concetti generali espressi nella sezione precedente ed esempi illustrativi su sistemi biomolecolari.

3.1 Livello di “piccola molecola”: simulazioni QM

Per “piccola molecola” si intende un insieme di atomi (di solito da 10 a 100) che ha una parte importante nel sistema in esame: la parte chimicamente attiva, il sito attivo, un ligando, che abbiano necessità di una trattazione quantomeccanica. Anche restringendoci all'approssimazione di BO (si veda sez 2.2), sono stati utilizzati nel corso degli anni innumerevoli diversi metodi QM per trattare i sistemi biologici, caratterizzati da accuratezze diverse. Uno che è emerso nel corso degli anni come compromesso ottimale tra accuratezza e contenuto costo di calcolo è il metodo Car-Parrinello. La parte elettronica del problema viene affrontata in uno schema DFT (Teoria del funzionale di Densità) che riduce il problema multi-elettronico ad un problema di singoli elettroni disaccoppiati: ciascun elettrone si comporta come fosse singolo, ma immerso in un potenziale esterno dovuto alla nube elettronica formata dagli altri, oltre che del potenziale nucleare. Il risultato finale è che la singola equazione di Schrödinger a N elettroni si separa in N equazioni per elettroni singoli, con funzione d'onda ϕ

$$E[\mathbf{R}, \phi] \phi = \left\{ -\frac{1}{2} \hbar^2 \nabla^2 + U[\mathbf{R}, \phi] \right\} \phi = \varepsilon \phi$$

dove $U[\mathbf{R}, \phi]$ è l'energia potenziale ed $E[\mathbf{R}, \phi]$ l'energia totale dell'elettrone. La ϕ dello stato fondamentale minimizza $E[\mathbf{R}, \phi]$, e più in generale tutti gli stati si possono trovare come minimi con la aggiuntiva condizione che siano ortonormali. Partendo da questo, Car e Parrinello hanno dato una formulazione alternativa all'equazione di Schrödinger, trasformandola in una equazione di evoluzione di tipo Newtoniano⁶

$$\mu \ddot{\phi} = - \frac{\partial E[\mathbf{R}, \phi]}{\partial \phi} + \varepsilon \phi$$

dove la ϕ assume il ruolo di una variabile classica con una massa μ ad essa associata. La derivata parziale di E rispetto a ϕ esprime il principio di minimo e gli autovalori ε compaiono come moltiplicatori di Lagrange per l'ortonormalità degli stati. Per $\mu=0$ questa equazione è esattamente equivalente all'equazione di Schrödinger, e il vantaggio di scriverla in questa forma è che può essere facilmente accoppiata all'equazione del moto (Newtoniana) dei nuclei

$$M \ddot{\mathbf{R}} = - \frac{\partial E[\mathbf{R}, \phi]}{\partial \mathbf{R}}$$

in un unico sistema “classico” accoppiato nuclei+elettroni. Nella pratica, per risolvere il moto, si parte da una configurazione nucleare data \mathbf{R}_0 e si calcola la ϕ_0 risolvendo l'equazione di Schrödinger. Poi si evolve il sistema secondo il sistema di equazioni “classiche” accoppiate come sopra descritto, usando uno qualsiasi degli algoritmi e metodi descritti in sez. 2.3. Notiamo che ai passi successivi al primo non ci sarà più bisogno di risolvere l'equazione di Schrödinger degli elettroni, perché questi seguiranno adiabaticamente il moto nucleare, oscillando vicino al minimo della PES man mano che il sistema evolve (Fig 3.1). Questo è un punto cruciale, perché rende il metodo Car-Parrinello più computazionalmente snello rispetto ad altri. L'adiabaticità è regolata da μ che per questo (e per non discostarsi troppo dalla “reale” equazione di Schrödinger) deve essere molto piccola. D'altra parte, μ determina il valore delle frequenze più alte del sistema (quelle elettroniche) e quindi è strettamente correlata al timestep. Valori ottimali di μ portano a timestep dell'ordine di 0.1fsec e consentono la simulazione di sistemi di 10-100 atomi per tempi dell'ordine di 10-100 ps.

In Fig 3.2 sono riportati due esempi di simulazioni Car-Parrinello. Entrambi coprono tempi di decine di ps e descrivono trasferimenti protonici. Fig 3.2 (a) riporta la simulazione di una “scatola” di molecole di acqua (una cinquantina) e mostra il trasferimento di un protone attraverso una rete di legami a idrogeno da una molecola di acqua all'altra con la formazione temporanea di ioni idronio. Nella sequenza a fondo nero sono anche rappresentati gli orbitali coinvolti nel passaggio. In (b) è mostrato il trasferimento protonico che si presume possa avvenire nel sito attivo della

GFP dalla Glu222 all'anello eterociclico del cromoforo a formare lo stato zwitterionico, ritenuto coinvolto nel blinking. La simulazione indicherebbe che il trasferimento possa avvenire attraverso la Ser65 in uno stato eccitato del cromoforo, anche se fin'ora non ci sono evidenze sperimentali dirette dell'esistenza dello stato zwitterionico.

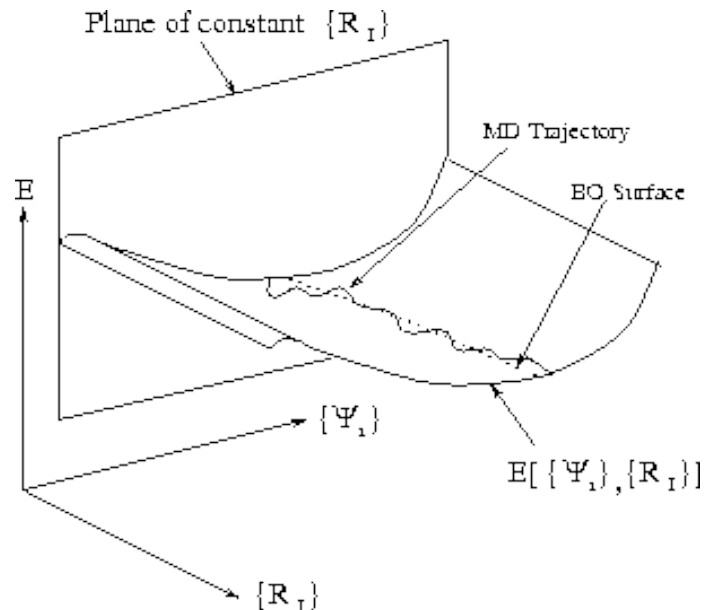
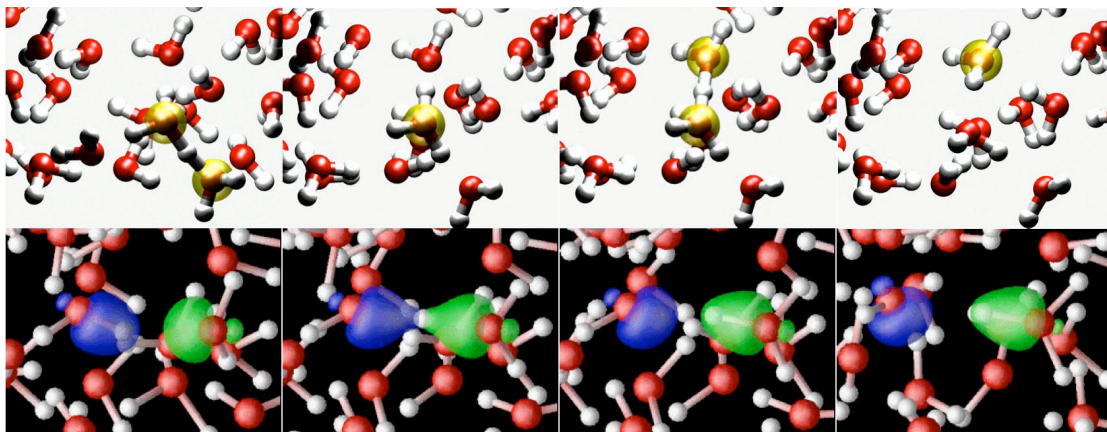
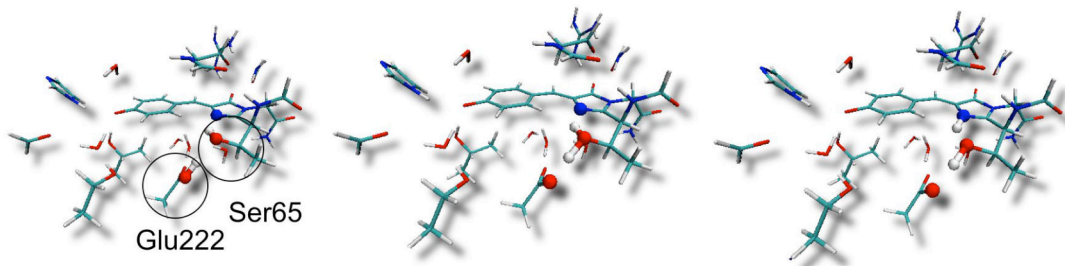


Fig 3.1 Illustrazione schematica del metodo Car-Parrinello.



