

Corso di
Chimica delle Macromolecole I

Prof. R.URBANI

Dipartimento di Scienze Chimiche e Farmaceutiche
a.a. 2021-2022

PREDIZIONE DELLA CONFORMAZIONE:
Meccanica Molecolare (MM)
Statistica conformazionale

CALCOLO CONFORMAZIONALE

Individuazione della
struttura molecolare
rappresentativa della catena

scelta del metodo:
- calcolo "ab initio"
- calcolo semiempirico
- meccanica molecolare
- dinamica molecolare

scelta del campo di
forze per E_{tot}

scelta delle variabili
descrittive del sistema

**PREDIZIONE DELLE
STRUTTURE ORDINATE**
(stato cristallino, struttura
secondaria dei biopolimeri,...)

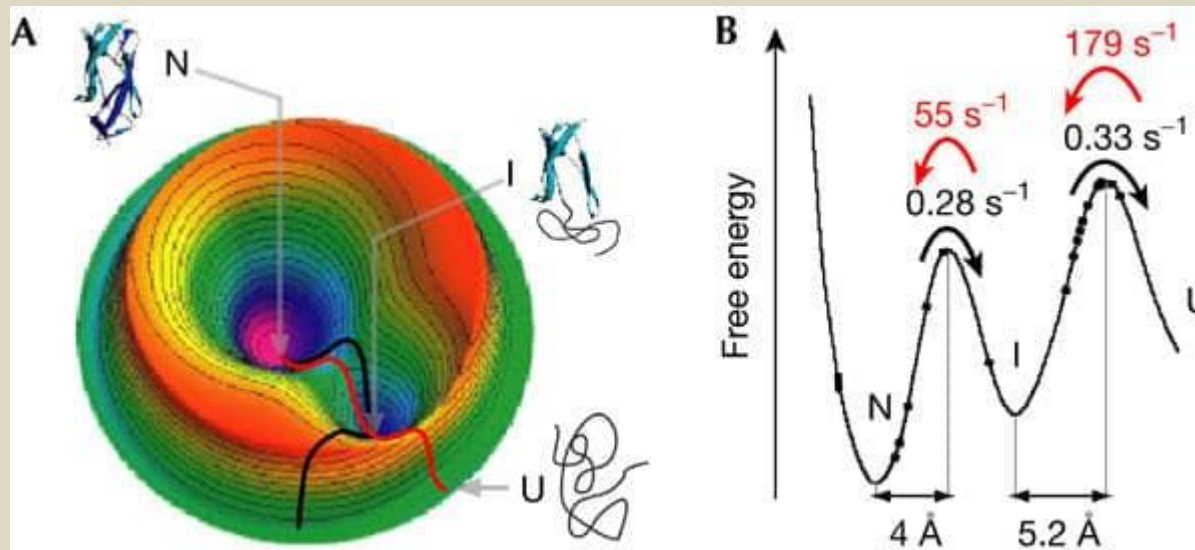
**STATISTICA
CONFIGURAZIONALE**
Stato disordinato (amorfo, in
soluzione)

scelta dei legami descrittivi
della catena (reali o virtuali)

scelta del modello probabilistico:
- catena Bernoulliana
- catena Markoviana
- catena di Ising

scelta del metodo:
- metodo matriciale di Flory
- metodo Monte Carlo

Predizione della struttura: Meccanica e Dinamica Molecolare



MM

Molecular
Mechanics



minimizzazione
dell'energia

$$V(q)$$

q = coordinate degli
N atomi



Osservabile A

$$A = f(q)$$

$\langle A \rangle$ = media su N
stati (insieme)

MD

Molecular
Dynamics



integrazione dell'equazione
di Newton (in vacuo)

$$P_i = F_i \quad (i = 1, \dots, N)$$
$$F_i = \partial V(q) / \partial q_i$$


Osservabile A

$$A = f(q, t)$$

$\langle A \rangle$ = media
temporale

definizione di un campo di forza:

$$V(q) = k_b(b-b_0)^2 + k_\theta(\theta-\theta_0)^2 + k_\phi[1 + \cos(n\phi)] +$$
$$+ A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^6 + q_i q_j / r_{ij} + \dots$$

Predizione della struttura:

Meccanica Molecolare (MM)

Dinamica Molecolare (MD)

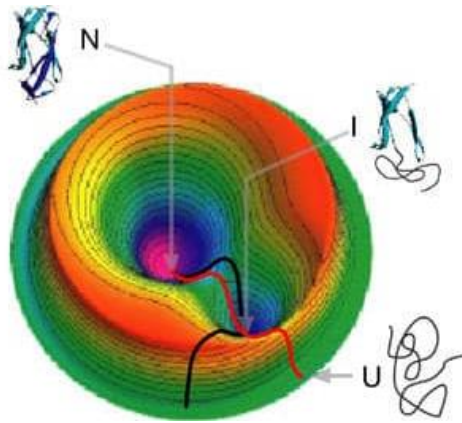
MM ed MD sono approcci computazionali molto usati sia nella scienza dei materiali che nello studio di biomolecole.

Con questi metodi viene seguita passo per passo e simulata in dettaglio la **dinamica e l'energetica del sistema**.

Mediante l'integrazione delle equazioni del moto, permettono di studiare la dinamica di evoluzione di un sistema fisico e chimico a livello atomico e molecolare.

Di conseguenza, le simulazioni MM/MD possono dare informazioni molto dettagliate sulla **stabilità conformazionale** e su fenomeni legati ai **moti interni** delle macromolecole ed alle transizioni conformazionali

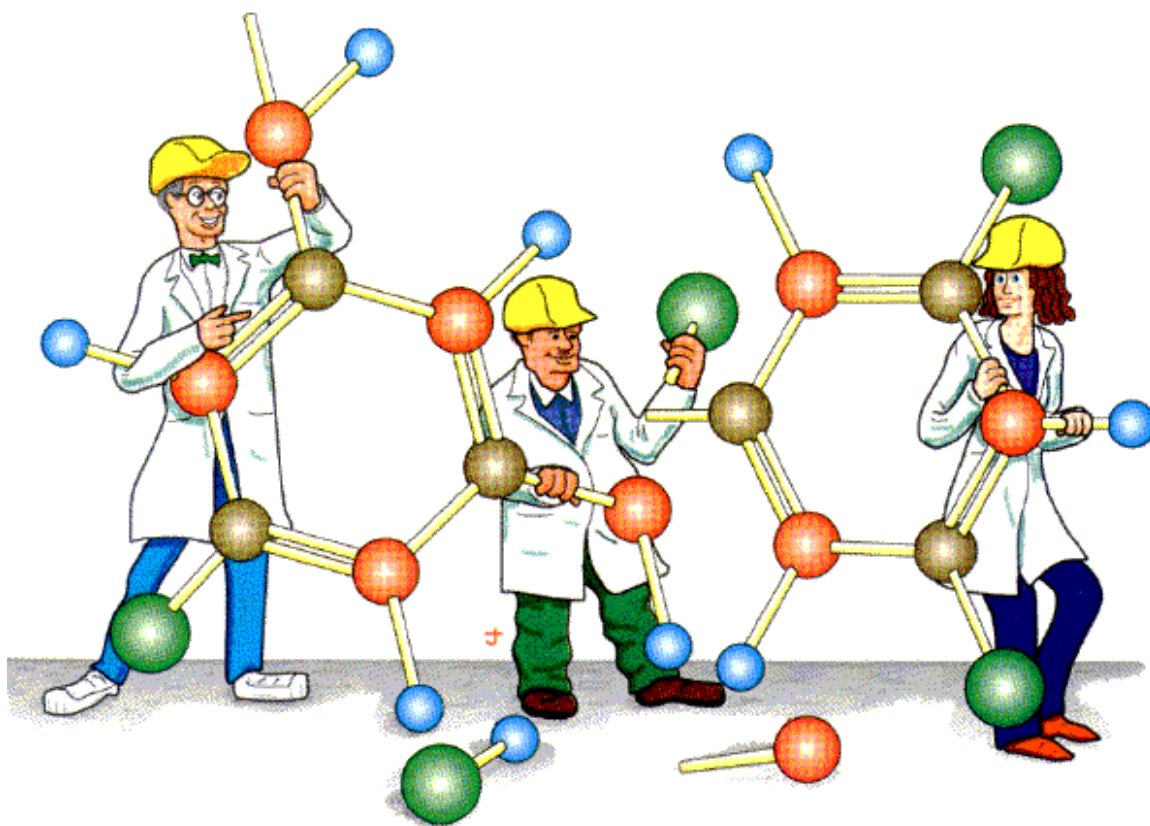
Con la **MM/MD** possono essere studiati in particolare:



- stabilità di piccole molecole(oligomeri);
- cambi conformazionali
- polymer folding
- soluzioni polimeriche
- interazioni tra segmenti polimerici
- complessi
- trasporto ionico in sistemi biologici
-

La tecnica fornisce anche un valido aiuto nel design di nuovi composti complessi (es. dendrimeri, farmaci, ...) e nella determinazione di strutture di molecole complesse, in combinazione con tecniche sperimentali quali diffrazione di raggi X e NMR.