(a)



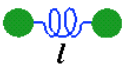
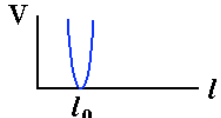

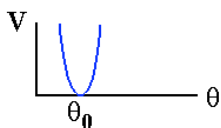
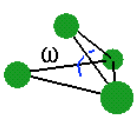
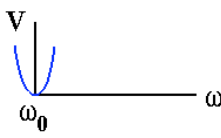
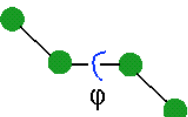
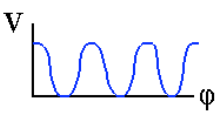

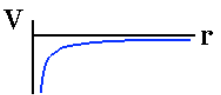

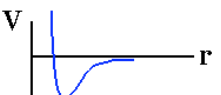
(b)

Fig 3.2 Esempi di simulazioni Car-Parrinello (a) Trasferimenti protonici e formazioni di ioni idronio in acqua. Sotto, con sfondo in nero, sono mostrati anche gli orbitali coinvolti. (b) Trasferimento protonico dalla Glu 222 al cromoforo a formare lo zwitterione in GFP.

3.2 Livello di Bio-molecola: simulazioni MM

Il livello di biomolecola comprende sistemi delle dimensioni di intere proteine. Includendo una completa idratazione, si arriva facilmente su ordini di 10000-100000 atomi e dimensioni lineari

intorno a 10nm. Questi sistemi si affrontano, come detto, approssimando la PES con una somma di termini analitici, globalmente chiamata Campo di Forza (FF). Esiste un'enorme varietà di FF, adatti ai più disparati sistemi fisici, che differiscono per la forma e il numero di termini e per la parametrizzazione. A sua volta la parametrizzazione può differire per la quantità di "tipi chimici" di atomi definiti nel campo di forze e per la filosofia con cui i valori numerici dei parametri vengono determinati. Facciamo alcuni esempi. Solitamente i FF per i sistemi inorganici hanno un numero limitato di "tipi" di atomi. Il "tipo" è definito sia dall'elemento chimico (ad es, carbonio o idrogeno), sia dall'ambiente chimico circostante in cui l'atomo si trova (ad esempio, un carbonio planare del benzene avrà ovviamente un "tipo chimico" diverso dal carbonio tetraedrico del metano, ma anche dal carbonio planare della grafite, perché legato a due carbonii e un idrogeno, mentre nella grafite ogni carbonio è legato solo a carbonio). Quindi, siccome nei sistemi inorganici solitamente la varietà e la topologia di legame sono molto più semplici che nei sistemi biologici, il numero di tipi di atomi è molto minore. Però questi sistemi sono spesso caratterizzati da interazioni in cui la correlazione a molti corpi è molto rilevante, e per questo necessitano forme funzionali molto complesse, con termini spesso non semplicemente a singola variabile o additivi. Al contrario, nei sistemi biologici il numero di tipi di atomi è dell'ordine delle centinaia (nonostante il numero di componenti chimiche sia essenzialmente 4 – C H O ed N), ma le forme funzionali possono essere mantenute relativamente semplici perché la complessità delle interazioni è principalmente imputata ad effetti chimico-fisici locali. I campi di forza più comunemente usati per proteine e acidi nucleici sono Amber, Charmm e Gromos, e hanno i termini elencati in Tab 3.1.

$\sum_{\text{bonds}} k_b (l - l_0)^2$		
$\sum_{\text{angles}} k_\theta (\theta - \theta_0)^2$		
$\sum_{\text{imp dih}} k_\omega (\omega - \omega_0)^2$		
$\sum_{\text{dihedrals}} k_\phi [1 + \cos(n\phi + \delta)]$		
$\sum_{\text{non-bonded pairs AB}} \frac{1}{4\pi\epsilon_0} \frac{q_A q_B}{r_{AB}}$		
$\sum_{\text{non-bonded pairs AB}} \epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{r_{AB}} \right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}} \right)^6 \right]$		

Tab 3.1 Forme analitiche tipiche e significato fisico dei termini più comuni nei campi di forze empirici per biomolecole.

Questi campi sono stati ottimizzati negli ultimi decenni, e hanno un'accuratezza globale circa equivalente. Tuttavia presentano alcune differenze e possono quindi riprodurre con diversa accuratezza certe classi di proprietà. Anzitutto ci sono differenze (non sostanziali) nel numero di tipi di atomi definiti, nella forma funzionale dei diedri impropri, nelle regole di composizione dei parametri per le interazioni Van der Waals. Inoltre, i diversi campi sono stati parametrizzati seguendo delle filosofie leggermente diverse. In ogni caso, i parametri vengono fittati su dati

relativi a piccole molecole, in fase gassosa o condensata che possono essere di varia natura: geometria, frequenze vibrazionali, dati energetici o termodinamici, come barriere di attivazione, calori di vaporizzazione, energie libere di solvatazione, e di origine sperimentale o teorica⁷. I dati di origine teorica sono generalmente geometrie, frequenze vibrazionali oppure interi tratti di PES per fittare barriere e minimi di energia, come abbiamo visto nelle sezioni precedenti. I campi differiscono per le classi di dati su cui viene operato il fit (Tab 3.2). È abbastanza difficile dire, in assoluto, quale dei tre campi sia il migliore: hanno diverse accuratèzze a seconda del sistema e del tipo di condizioni in cui si simula. In molti casi essi danno risultati essenzialmente equivalenti, e la scelta può essere guidata da motivi di tipo pratico, ad esempio la maggiore o minore maneggevolezza dei codici che li implementano.

	Bonded interactions	Van der Waals	Coulomb cariche parziali
Amber http://amber.scripps.edu/ UCSF	da dati geometrici e vibrazionali sperimentali Barriere e profili di PES fittati su dati QM	Riproduzione di densità e calori di vaporizzazione di vari liquidi in fase condensata	Fittate su potenziali elettrostatici da calcoli teorici
CHARMM http://www.charmm.org/ Harward	da dati geometrici e vibrazionali sperimentali e QM	Riproduzione di densità e calori di vaporizzazione di vari liquidi in fase condensata	Fittate su energie di dimerizzazione di piccole molecole da calcoli QM
Gromos http://www.igc.ethz.ch/gromos/ University of Groningen and the ETH Zurich	da dati geometrici e vibrazionali sperimentali e QM	Come sopra, ma vengono incorporate anche le polarizzabilità atomiche sperimentali	Fittate su dati energetici e termodinamici sperimentali in fase condensata, come il parm vdw

Tab 3.2 Caratteristiche dei tre principali campi di forze per bio-molecole.

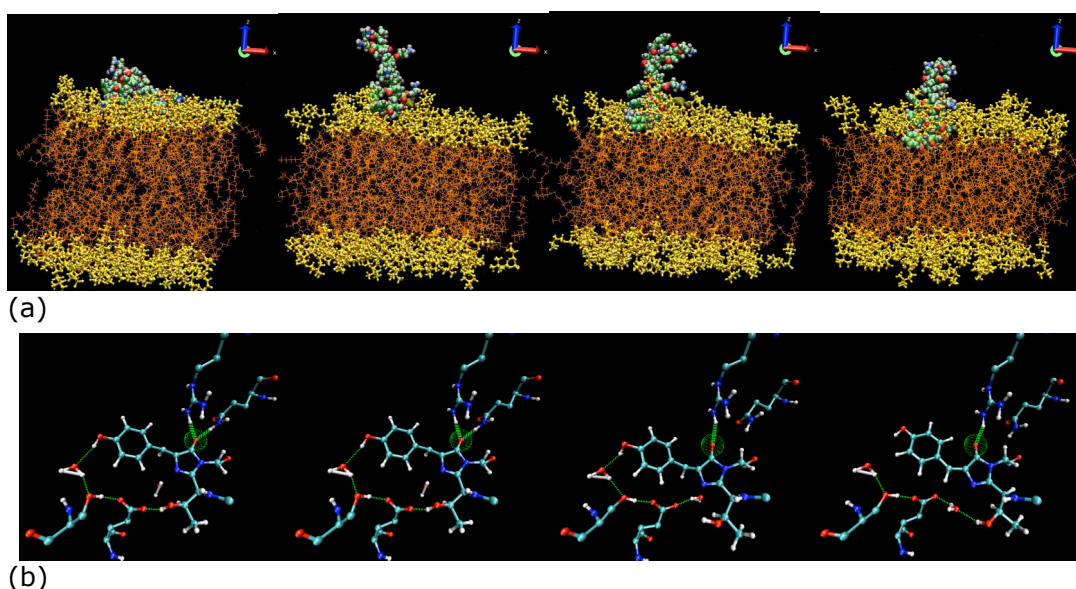
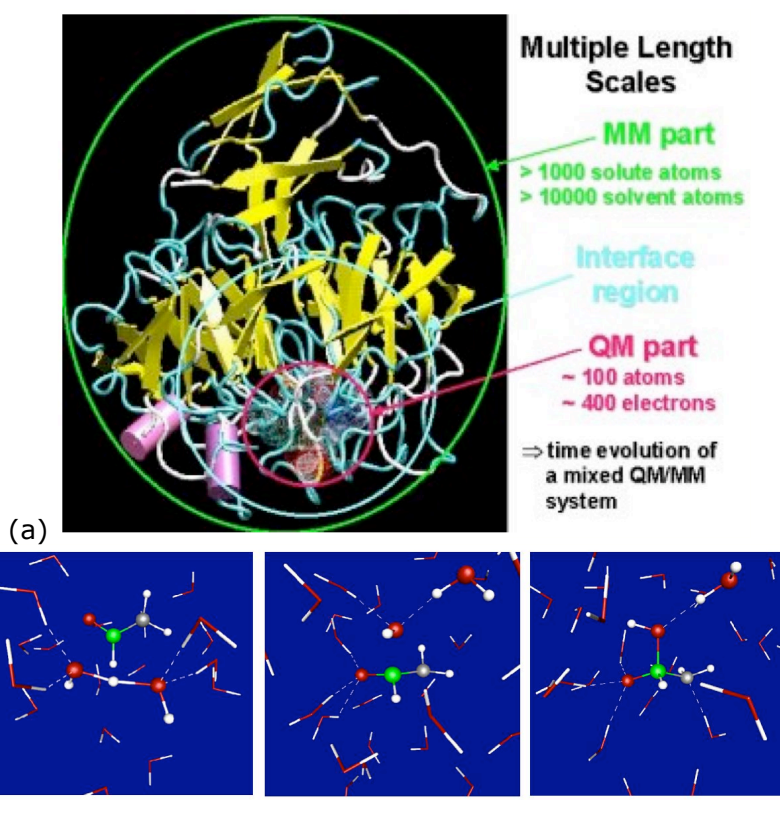


Fig 3.3 Esempi di simulazioni di dinamica molecolare classica. (a) Molecola antimicrobica che penetra nella membrana cellulare (b) cromoforo della GFP che isomerizza dallo stato cis allo stato trans.

Nelle simulazioni di dinamica molecolare classica con i FF all-atom, il timestep è limitato dalle frequenze di vibrazione dei legami coinvolgenti atomi di idrogeno, e tipicamente viene scelto

intorno a 1 fsec. Con questa configurazione e per sistemi grandi quanto un'intera proteina idratata, run di 10-100 nsec necessitano tempi macchina dell'ordine di mesi. In Fig. 3.3 sono riportati un paio di esempi tipici di simulazioni che coprono tempi dell'ordine di decine di nsec. Nel caso del farmaco peptide-mimetico che entra la membrana, sono rappresentati esplicitamente oltre al farmaco, un tratto di membrana e l'acqua circostante (dentro e fuori dalla membrana), non mostrata per chiarezza. Nel caso dell'isomerizzazione cis-trans del cromoforo è simulata esplicitamente tutta la proteina con uno strato di idratazione di diversi Å, anche se in figura è mostrato solo il sito attivo per chiarezza. Inoltre in questo caso, siccome l'isomerizzazione avviene in seguito ad eccitazione, è stato necessario riparametrizzare alcuni termini del campo di forze del cromoforo per rappresentare la PES dello stato eccitato⁸.

Fig 3.4 Simulazioni ibride QM/MM. A sinistra, uno schema della suddivisione del sistema per un approccio ibrido ad una proteina. Sono evidenziate le diverse zone QM, MM e interfaccia. Sotto, simulazione dell'attacco nucleofilo alla formammide. Qui la parte QM è costituita dalla formammide e dalle due molecole di acqua rappresentate come ball&sticks.

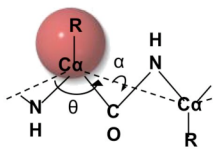
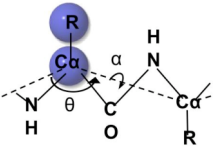
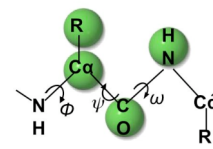
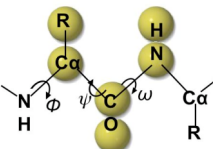


Abbiamo già accennato alle simulazioni ibride QM/MM in cui una parte piccola del sistema viene rappresentata con metodi quantistico (ad es Car-Parrinello) e il resto con FF empirici. I problemi da risolvere per la doppia rappresentazione del sistema riguardano il matching delle forze all'interfaccia tra le due rappresentazioni, e sono stati risolti nel corso degli anni con diverse "ricette". In generale, le forze vengono separate nelle loro componenti "bonded" e "non bonded". Nel trattare la parte "bonded" si ha a che fare con legami chimici che attraversano l'interfaccia QM/MM e quindi bisogna, da una parte troncare in maniera opportuna la densità elettronica che si deve annullare passando dalla parte QM a quella MM, simulare il legame tagliato. Questi problemi vengono risolti in una zona di interfaccia tra la zona QM e la MM per mezzo di artifici quali "atomi fittizi" di idrogeno (per saturare senza traumi la densità elettronica) e termini di forze aggiuntivi opportunamente parametrizzati per i legami "ibridi". Un'analoga riparametrizzazione necessitano i termini non bonded di tipo VdW. La parte elettrostatica, in linea di principio non ha bisogno di riparametrizzazioni: i nuclei ed elettroni della parte QM sentono il campo elettrostatico delle cariche sugli atomi della parte MM, mentre gli atomi MM risentono il campo elettrostatico generato dai nuclei e dalla nube elettronica della parte QM. Qui il problema è trovare un modo computazionalmente efficiente di calcolare i campi elettrostatici ad ogni passo della simulazione, specialmente quello della nube elettronica, che è stato generalmente risolto con un'espansione in multipoli della stessa⁵. In Fig 3.3 è riportata la suddivisione tipica di una proteina a cui è applicato questo approccio. In altri casi, come la simulazione di una reazione chimica in soluzione acquosa,

la suddivisione del sistema in parte QM e MM è più semplice: i reagenti vengono rappresentati come QM, l'acqua del solvente come MM, e non vengono tagliati legami covalenti, quindi non c'è bisogno di artifici per tagliare la densità elettronica (Fig 3.3). In ogni caso il timestep nelle rappresentazioni ibride è quello della parte QM, ovvero tipicamente 0.1fsec, quindi la lunghezza tipica della simulazioni è al massimo centinaia di ps.

3.3 Livello di aggregato macromolecolare: modelli Coarse Grained (CG_

Per aumentare le dimensioni dei sistemi simulabili e/o per allungare le traiettorie di simulazione è necessario diminuire i gradi di libertà semplificando la descrizione del sistema. Un modo per farlo è condensare molti atomi in singoli centri interattivi (Coarse Graining). Questo si può fare in molti modi diversi³. Ad esempio, per le proteine, i modelli CG più diffusi sono riportati in Tab 3.3. Risulta chiaro che minore il numero delle bead, maggiore il guadagno sul costo computazionale, ma d'altra parte più complicata la rappresentazione dell'energia potenziale. Ad esempio, togliendo centri interattivi posti sui gruppi CO e NH del backbone, si toglie la possibilità di rappresentare esplicitamente i legami a idrogeno, che vanno quindi inseriti implicitamente negli altri termini del FF, che sono di conseguenza più complessi e difficili da parametrizzare.

	one-bead	two-beads	four-beads	six-beads
				
Legame peptidico	implicito	implicito	esplicito	esplicito
Variabili conf del backbone	θ, α	θ, α	ϕ, ψ Mappa di Ramachandran esplicita	ϕ, ψ Mappa di Ramachandran esplicita
Legami a idrogeno del backbone	impliciti	impliciti	espliciti, ma complessi	espliciti
Effetti di volume catena laterale	impliciti	espliciti	espliciti	espliciti
Idrofobicità elettrostatica	trattate in un unico potenziale	trattate in potenziali separati per catena laterale e backbone	Termini separati e distribuiti sulle beads	Termini separati e distribuiti sulle beads

Tab 3.3 Tipi principali di modelli CG per proteine

Ci restringeremo d'ora in avanti ai modelli one-bead, sia per proteine – con centro interattivo sul $C\alpha$ – che per acidi nucleici – con centro interattivo sul P del gruppo fosfato. Questi modelli sono infatti rappresentativi, perché presentano, a volte amplificate, molte delle problematiche comuni a questi tutti i modelli CG, e in più sono in un certo senso "naturali": le bead rappresentano le unità base di proteine e acidi nucleici (rispettivamente amminoacidi e nucleotidi). Inoltre, hanno un legame diretto con i dati sperimentali: nei modelli a bassa risoluzione ottenuti con microscopia crio-elettronica, gli unici atomi risolti sono proprio i $C\alpha$ e i P.

Anche restringendoci ai modelli one-bead per proteine, la letteratura ne presenta un'ampia scelta⁹, differenti per tipo di termini presenti nel FF e loro forma funzionale e per la filosofia della parametrizzazione. Esse influenzano accuratezza e potere predittivo del modello, che normalmente sono due caratteristiche concorrenti, difficili da raggiungere contemporaneamente. In Fig 3.5 i vari modelli sono piazzati a seconda delle loro caratteristiche. Ad un estremo troviamo (in verde) i modelli a "bias totale": modelli a rete e di tipo Go. Riferendoci alla formula generica del FF riportata in Fig 3.5, nei modelli a rete¹⁰ sono presenti solo termini di tipo U^b , e sono considerate "legate" tutte le coppie di bead a distanza minore di una certa distanza di cutoff, che, nell'approssimazione più semplice, interagiscono tra loro con un potenziale armonico con costante k uniforme. Le distanze di equilibrio sono quelle di una struttura di riferimento, di solito

cristallografica. In pratica la proteina viene rappresentata come una rete di legami elastici. L'accuratezza strutturale è alta nel senso che la struttura di equilibrio è, per costruzione, quella sperimentale. Gli unici due parametri del FF, k e il raggio di cutoff, sono aggiustati in modo che siano riprodotte le fluttuazioni termiche sperimentali. Nonostante l'estrema semplicità, questo modello è in grado di dare molte informazioni. Ad esempio, per mezzo di esso, con una metodologia chiamata Analisi dei Modi Principali (PCA), che consiste sostanzialmente nell'analizzare i modi normali più ampi del sistema, si possono studiare i modi lenti, che di solito coincidono con quelli associati alla funzionalità della proteina (ad es, apertura e chiusura di siti attivi, binding modes etc). Il limite di questo modello è evidente: esso ha scarso potere predittivo, perché non può esplorare configurazioni lontane da quella di equilibrio, che per altro è un input. Sono state studiate estensioni del modello base che includono la dipendenza di k dal tipo di amminoacido, che aumenta la predittività, usano di potenziali anarmonici o a doppia buca e termini non bonded, per consentire una più ampia esplorazione dello spazio configurazionale¹¹, e infine reti elastiche a risoluzione multipla e adattabile, per migliorare l'efficienza computazionale¹².