Modellizzazione Molecolare



Limiti delle tecniche sperimentali

risoluzione spaziale



misure ad alta risoluzione di strutture molecolari sono possibili solo per sistemi relativamente rigidi (cristalli, fibre,...)

risoluzione energetica



analisi delle energie di interazione atomica difficoltosa

risoluzione temporale



i primissimi eventi nelle reazioni chimiche e nei processi biologici sono spesso di difficile misurazione

Limiti delle tecniche computazionali

Sistemi molecolari troppo complessi  meccanica classica con funzioni di interazione **semi-empiriche** per descrivere le interazioni tra gli atomi di un sistema molecolare

Simulazione del comportamento di un sistema molecolare

 solo per un **numero limitato** di atomi o di gradi di libertà (di solito 10^2 - 10^5 atomi), per un limitato periodo di tempo (10^2 - 10^4 picosecondi) $\Rightarrow$ piccoli sistemi, con tempi di rilassamento brevi

Campionatura limitata dello spazio delle conformazioni di una macromolecola  utilizzo dei dati sperimentali per restringerlo

Ottimizzazione del modello

```
graph TD; A[Ottimizzazione del modello] --> B[Regolarizzazione di legami, angoli e torsioni]; A --> C[Minimizzazione energetica]; A --> D[Eliminazioni di clash strutturali];
```

Regolarizzazione
di legami, angoli e
torsioni

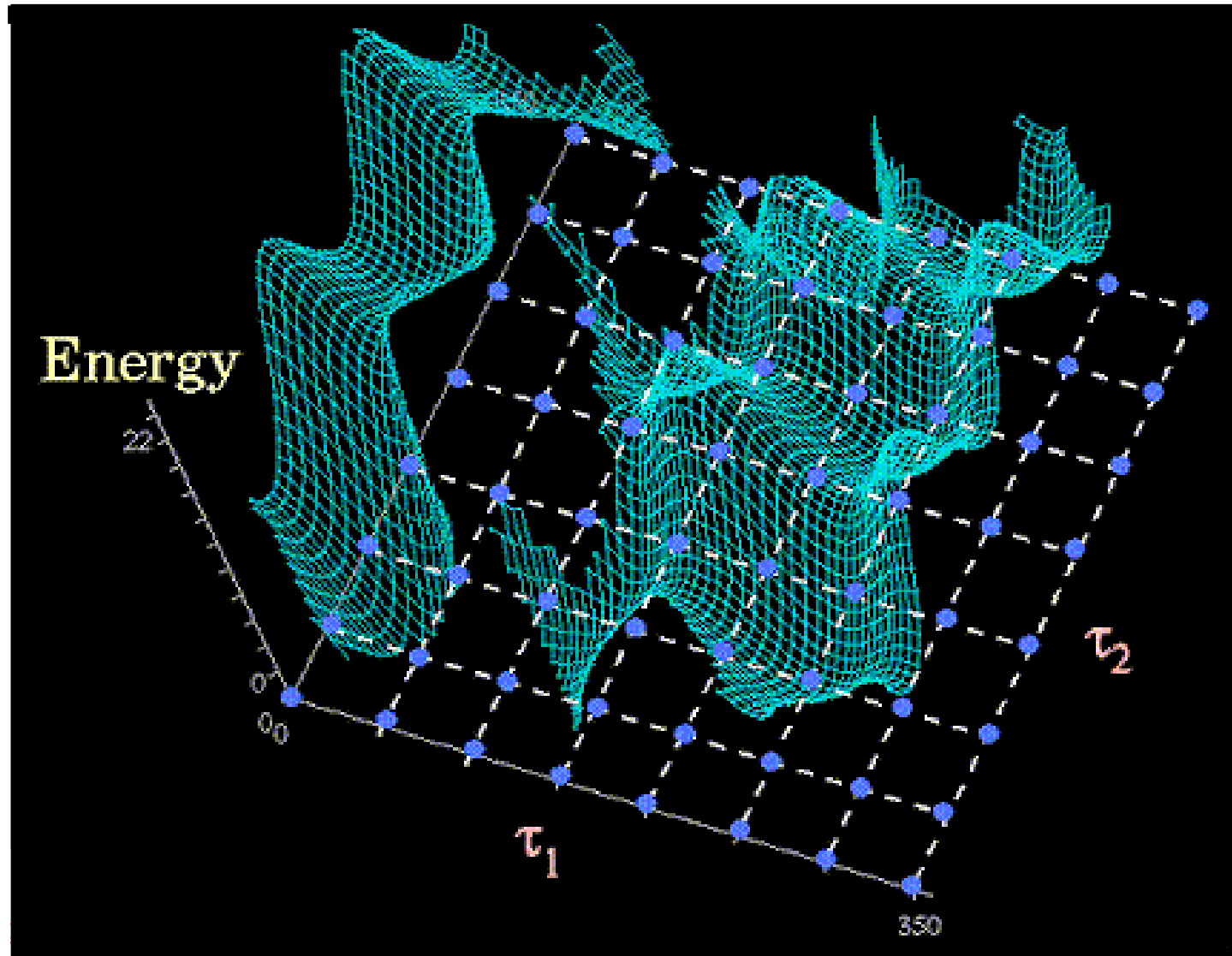
Minimizzazione
energetica

Eliminazioni di
clash strutturali

Meccanica molecolare

**RICERCA dei MINIMI ASSOLUTI e RELATIVI
DELL'ENERGIA CONFORMAZIONALE**

Ad ogni conformazione molecolare è associata un'ENERGIA



Conformazione energeticamente preferita



minimo globale dell'energia potenziale



$$F_i = -\partial V / \partial r_i$$



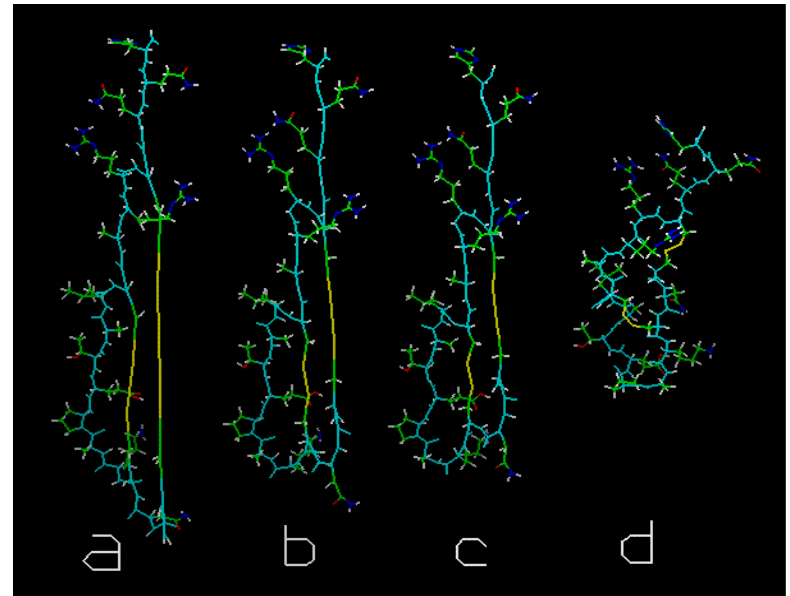
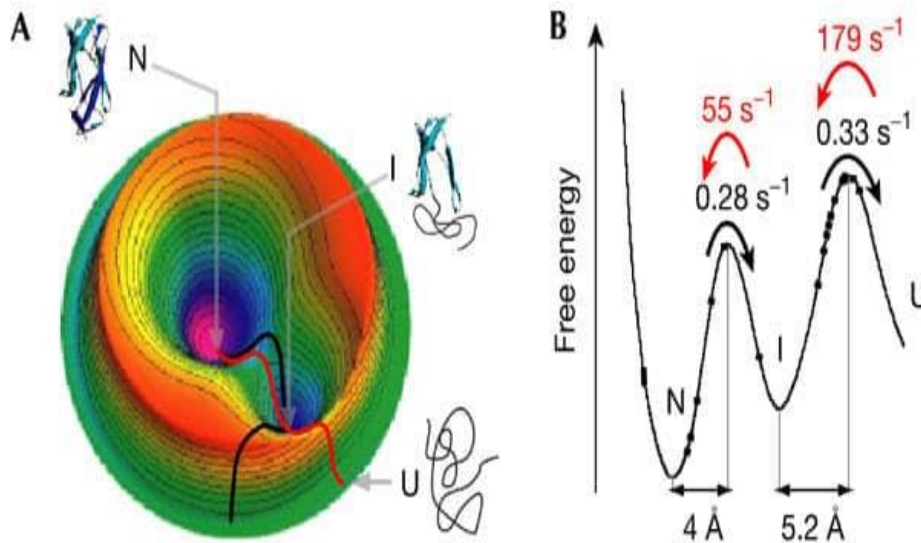
equilibrio



stato più popolato

Minimizzazione dell'energia

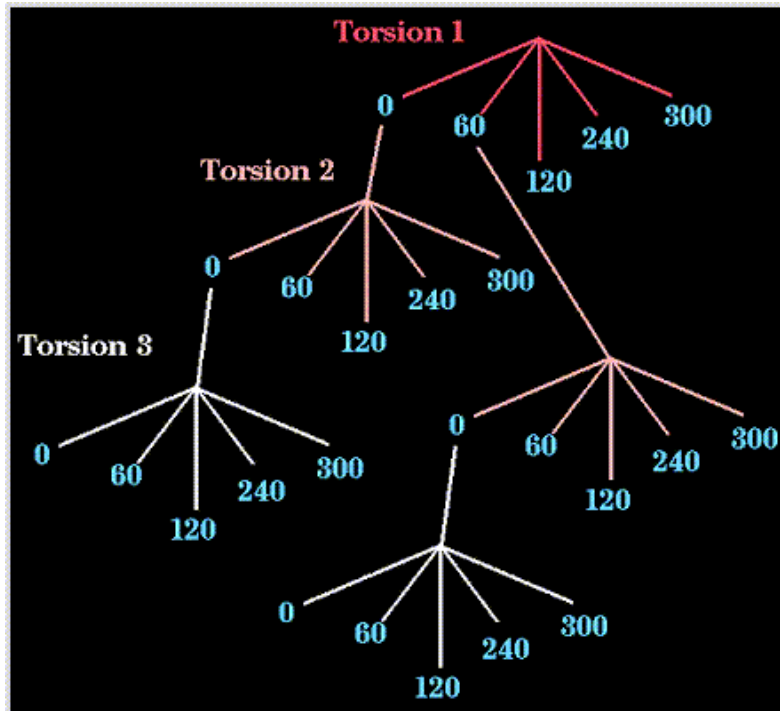
Minimizzare l'energia potenziale di una molecola significa trovare un percorso (costituito dalle variazioni dei gradi di libertà intramolecolari) che conduca da una conformazione iniziale alla **conformazione a minima energia più vicina** (MINIMO LOCALE), usando il **minor numero di calcoli** possibile.



Strategie di ricerca del minimo globale

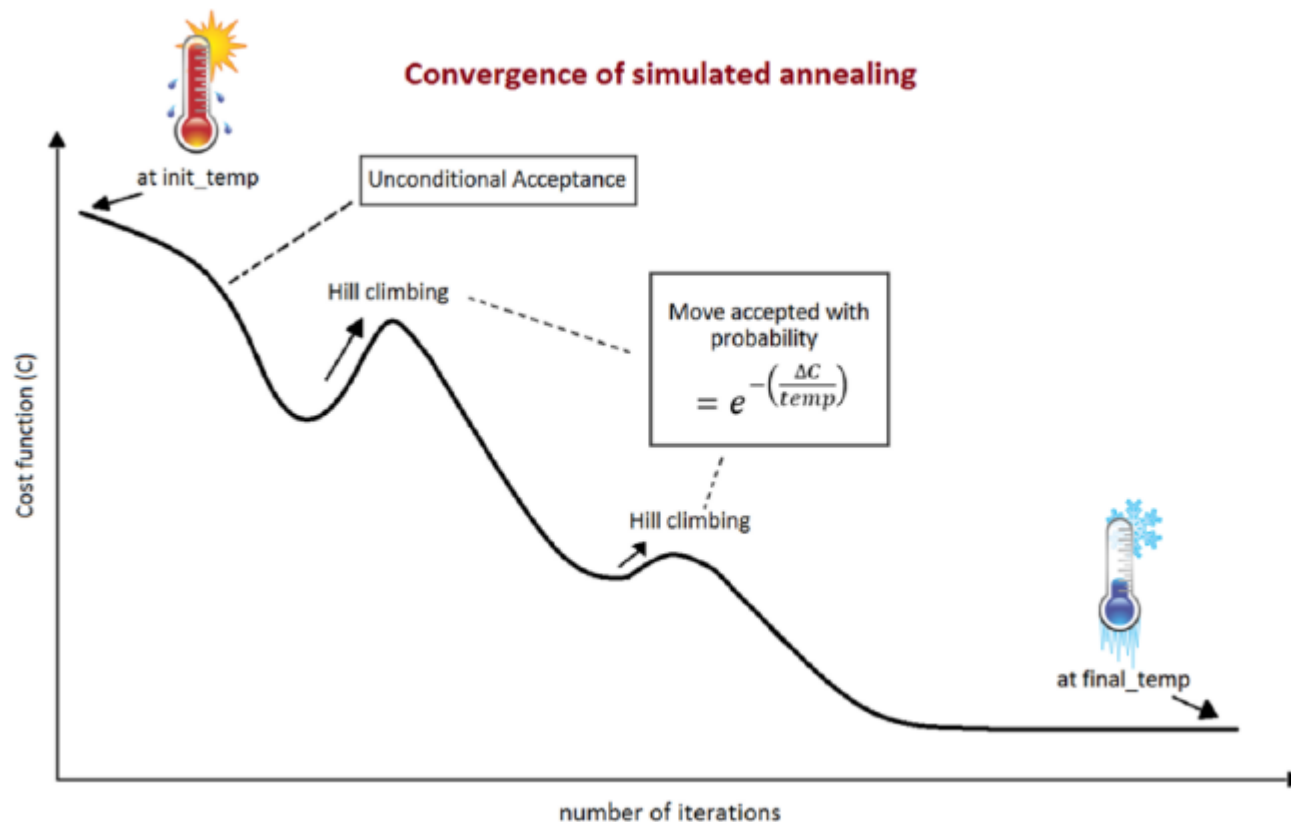
➤ Campionamento energetico sistematico:

L'energia viene campionata ad intervalli regolari sull'intera estensione di ciascun grado di libertà (tipicamente rotazioni dei legami).



➤ Annealing simulato:

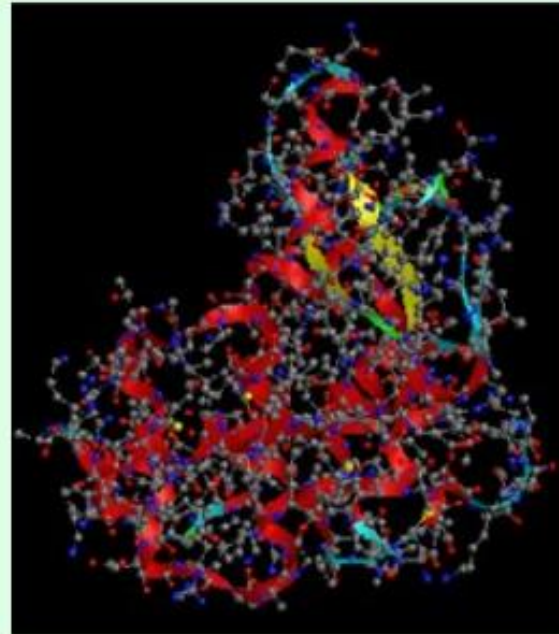
riscaldamento ad alta temperatura che modifica la struttura superando le barriere energetiche, seguito da un raffreddamento lento.



E' impossibile visualizzare l'intera superficie di energia potenziale, tranne che in casi molto semplici in cui l'energia è funzione di solo una o due coordinate.

Lisozima T4, 1328 atomi circa
Superficie di energia molto complicata.

Ipersuperficie 3984-dimensionale!!



Ricerca del minimo (minimizzazione o ottimizzazione)

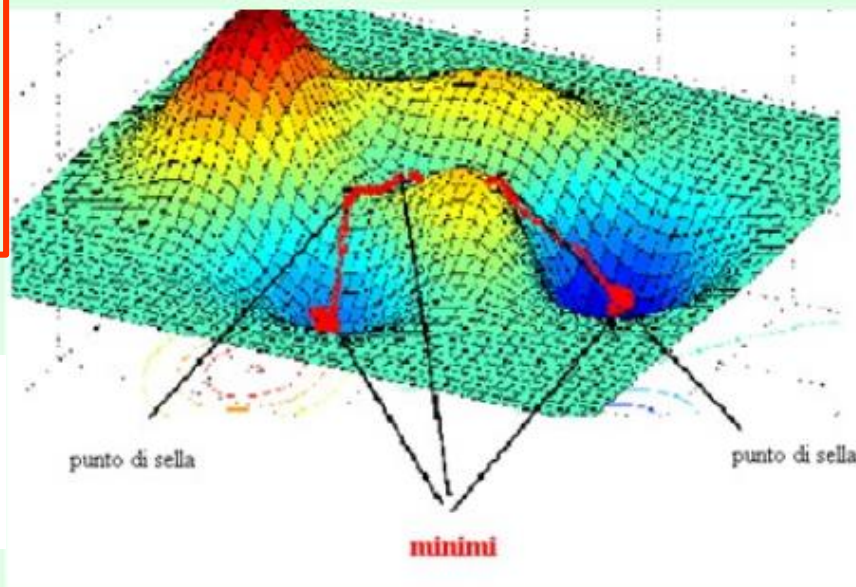
Quello della ricerca di un algoritmo di minimizzazione efficace è un problema molto noto in matematica. Nei casi complessi per determinare i minimi sulla superficie di energia potenziale in genere si ricorre ad algoritmi matematici basati su metodi numerici. C'è una vasta letteratura su questi metodi, qui ci soffermeremo su quegli approcci più comunemente usati in meccanica molecolare.

Punti di minimo

Un punto di minimo oltre ad avere le derivate prime rispetto a tutte le coordinate uguali a zero deve avere tutte le derivate seconde positive.

$$\frac{\partial V}{\partial r_i} = 0; \quad \frac{\partial^2 V}{\partial r_i^2} > 0$$

Se la derivata seconda è >0 allora è un punto di minimo



Ricerca dei punti di minimo

In termini matematici il problema è dunque quello di minimizzare una funzione -appunto l'energia- di più variabili -le coordinate interne degli atomi.

Qualora l'energia di un sistema di particelle possa essere rappresentata da una funzione analitica, il minimo può essere trovato con metodi di calcolo standard. Questo non è però generalmente possibile per sistemi molecolari complicati nei quali l'energia varia con le coordinate in maniera complessa. In questi casi i minimi sono localizzati usando metodi numerici nei quali le coordinate vengono variate gradualmente producendo configurazioni con energie sempre più basse fino al raggiungimento del minimo.

Scelta dell'algoritmo.