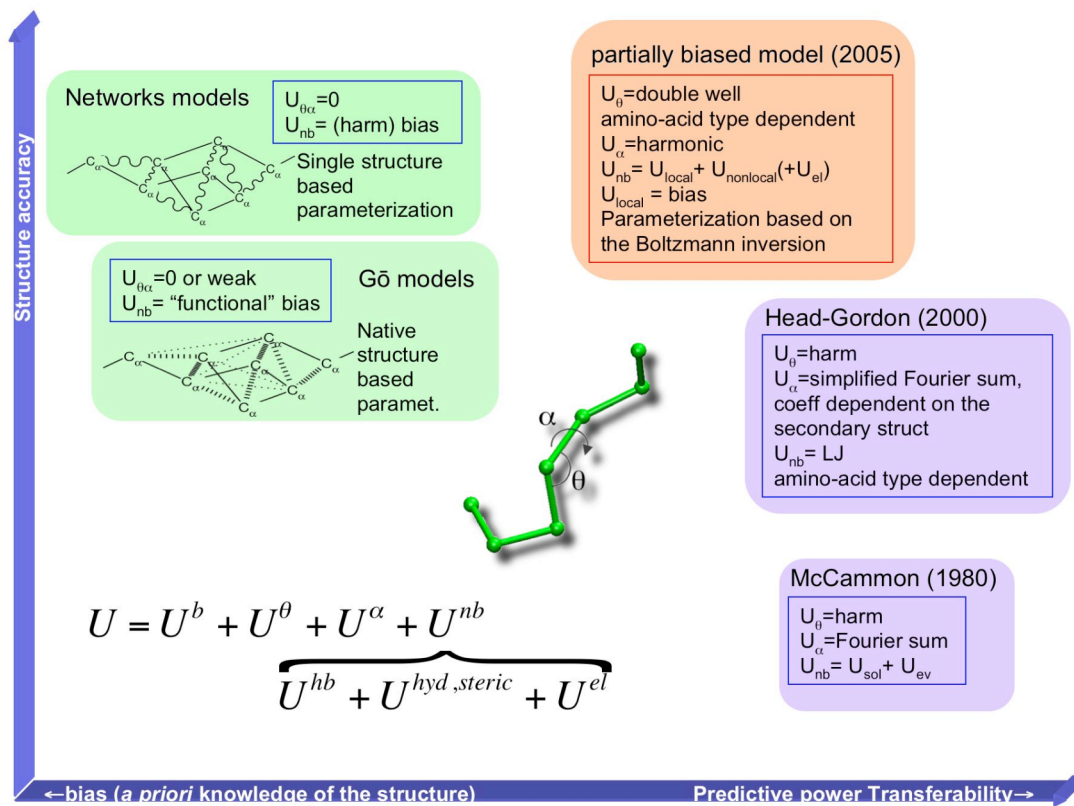


Fig 3.5 Complessità, accuratezza e potere predittivo di diversi tipi di modelli one-bead per proteine

Sempre nell'ambito dei modelli con bias, i modelli di tipo Go sono stati pensati per uno scopo diverso, cioè quello di simulare il folding. Anche in questo caso la parametrizzazione è basata su una struttura di riferimento, che la è struttura correttamente ripiegata (nativa). Il FF contiene termini $U^b + U^{\theta} + U^{\alpha} + U^{nb}$ separati. I primi tre termini sono di solito trattati come armonici e hanno i valori di equilibrio della struttura nativa (bias totale). Il termine non bonded è diviso in due parti: tutte le coppie non-bonded che si trovano all'interno di un cutoff (di solito settato intorno a 8 Å) nella struttura nativa (dette "contatti nativi") interagiscono con un potenziale tipo Lennard Jones con valore di equilibrio preso dalla struttura nativa stessa, che costituisce un ulteriore bias. Tutte le altre interagiscono con un potenziale repulsivo: i contatti non-nativi tendono a staccarsi. Quindi il modello ha, per costruzione, un minimo molto profondo nella struttura nativa. Anzi, il modello Go viene di solito detto modello per il folding completamente non-frustrato: a partire da una struttura sfoldata, il sistema si ripiega verso la struttura nativa senza rimanere intrappolato in intermedi di folding. Ovviamente questo tipo di modelli non serve per prevedere la struttura nativa, che è un input, ma può dare informazioni sulla dinamica di

formazione dei contatti nativi, e qualche informazione di tipo termodinamico, una volta che i parametri energetici (le costanti di forza e le energie di legame dei contatti nativi) sono opportunamente settati. È ovvio che la sua totale assenza di frustrazione è uno dei principali limiti del modello, che dà generalmente dinamiche di folding irrealistiche, prive di intermedi metastabili, che sono invece di solito presenti nel folding. Quindi le versioni più sofisticate del modello hanno nel corso degli anni, mirato ad introdurre la frustrazione in vari modi, ad esempio togliendo il bias dai termini angolari del potenziale e rendendoli indipendenti dalla struttura nativa e più fisicamente realistici, o introducendo potenziali non-bonded che tenessero conto dell'idratazione/disidratazione in modo da poter descrivere più accuratamente una prima fase che è spesso presente nel folding delle proteine, che consiste nell'espellere l'acqua per formare strutture secondarie e terziarie compatte.

All'altro estremo del grafico di Fig 3.5 troviamo i modelli totalmente privi di bias, il cui scopo è predire strutture (anche folding, quindi) senza averne alcuna conoscenza a priori. Quello di Head-Gordon (HG) può essere considerato paradigmatico per questa classe di modelli. Il FF contiene tutti i termini elencati in Fig 3.5, e le forme funzionali sono ispirate a quelle dei FF all-atom. Il termine U^b è armonico o un semplice vincolo a lunghezza fissata, U^θ è armonico, U^α è una somma di termini di tipo coseno e coseno triplo e U^{nb} contiene una parte non bonded standard (elettrostatica + idrofobicità) e un termine separato per il legame a idrogeno, che nell'ultima versione è non isotropo¹³. La parametrizzazione del modello HG è totalmente indipendente da strutture di riferimento, ma è inserita una certa conoscenza a priori nei parametri energetici del termine diedrale e nei termini non bonded. Specificamente,

$$U^\alpha = \sum_{dih} A[1 + \cos(\alpha + \alpha_0)] + B[1 - \cos(\alpha + \alpha_0)] + C[1 + 3\cos(\alpha + \alpha_0)] + D[1 + \cos(\alpha + \alpha_0 + \pi/2)]$$

dove i parametri A, B, C, D (e in minor misura, α_0) dipendono dalla struttura secondaria assegnata sulla base della sequenza. Si noti che praticamente tutta la determinazione configurazionale secondaria (cioè l'appartenenza ad un elica o un foglietto, ad esempio) è contenuta in questo termine, poco o niente nel termine U^θ . La parte idrofobica+elettrostatica di U^{nb} ha una forma tipo lennard jones a due parametri $\Sigma 4\epsilon S_1[(\sigma/r)^{12} - S_2(\sigma/r)^6]$ dove $\sigma \sim 1.6\text{\AA}$ ed ϵ sono uguali per tutti gli amminoacidi e assumono la funzione di unità lunghezza ed energia, rispettivamente, mentre i parametri S_1 ed S_2 dipendono dal tipo di amminoacido: tutti gli amminoacidi sono suddivisi in 4 classi a seconda della loro natura più o meno idrofobica o idrofilica e di conseguenza interagiscono con potenziali più o meno attrattivi o repulsivi. Infine, il termine di legame a idrogeno del backbone ha una forma anisotropa con pochi parametri uguali per tutti gli amminoacidi e assegnati sulla base della propensione a formare eliche o foglietti, come i parametri del termine diedrale. Il modello HG, come altri suoi predecessori più antichi e primitivi a cui esso si ispira, ha il pregio di non necessitare alcuna conoscenza a priori sul sistema, e riesce a predire generalmente il fold globale corretto di proteine (se non sono troppo grandi). Il problema di questo modello è la scarsa accuratezza strutturale locale, che gli deriva dall'eccessiva semplificazione dei potenziali e dei parametri, e dalla filosofia della parametrizzazione, che include molta conoscenza di tipo chimico-fisico e pochi dati strutturali.

Come si è detto, quello di combinare nei modelli minimalisti l'accuratezza strutturale e la predittività è alla fine, il nodo cruciale della questione. Tenendo sempre in mente che la meta finale – non ancora raggiunta – nello studio di questi modelli è la combinazione delle due qualità, presenteremo nel seguito possibili vie verso di essa e alcune soluzioni di compromesso.