Ci sono molti fattori che determinano la scelta dell'algoritmo più appropriato per un dato problema:

- veloce
- robusto
- che richiede poca memoria

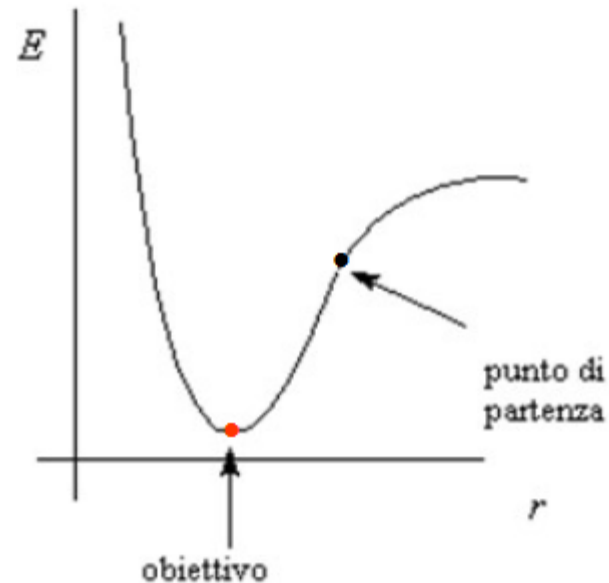
La **robustezza** indica la capacità di un metodo di non risentire degli effetti delle variazioni operative e della configurazione del sistema.

Quindi può considerarsi come la capacità del metodo di non essere significativamente influenzato da variazioni delle condizioni analitiche.

Algoritmi di minimizzazione

Si possono classificare gli algoritmi di minimizzazione in due gruppi:

- 1) quelli che non usano le derivate dell'energia rispetto alle coordinate
- 2) quelli che usano le derivate dell'energia rispetto alle coordinate.



Criteri di Convergenza

A differenza di quello che succede con le funzioni analitiche nelle applicazioni "reali" di molecular modelling è possibile identificare l'esatta posizione dei minimi solo raramente, possiamo solo sperare di raggiungere una buona approssimazione. E' necessario quindi avere dei criteri per decidere quando il calcolo di minimizzazione è sufficientemente vicino ad un minimo in modo da poter essere terminato.

Ogni calcolo è ovviamente limitato dalla precisione con cui vengono immagazzinati i numeri, ma nella maggior parte dei casi ci si ferma ben prima di questo limite.

Ci sono vari criteri tra cui scegliere (o applicare quello che si verifica per primo). Ci si ferma quando:

- si raggiunge un massimo numero di iterazioni
- quando la differenza di energia tra due passi successivi è minore di una certa tolleranza
- quando il passo (cioè la differenza tra due configurazioni successive) è piccola.

Scelta del punto di partenza

Dipende da quello che vogliamo ottenere:

Se vogliamo ottenere la struttura di minima energia (minimo globale) conviene partire da una struttura che probabilmente è già vicina a quel minimo:

- es. struttura a raggi X (dai database)
- se vogliamo utilizzare i metodi della meccanica quantistica (più pesanti) ci conviene partire da una struttura già minimizzata con i metodi della meccanica molecolare.

Se vogliamo la geometria di tutte le strutture a bassa energia (più minimi locali) allora possiamo partire da diverse strutture casuali ("random"). (

Minimizzazione con l'uso di derivate

Le derivate danno informazioni riguardo la forma della superficie di energia potenziale e se ben usate possono aumentare in maniera significativa l'efficienza nella localizzazione del minimo. Pertanto esse, qualora siano disponibili, vengono impiegate dai più comuni metodi di minimizzazione.

La direzione della derivata prima dell'energia (il gradiente) indica dove abbiamo un minimo visto che la grandezza del gradiente è relazionata alla pendenza della curva. Poiché la forza cui è sottoposto un atomo è uguale al gradiente cambiato di segno, l'energia del sistema può essere abbassata dal movimento di ciascun atomo in risposta all'azione della forza agente su di esso.

La derivata seconda indica la curvatura della funzione, e può essere usata per predire dove la funzione cambierà direzione. Per usare questo tipo di metodi è ovviamente necessario poter calcolare le derivate dell'energia rispetto alle variabili.

Le derivate possono essere ottenute analiticamente o numericamente. L'uso di derivate analitiche è preferibile perché sono esatte ed in più possono essere calcolate più velocemente. In alcuni casi può essere più efficace un algoritmo di minimizzazione non basato sulle derivate e fare ricorso a derivate numeriche

Minimizzazione con l'uso di derivate: metodo "steepest descent"

(Cauchy, 1847)

Il metodo di discesa del gradiente (detto anche metodo del gradiente, metodo **steepest descent** o metodo di discesa più ripida) è una tecnica che consente di determinare i punti di massimo e minimo di una funzione di più variabili.

Nel metodo "steepest descent" ci si muove nella direzione parallela alla forza netta, che equivale a muoversi in linea retta per la discesa. Il punto d'inizio per ogni iterazione k è la configurazione molecolare ottenuta dal passaggio precedente che è rappresentata da un vettore multidimensionale $\mathbf{x}_{k-1}$ che contiene le coordinate di tutti gli atomi del sistema.

Per la prima iterazione il punto di partenza è la configurazione iniziale del sistema, il vettore $\mathbf{x}_1$ che è fornito dall'utente. La direzione dello spostamento dal punto iniziale è dunque definita dal gradiente $\mathbf{g}_k$ ed è data da un vettore unitario 3N-dimensionale (in coordinate cartesiane) $\mathbf{s}_k$:

$$\mathbf{s}_k = \frac{-\mathbf{g}_k}{|\mathbf{g}_k|}$$

Minimizzazione con l'uso di derivate: metodo "steepest descent"

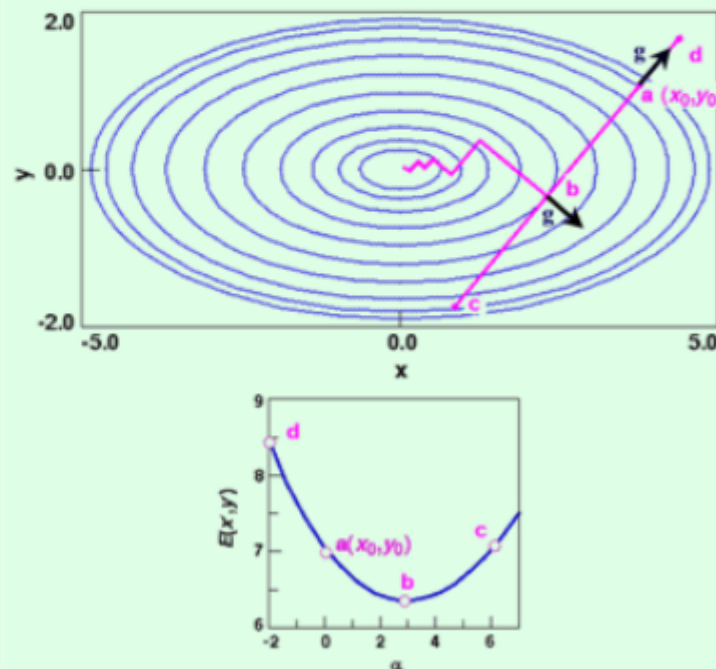
Dopo aver definito la direzione lungo la quale muoversi occorre decidere quanto muoversi lungo il gradiente.

Ci sono due metodi:

(in pratica si minimizza in una dimensione)

$$\nabla V = \left(\frac{\partial V}{\partial x_i}, \frac{\partial V}{\partial y_i}, \frac{\partial V}{\partial z_i} \right)$$

$$\mathbf{p}_k := -\nabla f(\bar{\mathbf{x}}).$$



la **direzione di massima discesa** in un assegnato punto corrisponde a quella determinata dall'opposto del suo **gradiente** in quel punto.

Questa scelta per la direzione di discesa garantisce che la soluzione tenda a un punto di minimo di f .

Il metodo del gradiente prevede dunque di partire da una soluzione iniziale $\mathbf{x}_0$ scelta arbitrariamente e di procedere iterativamente aggiornandola come:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

Minimizzazione con l'uso di derivate: metodo "conjugate gradient"

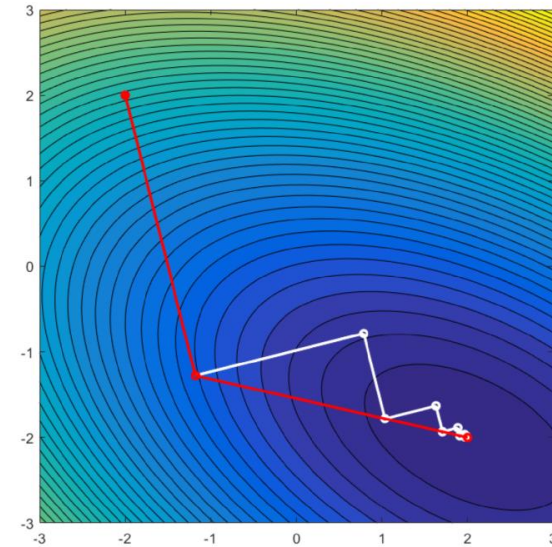
Il metodo è stato inizialmente proposto nel 1952 dai matematici [Hestenes](#) e [Stiefel](#) e costituisce una variante del metodo del gradiente in cui la direzione di discesa a ogni passo è scelta in modo tale da garantire la convergenza del metodo

In analisi numerica, il metodo del gradiente coniugato è un algoritmo per la risoluzione numerica di un sistema lineare la cui matrice sia simmetrica

Il metodo "conjugate gradient" produce un insieme di direzioni che non mostra il comportamento oscillatorio del metodo "steepest descent". Il metodo *conjugate gradient* usa un algoritmo che produce una serie di line search con direzioni mutuamente coniugate così che ad ogni successivo step si ha un raffinamento della direzione verso il minimo.

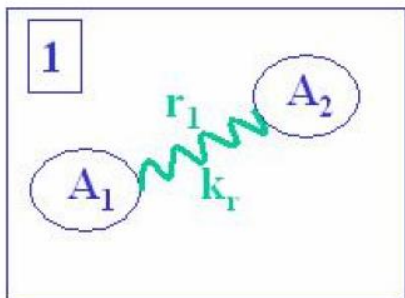
Minimizzazione con l'uso di derivate: metodo al secondo ordine

Sono i metodi che fanno uso di derivate seconde (matrice Hessiana), oltre alle derivate prime. Le derivate seconde danno informazioni sulla curvatura della superficie. Tra questi metodi ricordiamo il [Newton-Raphson](#) e varianti su questo metodo, che introducono semplificazioni nel calcolo della matrice Hessiana.

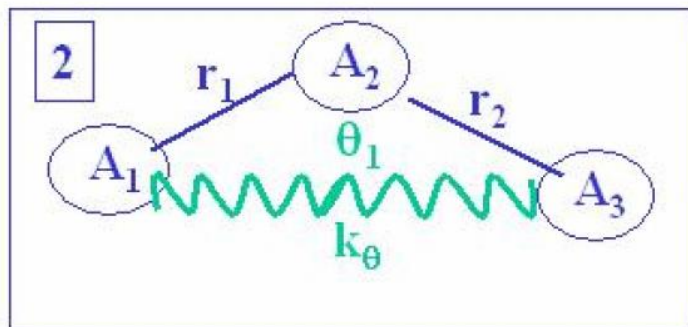


Applicazioni della minimizzazione

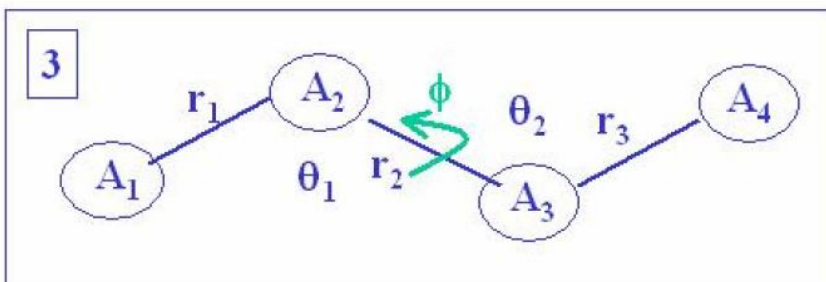
- per trovare il minimo globale
- per preparare il sistema per altre simulazioni:
 - rilassare le conformazioni ad energia più alta
 - avere una geometria di partenza buona
- ricerca conformazionale (conformational search)



$$U(r) = \frac{1}{2} K_r (r - r_0)^2$$



$$U(\vartheta) = \frac{1}{2} K_\vartheta (\vartheta - \vartheta_0)^2$$



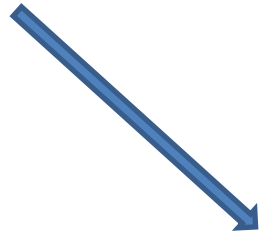
$$U(\varphi) = \sum_k c_k \cos^k \varphi \approx \frac{1}{2} [1 + \cos(3\varphi)]$$

Potenziale atomi legati

Potenziale atomi non legati

$$E = \sum_i^{\text{bonds}} k_i (r_i - r_{i,e})^2 + \sum_i^{\text{angles}} k_i (\theta_i - \theta_{i,e})^2 + \sum_i^{\text{some dipoles}} v_i [1 \pm \cos(n_i \phi_i)]^2 + \sum_i^{\text{atoms}} \sum_{j>1}^{\text{atoms}} \left(-\frac{A_i A_j}{r_{ij}^6} + \frac{B_i B_j}{r_{ij}^{12}} + \frac{q_i q_j}{r_{ij}} \right) + E_{HB}$$

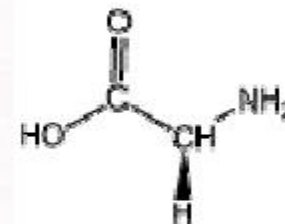
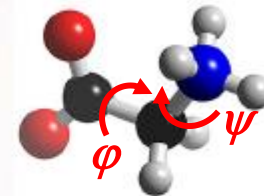
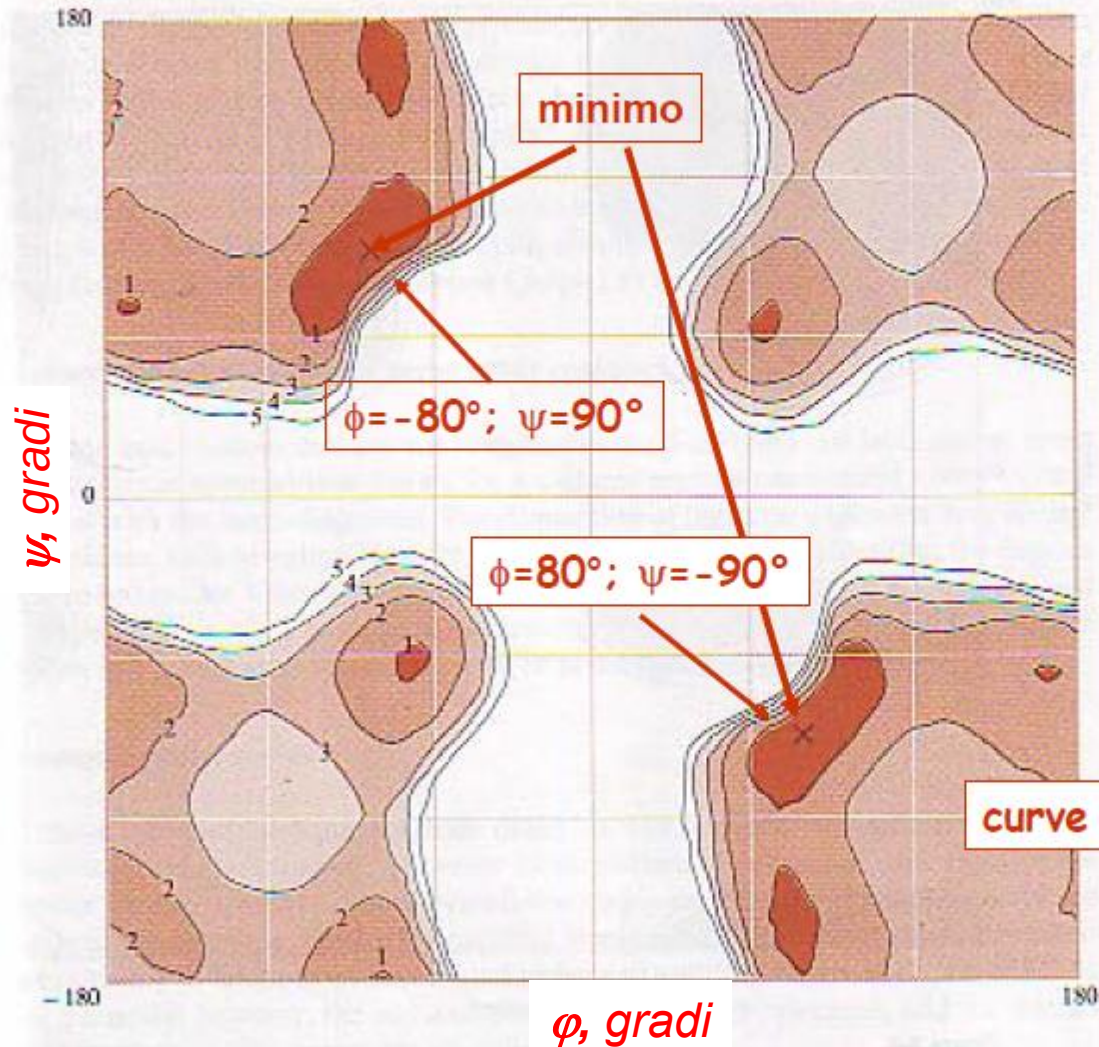
$E_{\text{bond}} + E_{\text{ang}} = \text{costante}$



Rigid geometry approximation

Mappe di Ramachandran

$$E(\varphi_i, \psi_i) = E_{LJ}(\varphi_i, \psi_i) + E_C(\varphi_i, \psi_i) + E_{tor}(\varphi_i, \psi_i)$$

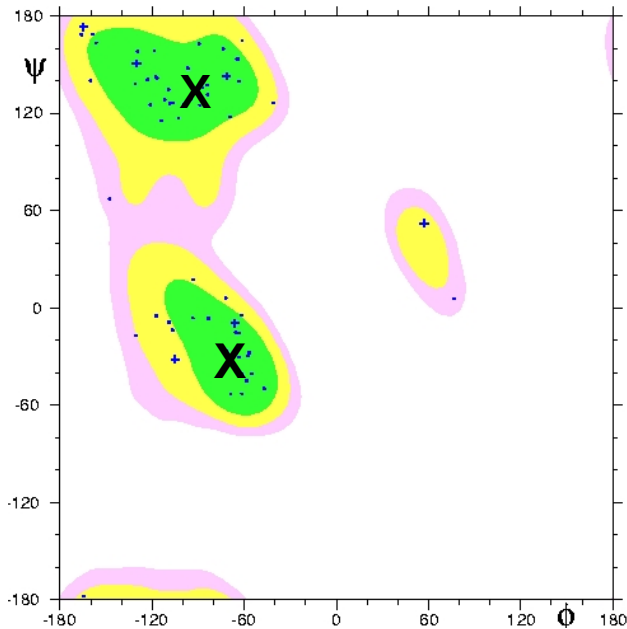
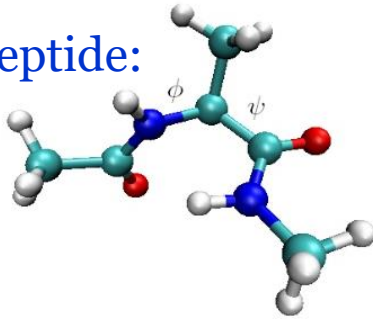