La scarsa accuratezza del modello HG deriva in parte dal non aver introdotto informazioni strutturali nei termini di potenziale. Come abbiamo visto, sia nei termini conformazionali che nei termini non bonded la specificità è introdotta solo attraverso la parte energetica, mentre la parte "strutturale" (i parametri σ , α_0 e θ_0) sono trattati molto meno accuratamente. La parametrizzazione accurata dei termini $U^\theta + U^\alpha$ passa per un'analisi preliminare di come le variabili interne conformazionali rappresentano la struttura secondaria. Come è noto, nelle rappresentazioni all-atom la conformazione del backbone è determinata dagli angoli diedri ϕ e ψ che precedono e seguono il C α lungo lo scheletro della proteina. Nei modelli one-bead (o two-bead) queste variabili interne scompaiono in favore delle nuove variabili conformazionali α e θ

(Fig. 3.6 (b)). La mappatura $(\phi, \psi) \rightarrow (\theta, \alpha)$ può essere ricavata analiticamente¹⁴ (Fig. 3.6 (a)) e usata per mappare il plot di Ramachandran tradizionalmente rappresentato nel piano (ϕ, ψ) in un nuovo equivalente plot nel piano (θ, α) (Fig. 3.6 (c)). È interessante notare che, nonostante la mappa $(\phi, \psi) \rightarrow (\theta, \alpha)$ non sia biunivoca (punti appartenenti simmetrici rispetto alla diagonale nel piano (ϕ, ψ) vengono mappati nello stesso punto del piano (θ, α)), le zone corrispondenti a strutture secondarie differenti (colori differenti in Fig 3.6 (c)) restano separate, assicurando che le variabili (θ, α) sono ancora variabili utili per descrivere accuratamente le conformazioni proteiche.

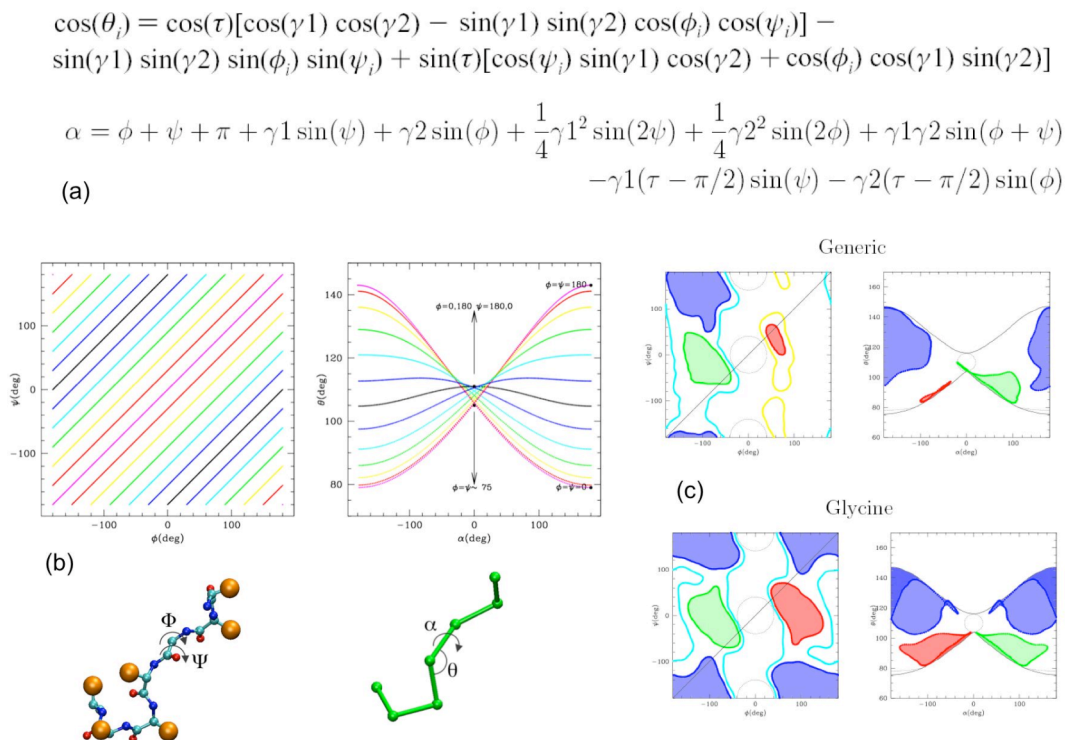


Fig 3.6 Mappatura delle variabili configurazionali dai modelli all-atom a one-bead. (a) formula analitica della mappatura (b) rappresentazione della mappatura del piano (ϕ, ψ) nel piano (θ, α) (c) mappatura del plot di Ramachandran per l'amminoacido generico e per la glicina.

Le mappe nel piano (θ, α) rappresentano distribuzioni di probabilità delle variabili collettive interne θ e α , e pertanto possono essere usate per ricavare i potenziali $U^{\alpha\theta} = U^\theta + U^\alpha$ tramite inversione di Boltzmann. La procedura standard consiste nel considerare come disgiunte le probabilità per le due variabili. Consideriamo ad esempio $P(\theta)$, la proiezione della mappa sull'asse θ , (Fig. 3.7). Essa presenta una struttura multipiccata. Però quando vengono considerate le distribuzioni separate per tipo di amminoacido, i picchi si riducono a due, uno localizzato a 90° e uno in posizione variabile dipendente dal tipo di amminoacido, da 120° a 150° . I due picchi hanno una interpretazione fisica molto semplice: rappresentano le conformazioni di elica o estesa. Invertendo questa distribuzione si ottiene un potenziale U^θ a doppia buca, con buche a profondità relativa diversa a seconda del tipo di amminoacido, che è in grado di rappresentare, a seconda della propensità secondaria dell'amminoacido in questione, eliche, o strutture estese, ma anche correttamente amminoacidi che hanno la possibilità di formare entrambe le strutture, come la glicina. Questo ci dice già che l'adozione di un potenziale armonico per il termine U^θ come nel modello HG è una scelta poco opportuna. Invece un modello minimalista con un potenziale a doppia buca è in grado di simulare la transizione tra elica e hairpin (Fig 3.8) e rappresentare le due strutture con alta accuratezza, come dimostrato dalla buona sovrapposizione tra la mappa conformazionale ottenuta dalla simulazione e quella "sperimentale".

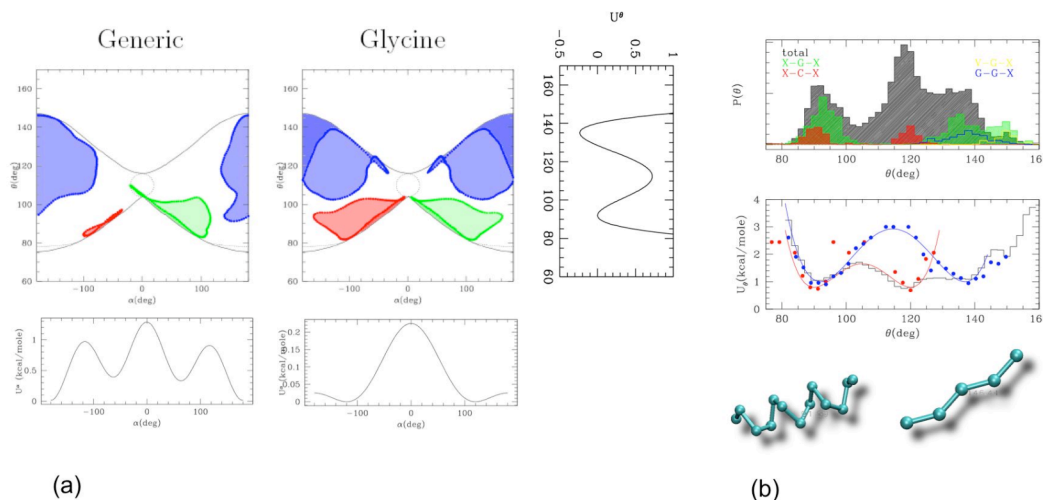


Fig 3.7 Inversion di Boltzmann applicata ai potenziali $U^{\alpha\theta} = U^\theta + U^\alpha$. (a) Proiezione della mappa bidimensionale sui due assi per ottenere le probabilità separate per i due angoli. (b) Inversione della probabilità per l'angolo θ e fit con un potenziale a doppia buca.

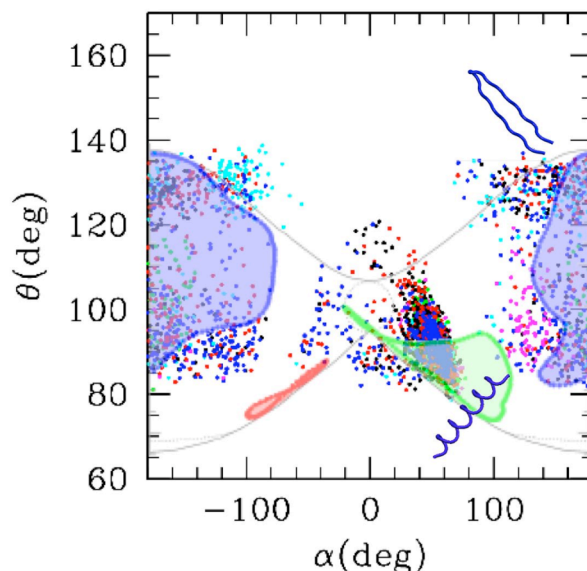


Fig 3.8 Simulazione della transizione da elica a hairpin in un modello one-bead.