Glicina

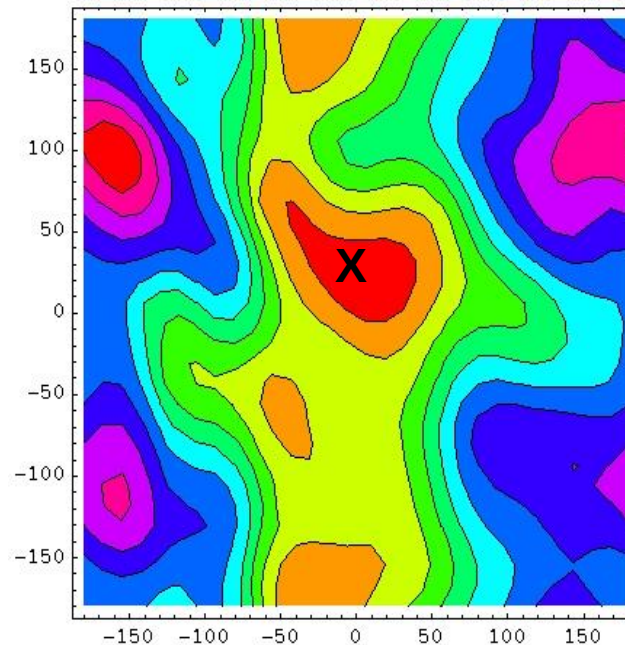
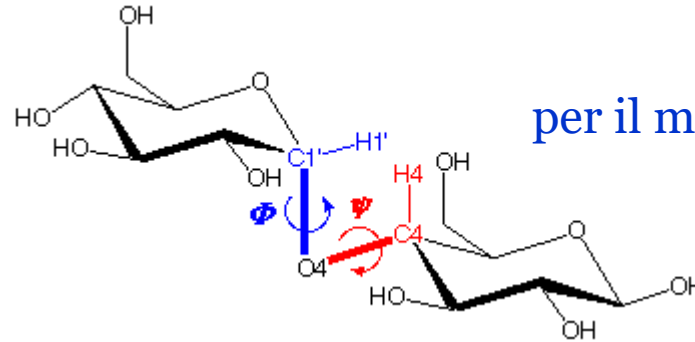
La libertà conformazionale è data dalle regione a bassa energia delle **mappe di Ramachandran**.

Esempio: le mappe relative a disaccaridi sono molto diverse da quelle relative a dipeptidi

per un dipeptide:



per il maltosio:



*La probabilità che un dato stato conformazionale venga popolata ad una certa temperatura è data dalla **Distribuzione di Boltzmann**:*

$$p(\phi) = \frac{\exp\left[-\frac{E(\phi)}{RT}\right]}{\int_0^{2\pi} \exp\left[-\frac{E(\phi)}{RT}\right] d\phi}$$

$$\int_0^{2\pi} p(\phi) d\phi = 1$$

Simulated Annealing

- L'algoritmo è così chiamato perché simula l'annealing, cioè il processo *graduale* di raffreddamento di una sostanza che passa dallo stato liquido allo stato solido.
- Se la temperatura viene diminuita abbastanza lentamente, allora il liquido raggiunge una configurazione a minima energia.

Simulated Annealing

A ciascuna iterazione del ciclo principale, un nuovo punto, è generato casualmente all'interno del dominio indagato e la sua distanza da quello creato in precedenza è dato dal valore assunto dalla temperatura in quel momento.

L'algoritmo accetta sicuramente tutti i nuovi punti che migliorano il valore corrente della funzione da ottimizzare e, con una certa probabilità, le coordinate in corrispondenza delle quali si ottiene uno stato energetico ad energia superiore

Questo accorgimento permette alla tecnica di evitare di rimanere intrappolata in minimi locali, consentendo un'esplorazione più accurata dello spazio indagato soprattutto nella fase iniziale, ovvero a temperature elevate.

Con l'avanzare delle iterazioni uno schema di raffreddamento (cooling schedule) decrementa la temperatura, lo spazio di esplorazione delle migliori soluzioni diviene sempre meno esteso e la probabilità di accettare soluzioni che peggiorano la funzione obiettivo tende a diventare sempre più bassa, fintanto che si giunge a convergenza con l'individuazione della corretta zona di stazionarietà.

SIMULATED ANNEALING

L'**algoritmo SA** per un problema di ottimizzazione può essere riassunto nel seguente generico algoritmo:

I. Sia data una configurazione iniziale o soluzione x_0 con *energia* o valore della funzione obiettivo E_0 . Selezionare un valore iniziale T_0 per la temperatura.

II. Per ogni stadio della temperatura effettuare i seguenti passi :

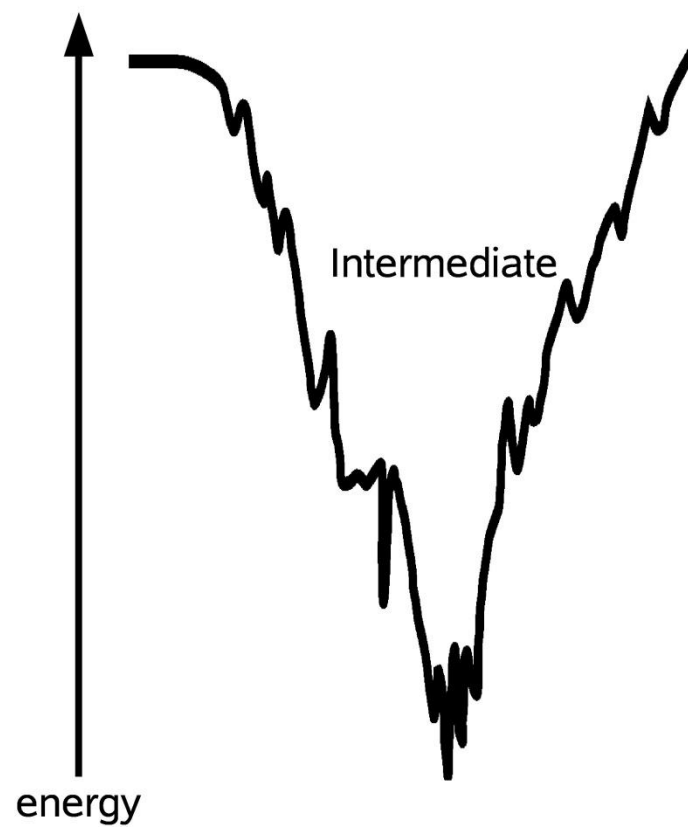
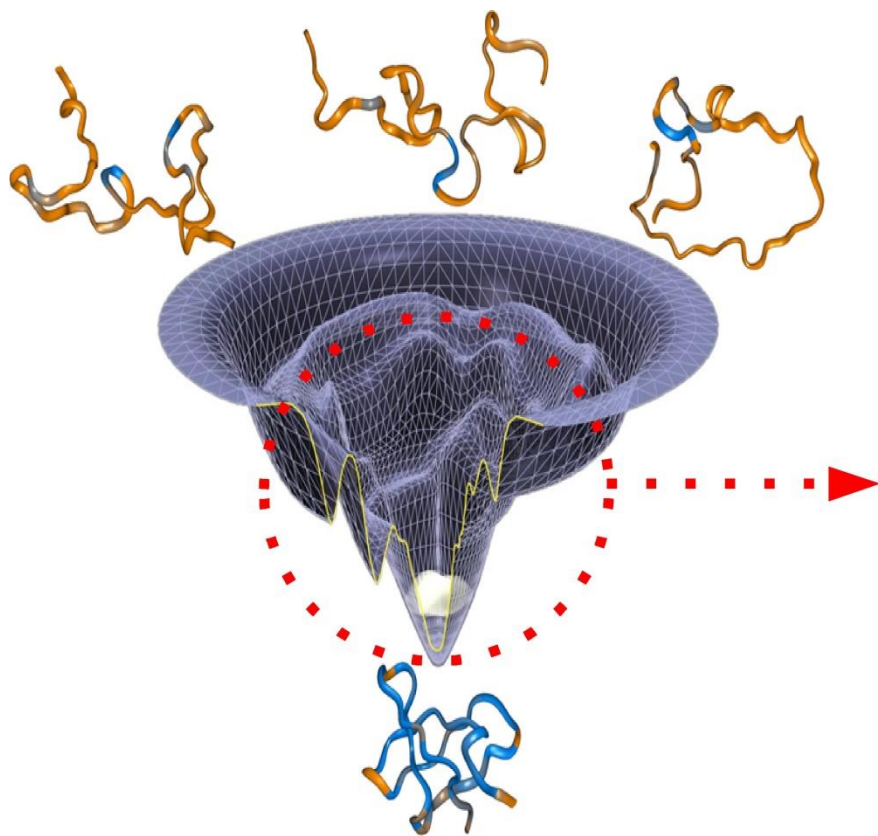
- **A.** Generare una configurazione candidato *ammissibile* tramite **una piccola perturbazione casuale della configurazione corrente**.
Valutare la differenza di energia ΔE fra le due configurazioni.
- **B.** Se $\Delta E < 0$, la configurazione candidato ha un valore della funzione obiettivo inferiore rispetto a quello della configurazione corrente. Accettare la nuova soluzione e sostituirla a quella corrente.
Se $\Delta E > 0$, la configurazione candidato ha un valore della funzione obiettivo peggiore rispetto quello della configurazione corrente.
Accettare tale soluzione con una probabilità $P(\Delta E)$

$$P(\Delta E) = \exp\left(-\frac{\Delta E}{k_b T}\right)$$

e aggiornare la configurazione corrente, se necessario.

- **C.** Se non è raggiunto l'equilibrio termico, torna al punto **A**. Altrimenti vai al punto **III**.