Con una procedura analoga si possono in linea di principio parametrizzare in maniera accurata ed amminoacido dipendente tutti i termini del potenziale. Inoltre, siccome l'input è costituito dalle distribuzioni di probabilità, ricavabili da dati strutturali, data l'enorme mole di dati cristallografici attualmente disponibili esse possono avere un'accuratezza e generalità praticamente illimitata. Le difficoltà sono più che altro di tipo pratico: man mano che i termini aumentano, i parametri si moltiplicano e diventa molto difficile fare una parametrizzazione coerente, tenendo in conto in maniera opportuna delle correlazioni tra i vari termini. Le difficoltà maggiori derivano, come è intuibile, dal termine U^{nb} . Come detto, esso può essere parametrizzato tramite l'inversione di Boltzmann della funzione di distribuzione di coppia $g(r)$. Come osservato in precedenza, in realtà, anche per un sistema molto semplice come ad esempio un fluido uniforme di particelle che interagiscono solo con il potenziale di Lennard Jones, il potenziale di forza media $W(r) = -kT \ln(g(r))$ e il potenziale "vero" $V(r)$ differiscono, anche sostanzialmente, a causa delle correlazioni dovute alle altre variabili interne. Questo ha l'effetto di aggiungere molti picchi e valli a $W(r)$ o accentua quelli che sarebbero presenti realmente anche in $V(r)$. Questi effetti di correlazione sono in realtà presenti in tutti i termini del potenziale, ma in generale sono abbastanza trascurabili, mentre, essendo una funzione di una variabile non locale, il potenziale non bonded è quello che ne risente

di piú. Quindi di solito $V(r)$ viene ricavato da $W(r)$ attraverso una procedura iterativa di aggiustamenti fino a che la $g(r)$ della simulazione non approssima bene quella sperimentale.

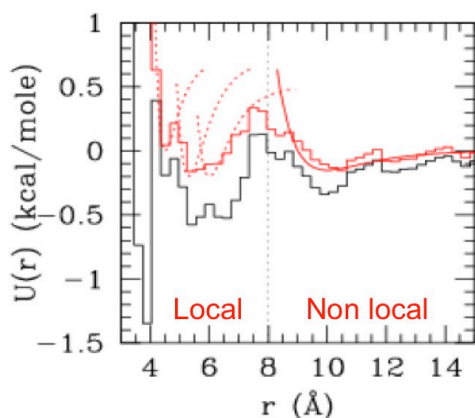
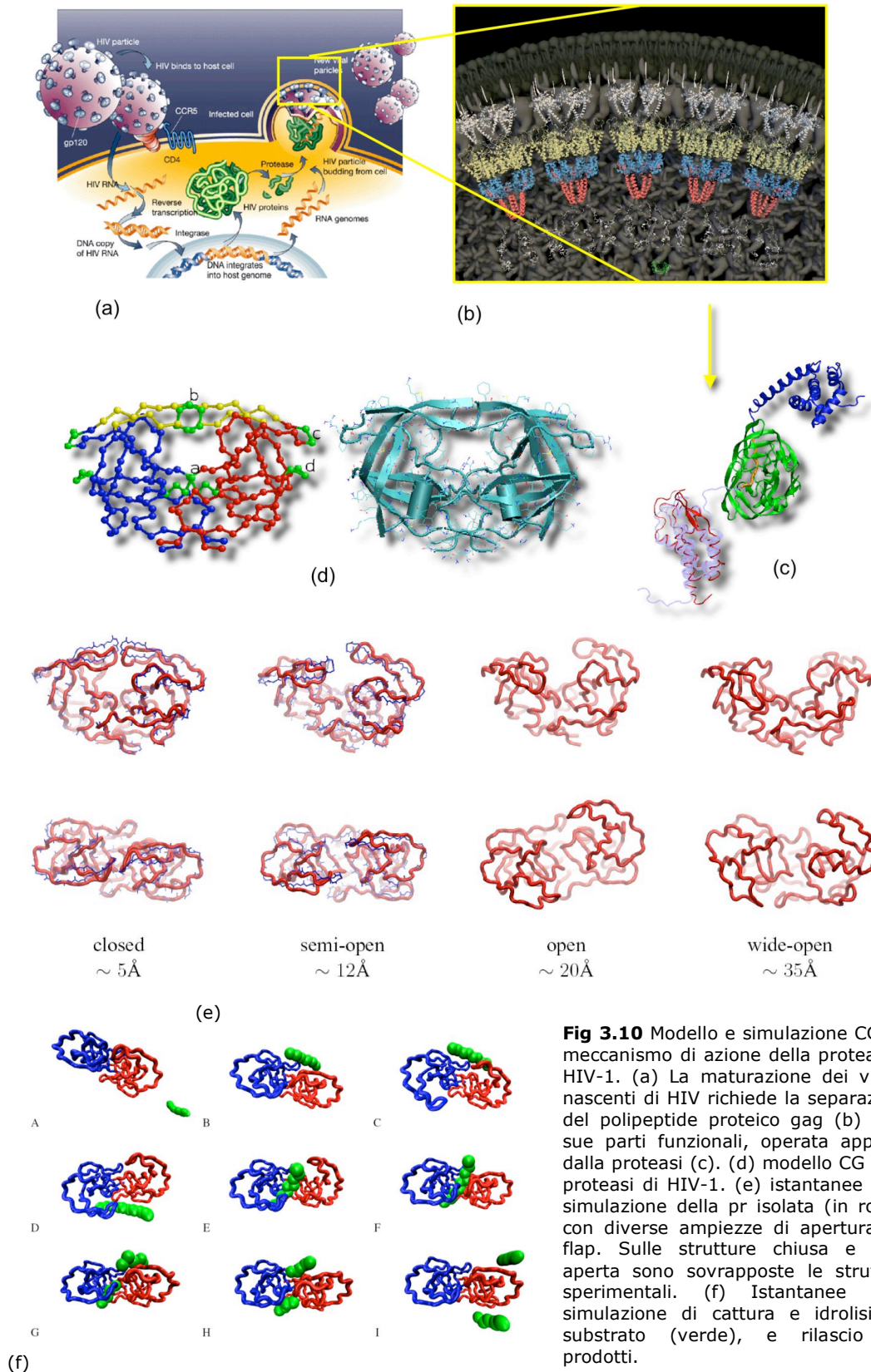


Fig 3.9 Separazione del potenziale non bonded in parte locale e non locale.

Altri problemi derivano dall'effettiva complessità di U^{nb} : esso include effetti elettrostatici, effetti di forma e conformazione della catena laterale, effetti di idratazione, e, se non sono scorporati, anche effetti di legame a idrogeno. Globalmente U^{nb} assume l'aspetto di Fig 3.9 (linea rossa): una parte "locale" (<8 Å) molto strutturata, contiene molti degli effetti piú complessi (legami a idrogeno, conformazioni e forma delle catene laterali, idratazione-disidratazione) e una parte non locale piú regolare che contiene sostanzialmente idrofobicità ed elettrostatica. Questa naturale suddivisione di U^{nb} in parte locale e non locale con caratteristiche differenti, offre una possibile soluzione di compromesso per combinare accuratezza e generalità nel modello senza complicare troppo il FF: U^{nb} viene scritto come la somma di due termini $U^{nb}_{loc} + U^{nb}_{non-loc}$ di cui la parte non locale viene descritta da potenziali isotropi tipo Lennard Jones o Morse, parametrizzati in maniera amminoacido-dipendente sulla base dell'inversione di Boltzmann; la parte locale viene descritta sempre tramite potenziali "staccabili" come LJ o Morse, ma la parametrizzazione mantiene il bias verso una struttura di riferimento. In questo modo si mantiene un'alta accuratezza tipica dei modelli a rete o Go: infatti il termine locale descrive in maniera semplice, efficace e accurata tutte quelle interazioni complesse e anisotrope che creano difficoltà nella parametrizzazione. Allo stesso tempo gli altri termini del potenziale non contengono bias, e quindi danno generalità e predittività al modello. L'equilibrio tra predittività e accuratezza viene regolato dall'ampiezza del cutoff. Questa semplice soluzione di compromesso, benché non ottimale, si è rivelata molto efficace in diversi contesti. Uno è stato lo studio delle basi molecolari del meccanismo di azione della proteasi di HIV-1¹⁵. Essa è uno dei tre enzimi virali essenziali per la replicazione del virus: durante la fase di maturazione del virus, opera la separazione del polipeptide virale gag in parti funzionali che andranno, tra l'altro a formare il capsido dei nuovi virus (Fig 3.10 (a,b)). I complessi coinvolti nel processo sono molto grandi (Fig 3.10 (b,c)), ma soprattutto i tempi coinvolti sono dell'ordine di micro-millisecondi, che rende molto improbabile l'osservazione di eventi interessanti con metodi basati su rappresentazioni all-atom. Nel modello CG invece la proteasi consiste di soli 198 centri interattivi, il che consente di raggiungere facilmente la scala delle decine o centinaia di microsecondi. Così è stato possibile osservare nella simulazione il movimento di apertura e chiusura dei flap, due strutture a forma di hairpin che coprono il sito attivo (Fig. 3.10 (e)). L'accuratezza strutturale del modello è stata confermata a posteriori da una struttura sperimentale semi-aperta cristallizzata dopo la pubblicazione del lavoro teorico, che si sovrappone molto bene con le strutture semi-aperte della simulazione. La simulazione ha reso possibile l'individuazione di siti allosterici il cui bloccaggio porterebbe all'inibizione dell'apertura del sito. Inoltre è stato possibile simulare l'intero processo di cattura di un substrato-modello (un oligopeptide, Fig 3.10 (f)), idrolisi e rilascio dei prodotti. La simulazione dello stesso processo sul substrato reale è in corso (Fig 3.10 (c)). Sono stati anche analizzati gli effetti delle molecole "crowders" (proteine ed altre molecole normalmente presenti nell'ambiente cellulare) sulla frequenza di apertura e chiusura dei flap. Questo è un esempio di CG multi-scala, dal momento

che i crowders vengono rappresentati come grandi sfere destrutturate.



Naturalmente i modelli CG sono stati usati per simulare anche sistemi molto più grandi di semplici proteine, come interi ribosomi, nucleosomi o altri aggregati macromolecolari (Fig 3.11)¹⁶. In

questo caso i tempi di simulazione massimi accessibili si aggirano sulle centinaia di nsec, ma sono stati sufficienti per analizzare i moti correlati alla traslocazione nel ribosoma, e le prime fasi dello svolgimento del DNA nel nucleosoma.

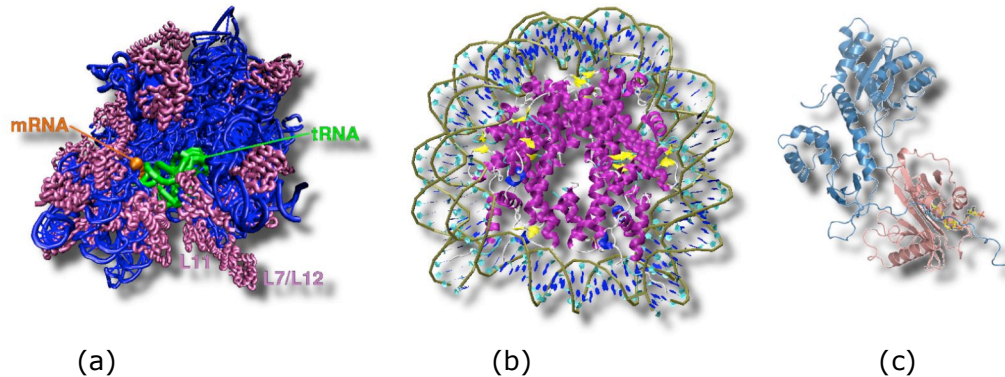


Fig 3.11 Tre sistemi macro-aggregati molecolari simulati con modelli CG (a) Il ribosoma batteri S70 (b) un nucleosome (c) Integrasi dell'HIV complessata con proteina HAT.

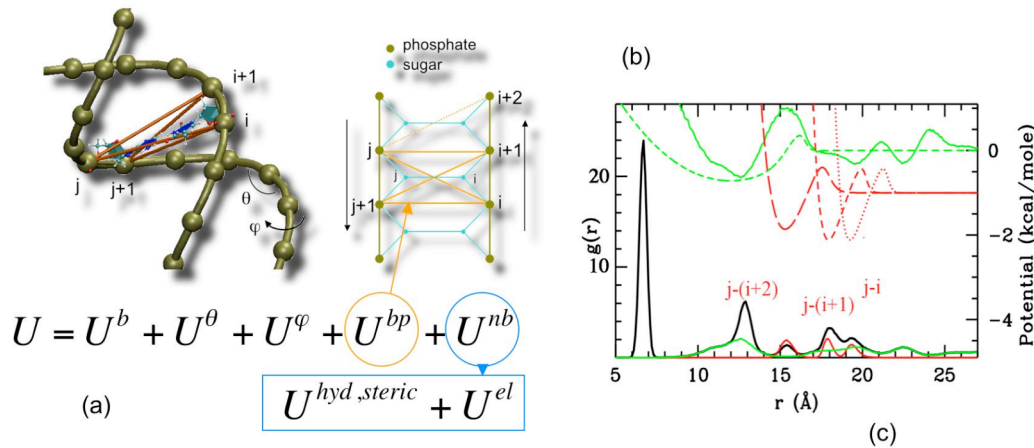


Fig 3.12 Modello one-bead per il DNA. (a) Descrizione dei termini del campo di forze e (b) derivazione dei termini U^{bp} e U^{nb} tramite inversione di Boltzmann. (c) istantanee dalla simulazione di un plasmide di ~ 900 coppie di basi.

In questi modelli per i nucleosomi e ribosomi anche gli acidi nucleici, come le parti proteiche, contengono un bias parziale. Tuttavia, per il DNA costruire un modello completamente privo di bias e con alta accuratezza è più semplice, data la sua maggiore semplicità strutturale rispetto alle proteine. Un modello minimalista per il DNA con queste caratteristiche è descritto in Fig 3.12¹⁷. Le interazioni di accoppiamento WC sono trattate separatamente dal resto delle non bonded, e la loro direzionalità è modellata con tre legami per coppia di nucleotidi con una geometria particolare, che mantiene la torsione e riproduce sia le componenti di forza parallele al piano delle basi, dovute ai legami a idrogeno, sia quelle ortogonali, dovute allo stacking. Inoltre sia il termine di accoppiamento WC sia l'altro termine non-bonded, contenente elettrostatica e idrofobicità sono rappresentati da un potenziale con una buca e una barriera che deve tenere conto della repulsione elettrostatica tra i fosfati oltre che delle interazioni attrattive locali. Questo semplice potenziale è

in grado di riprodurre accuratamente fenomeni complessi come il supercoiling e la denaturazione in sistemi anche molto grandi come interi plasmidi (Fig 3.12 (c)).

A conclusione di questa sezione, accenniamo ai modelli a risoluzione mista. Come nel caso del passaggio tra QM e MM anche nel passaggio da MM a CG è possibile (ed è stato fatto), considerare modelli ibridi MM/CG in cui una parte piccola del sistema è trattata in rappresentazione all-atom e il resto in rappresentazione CG. Anzi, più in generale sono stati studiati sistemi ibridi con più di due livelli diversi di risoluzione. Anche qui il problema è il corretto trattamento delle forze all'interfaccia tra due rappresentazioni diverse, che viene generalmente risolto con la metodologia del force matching¹⁸. L'uso di più rappresentazioni diverse in questo caso è in qualche modo incoraggiato dal fatto che almeno tutte le rappresentazioni sono di tipo classico, non si presenta il problema di come terminare densità elettroniche e la dinamica del sistema può essere trattata nell'ambito di un singolo approccio. Tuttavia notiamo che, come nel caso degli approcci QM/MM il timestep è limitato dalla rappresentazione a più alta risoluzione, e dunque le simulazioni non possono essere temporalmente molto estese.

3.4 Modelli mesoscopici e passaggio al continuo: cenni

Al limite del coarse graining più spinto troviamo modelli che abbandonano la rappresentazione "particellare" del sistema in favore di una rappresentazione "estesa". Qui descrivere le metodologie al livello generale è praticamente impossibile, perché i singoli modelli sono tutti uno differente dall'altro e hanno ciascuno le proprie particolarità. Ci limitiamo a dare un elenco con qualche caratteristica descrittiva, riportata poi schematicamente in Fig 3.13.

1. *Modelli mesoscopici*. Intere proteine o domini vengono rappresentate come singoli oggetti rigidi dalla forma semplice (sfere, ellissoidi). Abbiamo già visto un esempio di questa estrema semplificazione (la rappresentazione di proteine come sfere per rappresentare i "crowders" dell'ambiente cellulare), che però viene usata solo in casi molto particolari.
2. *Superfici molecolari*. Ad un livello un po' più sofisticato, molecola viene rappresentata tramite una superficie (normalmente rigida) che la delimita dall'ambiente esterno. Essa può essere definita in vari modi: tramite il volume escluso (superficie di VdW), superficie accessibile al solvente (luogo dei centri di una sfera-sonda del diametro delle molecole di solvente che rotola sulla superficie di VdW), superficie elettronica (superficie di livello della densità elettronica). Se l'oggetto viene trattato come un corpo rigido è descritto da soli 6 gradi di libertà (rotazioni più traslazioni), ma ovviamente servono molti parametri per definire la forma della superficie, che di solito viene descritta tramite un insieme di punti che la "triangolano", il cui numero dipende dalla precisione desiderata. Questo tipo di approccio viene usato soprattutto nei metodi di molecular recognition (interazione proteina-proteina) e docking (inserimento di piccole molecole in siti proteici): in questi casi la complementarità di forma è uno dei criteri principali per determinare l'affinità tra due biomolecole. Inoltre sulle superfici si possono mappare proprietà tipo l'idrofobicità e il potenziale elettrostatico in base alle quali si può determinare un punteggio aggiuntivo per l'affinità.
3. *Modelli "elastici"*. Per alcuni sistemi particolari è possibile reincludere la dinamica interna in maniera semplice anche in una rappresentazione di tipo continuo. È questo il caso della rappresentazione delle membrane cellulari come superfici (quasi) bidimensionali elastiche o del DNA o delle fibre di collagene come "corde" flessibili e con elasticità torsionale. In questo caso si devono definire (poche) ulteriori coordinate interne che descrivono le deformazioni.
4. *Solventi continui*. Esiste una vastissima letteratura sui metodi di rappresentazione di solventi come mezzi continui. Abbiamo già parlato di come alcune delle proprietà dovute alla "corpuscolarità" del solvente si possono reintrodurre implicitamente (dinamiche stocastiche). Ma molti solventi organici, e l'acqua sopra a tutti, sono dotati anche di proprietà elettrostatiche, che è fondamentale tenere in considerazione. Esiste una gran quantità di modelli di solvente polarizzabile, che in sostanza rinormalizzano le interazioni elettrostatiche tramite una funzione dielettrica eventualmente dipendente dalla distanza (modelli tipo Born). Questi modelli dielettrici sono un'approssimazione grossolana dell'approccio di Poisson-Boltzmann (PB), che tiene conto della presenza di ioni dissociati in acqua, i quali, a seconda della distribuzione delle cariche della molecola di soluto (ad es la

proteina), si distribuiscono intorno ad essa per schermarla. L'approccio PB è accurato ma è talmente dispendioso in termini di tempo di calcolo da non poter essere usato durante la dinamica.