III. Se il processo di annealing è incompleto, ridurre la temperatura e ripartire dal punto **II**.



CALCOLO CONFORMAZIONALE

Individuazione della
struttura molecolare
rappresentativa della catena

scelta del metodo:
- calcolo "ab initio"
- calcolo semiempirico
- meccanica molecolare
- dinamica molecolare

scelta del campo di
forze per E_{tot}

scelta delle variabili
descrittive del sistema

**PREDIZIONE DELLE
STRUTTURE ORDINATE**
(stato cristallino, struttura
secondaria dei biopolimeri,...)

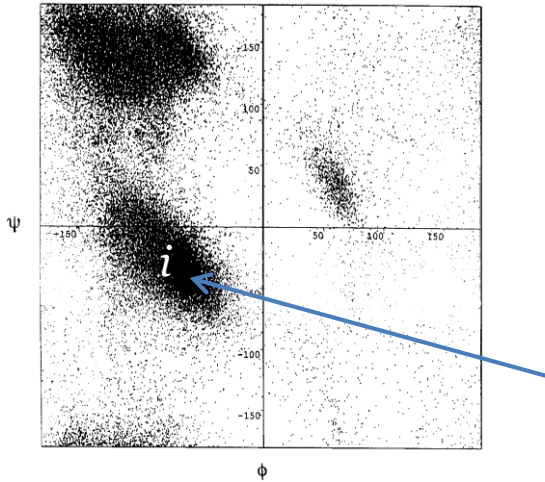
**STATISTICA
CONFIGURAZIONALE**
Stato disordinato (amorfo, in
soluzione)

scelta dei legami descrittivi
della catena (reali o virtuali)

scelta del modello probabilistico:
- catena Bernoulliana
- catena Markoviana
- catena di Ising

scelta del metodo:
- metodo matriciale di Flory
- metodo Monte Carlo

Dall'insieme di conformazioni accessibili (energie) alla valutazione delle proprietà (medie) delle catene (dimensione e forma, osservabili termodinamiche)



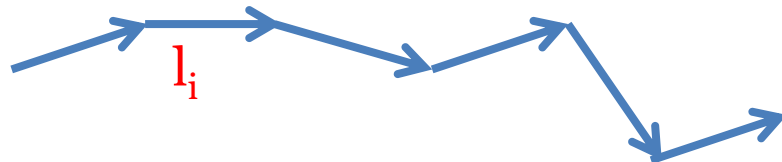
Energie degli stati conformazionali
 $E_i = E(\phi_i, \psi_i)$

Ad ogni stato viene assegnata una probabilità

$$P_i(\phi_i, \psi_i) = \frac{e^{-\frac{E_i}{RT}}}{\sum_i e^{-\frac{E_i}{RT}}}$$

Catena polimerica

$$\{(\phi_1, \psi_1)(\phi_2, \psi_2) \dots (\phi_n, \psi_n)\}$$



**Dagli stati conformazionali $E(\phi_i, \psi_i)$
alla predizione della topologia di catena**



**Metodi stocastici per la predizione
della topologia di macromolecole
nello stato disordinato**

I **modelli deterministici** sono i più semplici; in essi, le variabili di **input** **assumono valori fissi**.

E' vero che i risultati (*output*) generati da questi modelli possono tener conto, entro certi limiti, della *variabilità* e dell'effetto del caso; tuttavia, si tratta sempre di elaborazioni di tipo deterministico, in quanto non si tiene in considerazione l'incertezza associata alle variabili di *input*.

i **modelli stocastici** (stocastico = dovuto al caso, aleatorio, dal greco *stochastikòs*=congetturale) **tengono in considerazione le variazioni (causali e non) delle variabili di input**, e quindi forniscono risultati in termini di "**probabilità**".

È importante sottolineare che ciò che differenzia i modelli deterministici da quelli stocastici è che in questi ultimi si tiene conto della variabilità dei dati di input.

In genere i modelli stocastici hanno una struttura più complessa di quelli deterministici.

I metodi stocastici hanno trovato larga diffusione in vari settori, (scientifici e non):

Fisica (metodi quantistici, reattori nucleari, aerodinamica, ottica...),

Astrofisica (modelli di evoluzione di galassie),

Biologia (sistemi biologici),

Finanza (calcolo del valore di una compagnia, per prezzare i derivati e le opzioni non standard),

Ingegneria (variazioni in circuiti integrati, fluidodinamica,...),
altri.....

altri settori:

Traffico stradale,

Videogiochi, metodi scacchistici, ...

Meteorologia,

Progettazione grafica, ecc...

Disordine :

- Disordine nelle velocità'
- Disordine nelle posizioni:
 - Liquidi
 - Polimeri
 - Solidi disordinati (vetri, amorfi)

la distribuzione di Maxwell-Boltzmann

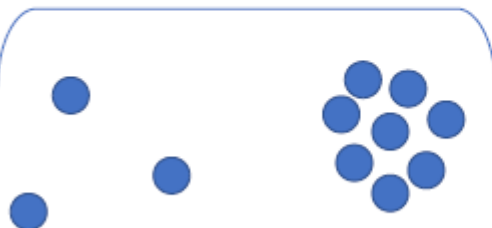
$$f_{MB}(v) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left[-\frac{mv^2}{2kT}\right]$$



f_{MB} predetta in sistemi in equilibrio classico, potenziali dipendenti dalla sola posizione e velocità non relativistiche. Confermata in:

Fluidi (alta mobilità')

Solidi (mobilità' trascurabile)



Gas ideali

Liquidi



Cristalli

Vetri

NB:

Le leggi dei sistemi disordinati sono di natura statistica

Liquidi: il moto browniano e l'esistenza degli atomi



J. Perrin

Jean Baptiste Perrin, The Nobel Prize in Physics 1926
"for his work on the discontinuous structure of matter..."

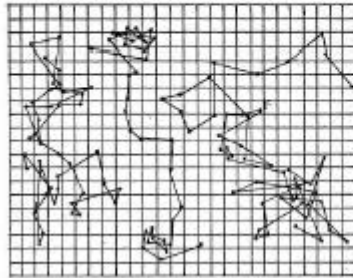


Fig. 7.



A. Einstein

Articoli 1905 (Annus Mirabilis), in ordine di invio:

- Effetto fotoelettrico (Nobel 1921)
- **Moto Browniano**
- Relatività speciale
- $E=mc^2$

$$\langle \mathbf{R}^2(t) \rangle = 6 D t$$

$$k_B = 6\pi a \frac{D \eta}{T}$$

D: coefficiente di diffusione
 η : viscosità

Legge di Stokes-Einstein

*Esistono diversi **modelli statistici** che vengono usati nella descrizione dei processi stocastici relativi a sistemi polimerici*

- La teoria dei processi stocastici riguarda lo studio di sistemi che evolvono nel tempo (ma anche più in generale nello spazio) **secondo leggi probabilistiche.**
- Tali sistemi o modelli descrivono **fenomeni complessi** del mondo reale che hanno la possibilità di essere **aleatori.**
- Tali fenomeni sono più frequenti di quanto si possa credere e si incontrano tutte quelle volte in cui le quantità alle quali siamo interessati **non sono prevedibili con certezza.**

Polimeri: analogia con il moto browniano

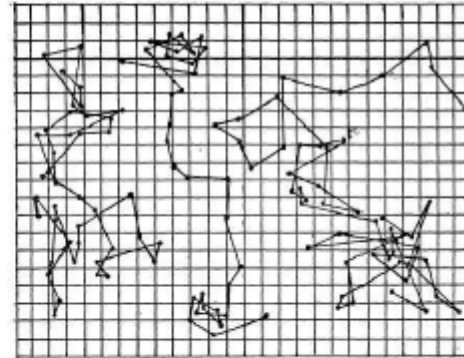
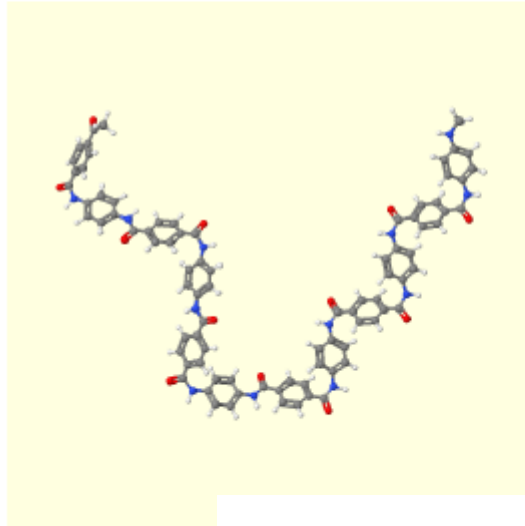
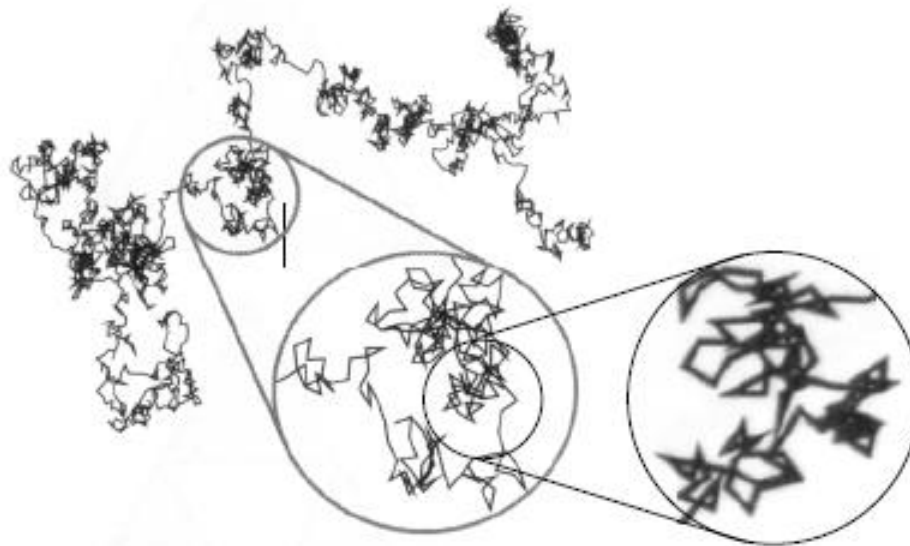


Fig. 7.