5. *Dinamica diffusiva di proteine.* Al livello estremo, un insieme statistico di proteine può essere trattato come un "fluido" continuo descritto da una variabile densità $n(r,t)$, la cui evoluzione temporale è guidata dalle equazioni di trasporto, la cui formula generale è molto complessa. In una forma un po' semplificata si può scrivere

$$\frac{\partial n}{\partial t} = \nabla \cdot [-n(r)F[n](r) + C\nabla n(r)]$$

dove $F[n](r)$ è la somma totale delle forze agenti sulla particella del fluido al punto r (incluente interazioni con il campo esterno e con le altre particelle) e il secondo termine è dovuto al fatto che le particelle tendono a diffondere dove sono meno concentrate. Sotto opportune approssimazioni, l'equazione sopra è equivalente alla dinamica Browniana, espressa in funzione della densità di particelle, anziché per la singola particella. Nel limite in cui $n(r)$ è uniforme si ottiene

$$\frac{\partial n}{\partial t} = -nK$$

la classica equazione per la cinetica del primo ordine che si applica ai cosiddetti "modelli a compartimenti" per la diffusione di fluidi o particelle: si divide il sistema in compartimenti (ad esempio i compartimenti cellulari), ciascuno caratterizzato da una variabile $n(r)$ che definisce la sua popolazione e si stimano le costanti cinetiche k che collegano i compartimenti. L'equazione diventa un sistema di n equazioni del primo ordine la cui soluzione darà l'evoluzione temporale delle popolazioni dei compartimenti. Questi modelli (sia quello diffusivo generale che quello a compartimenti) possono essere usati per lo studio di diffusione di proteine all'interno della cellula, e quindi per una diretta interpretazione delle osservazioni di esperimenti di FRAP e simili.

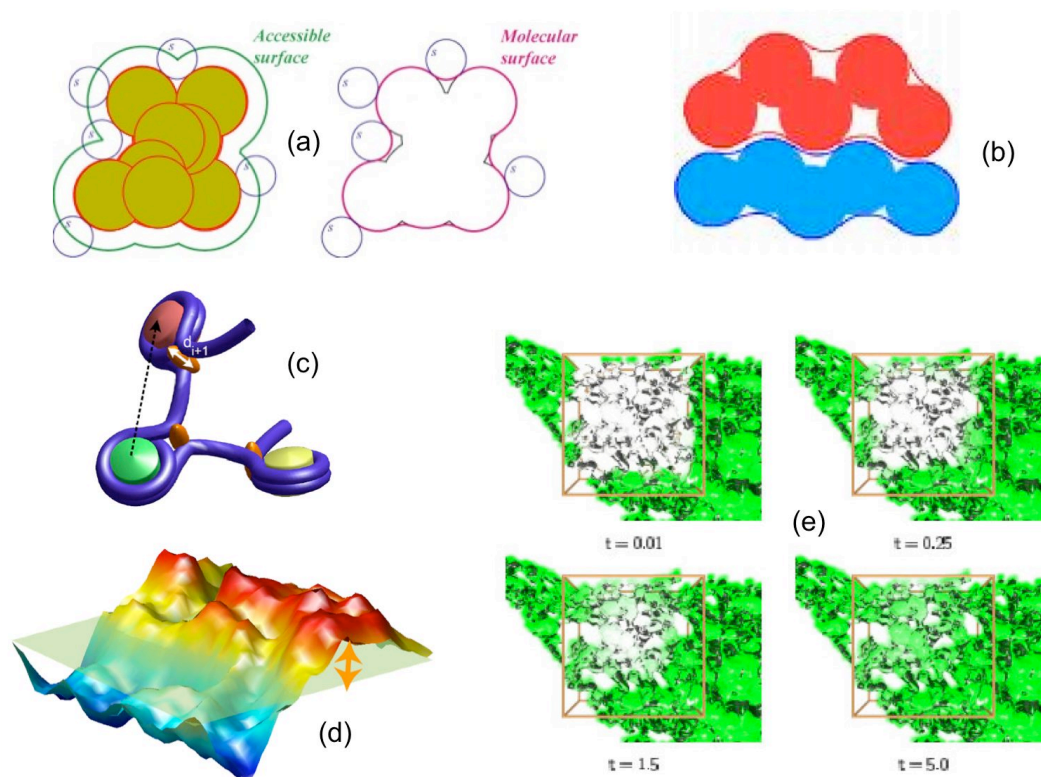


Fig 3.13 (a) Definizione di superfici molecolari e superfici accessibili al solvente (b) illustrazione della complementarità di forma (c) Modello mesoscopico e per la cromatina (d) modello continuo per la membrana (e) simulazione di un esperimento di FRAP tramite dinamica diffusiva.

Riferimenti Bibliografici

- ¹ Si veda ad esempio: D Marx and J Hutter "Ab initio molecular dynamics: Theory and implementation" in *Modern Methods and Algorithms of Quantum Chemistry* J Grotendorst Ed NIC Series Vol 1 p 301-449, 2000
- ² *Understanding Molecular Simulations* Frenkel, Smit, Academic Press Ed (1996)
- ³ V Tozzini *Coarse Grained Models for Proteins* Curr Opin Struct Biol 15 144-150 (2005)
- ⁴ Multiscale modeling of biomolecular systems: in serial and in parallel Gary S Ayton, Will G Noid and Gregory A Voth Current Opinion in Structural Biology 2007, 17:192-198
- ⁵ HM Senn, W Thiel, "QM/MM Methods for biological systems" To Curr Chem 268 173-290 (2007)
- ⁶ J Thar, W Reckien, B Kirchner Car-Parrinello Molecular Dynamics and Biological systems Top Curr Chem (2007) 268, 133-177
- ⁷ *Comparison of Protein Force Fields for Molecular Dynamics Simulations* O Guvench and A.D. MacKerell, Jr. in *Methods in Molecular Biology*, vol. 443, Molecular Modeling of Proteins Edited by Andreas Kukol Humana Press, Totowa, NJ; *A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6* C Oostenbrink A Villa A.E. Mark W.F. van Gunsteren J Comput Chem 25: 1656-1676, 2004
- ⁸ R Nifosí and V Tozzini *Cis-trans isomerization of the chromophore in the Green Fluorescent Protein variant E2GFP: a molecular dynamics study* Chem Phys, 323 358-368 (2006)
- ⁹ V Tozzini and J A McCammon *One-Bead Coarse Grained Models for Proteins* in "Coarse-graining of condensed Phase and Biomolecular Systems", G A Voth Ed, Taylor & Francis/CRC press, Chapter 19 (2008)
- ¹⁰ tyrion
- ¹¹ J Trylska, V Tozzini, J A McCammon *Exploring Global Motions and Correlations in the Ribosome* Biophys J 89 1455-1463 (2005)
- ¹² A multiscale coarse-graining method for biomolecular systems A Izvekov, GA Voth J Phys Chem B 109 2469-73 (2005)
- ¹³ *A coarse-grained α -carbon protein model with anisotropic hydrogen-bonding* Eng-Hui Yap, Nicolas Lux Fawzi, and Teresa Head-Gordon Proteins 2008; 70:626-638.
- ¹⁴ V Tozzini, W Rocchia, J A McCammon *Mapping all-atom models onto one-bead Coarse Grained Models: general properties and applications to a minimal polypeptide model* J Chem Theory Comput, 2 667-673 (2006)
- ¹⁵ V Tozzini, J A McCammon *A coarse grained model for the dynamics of flap opening in HIV-1 Protease* Chem Phys Lett 413 123-128 (2005); V Tozzini, J Trylska, C-E Chang, J A McCammon *Flap opening dynamics in HIV-1 protease explored with a coarse-grained model* J Struct Biol, 157 606-615 (2007); J Trylska, V Tozzini, C-E Chang, J A McCammon *HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics* Biophys J, 92 4179-4187 (2007)
- ¹⁶ J Trylska, V Tozzini, J A McCammon *Exploring Global Motions and Correlations in the Ribosome* Biophys J 89 1455-1463 (2005); K Voltz, J Trylska, V Tozzini, V Kurkal-Siebert, J Langowski, J Smith, *Coarse-grained force field for the nucleosome from self-consistent multiscaling* J Comput Chem 29, 1429-1439 (2008); A Di Fenza, W Rocchia, V Tozzini *Complexes of HIV-1 integrase with HAT proteins: multi-scale models, dynamics and hypotheses on allosteric site of inhibition* Proteins, in press
- ¹⁷ F Trovato, V Tozzini *Supercoiling and Local Denaturation of Plasmids with a Minimalist DNA Model* J Phys Chem B 112, 13197-13200 (2008)
- ¹⁸ Mixed Atomistic and Coarse-Grained Molecular Dynamics: Simulation of a Membrane-Bound Ion Channel Qiang Shi, Sergei Izvekov, and Gregory A. Voth *J. Phys. Chem. B*, 110, No. 31, 2006; Marilisa Neri, Claudio Anselmi, Michele Cascella, Amos Maritan, and Paolo Carloni *Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site* Phys. Rev. Lett. 95, 218102 (2005)