Alcuni esempi di modelli statistici sono:

- 1) **catena Bernoulliana**: in questo modello la probabilità di avere una determinata conformazione in catena è **indipendente** dallo stato conformazionale precedente e tale evento non ha alcun effetto sulla conformazione del residuo successivo in catena;
- 2) **catena Markoviana**: in una catena Markoviana di ordine **n** la probabilità di avere uno stato conformazionale ad un certo DP in catena è **dipendente dagli n stati** conformazionali che lo precedono. Quando $n=0$ la catena Markoviana coincide con quella Bernoulliana;
- 3) **catena di Ising**: la probabilità di avere un dato stato ad un certo punto della catena dipende dallo stato conformazionale dei **residui primi vicini**.

Metodo Monte Carlo

(Nicholas Constantine Metropolis)

Sono metodi probabilistici utilizzati per ottenere stime di grandezze d'interesse attraverso estrazioni di numeri casuali. Il metodo è usato per trarre stime attraverso simulazioni.



Con metodo (di) Monte Carlo si intende una classe di algoritmi computazionali che, a partire da un insieme di **numeri generati casualmente** o pseudo-casualmente, permetta di ottenere un risultato medio molto vicino al comportamento teorico.

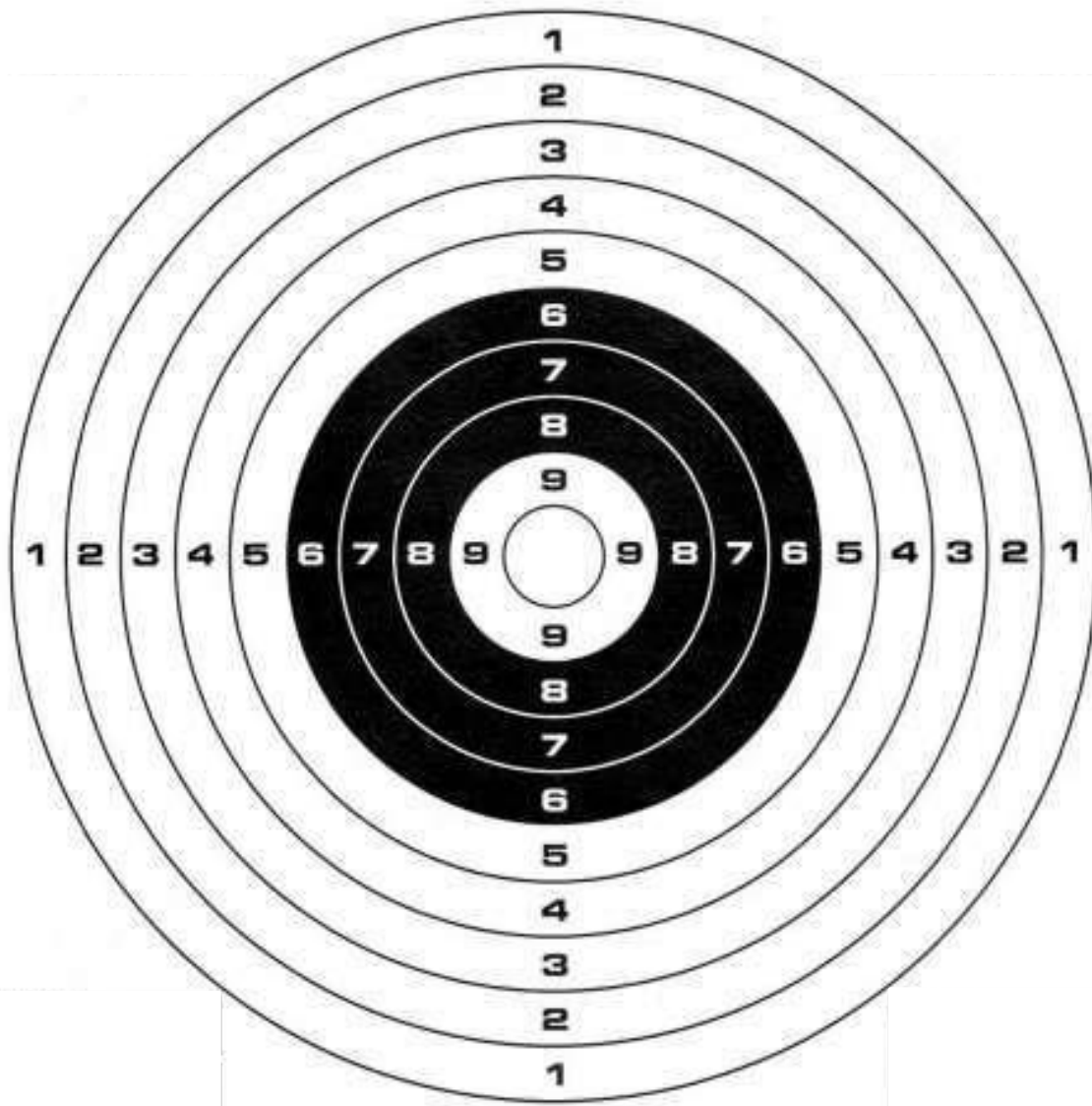
Prevede dunque la risoluzione di complicati problemi deterministici con metodi probabilistici.

Prende particolarmente piede, in campo scientifico, laddove l'osservazione diretta del fenomeno non sia più sufficiente per arrivare alla formulazione di un modello teorico (per es. in fisica subatomica).

Con il metodo Monte Carlo (MC) **si simula un processo stocastico associando agli stadi che si susseguono nel processo, una probabilità *a priori*.**

Per fare questo vengono generati dei **numeri casuali (random)** e confrontati con le **probabilità assegnate *a priori*** ai vari possibili eventi, in modo da determinare quale di questi avviene a quel particolare stadio del processo.

La simulazione Monte Carlo calcola una **serie di realizzazioni possibili** del fenomeno in esame, con il **peso** della probabilità di tale evenienza, cercando di esplorare in modo «denso» tutto lo spazio dei parametri del fenomeno



In generale viene generata una serie di **N** numeri random che determina un campione di popolazione di tutte le possibili conformazioni della catena ognuna **pesata** mediante l'opportuno **fattore probabilistico di Boltzmann** chiamato ***peso statistico*** :

$$q = \sum_{k=1}^M \exp \left[-\frac{E_k}{kT} \right]$$

Funzione di partizione

$$p_j = \frac{e^{-\frac{E_j}{kT}}}{q}$$

Probabilità di evento
dello stato j-esimo

Generatori Lineari Congruenziali (LCG)

Uno dei metodi matematici più noti per generare numeri casuali distribuiti uniformemente è rappresentato dai generatori lineari **congruenziali** (**numeri pseudo-casuali**)

Tali generatori vennero presentati per la prima volta da D.H. Lemer negli anni '40.

Il metodo LCG, come tutti i generatori di numeri casuali, ha bisogno di un **seed** (seme) per inizializzare la sequenza di numeri secondo la regola

$$x_{n+1} = (a x_n + c) \bmod m, \quad n \geq 0,$$

dove **a**, **c** ed **m** sono opportuni numeri interi costanti. La notazione “mod m”, ossia **modulo m**, significa che $ax_n + c$ viene diviso per **m** e x_{n+1} **posto uguale al resto intero della divisione**. Quindi x_{n+1} assume valori interi tra 0, 1, 2, ..., m-1.

Algoritmo di generazione di numeri random

$$x_i = (N_1 x_{i-1} + N_2) \bmod [(N_1 x_{i-1} + N_2), N]$$

$$\bmod [(N_1 x_{i-1} + N_2), N] = (N_1 x_{i-1} + N_2) - N \cdot \text{int} \left[\frac{(N_1 x_{i-1} + N_2)}{N} \right]$$

$$0 \leq N_1, N_2 < N$$

dove N_1 , N_2 ed N sono numeri interi ed $\text{int}[\dots]$ è una funzione che prende la parte intera del rapporto espresso in parentesi. Se come esempio consideriamo $x_0=4$, $N_1=5$, $N_2=3$ ed $N=8$, si ottiene la

ESEMPIO

Se come esempio consideriamo $x_0=4$, $N_1=5$, $N_2=3$ ed $N=8$, si ottiene la seguente serie di numeri:

$\{4, 7, 6, 1, 0, 3, 2, 5, 4, 7, 6, 1, \dots\}$

che si ripete ciclicamente dopo 8 numeri. La periodicità (numeri pseudo-random) è in genere uguale ad N , cosicchè prendendo N molto grande, la lunghezza del ciclo diventa più lunga della dimensione della serie di numeri random che usualmente si richiede nella maggioranza delle simulazioni Monte Carlo.

In effetti nessun calcolatore è in grado di generare **numeri puramente casuali**, ma solo numeri *pseudo-casuali*. Si tratta di numeri generati da algoritmi deterministici in grado di superare una serie di test statistici che conferiscono a tali numeri una **apparente casualità**.

Il problema della scelta dei migliori valori per **a**, **c** ed **m** è quindi il punto cruciale del metodo.

Un aspetto importante è la **lunghezza del periodo** della successione da cui segue che **m** dovrà essere molto grande.

Inoltre per un dato **m**, i valori di **a** e **c** dovranno essere tali che la successione abbia periodo massimo.

Una delle scelte più *popolari* è

m=2³¹-1,
a=75,
c=0.

Questo garantisce **un periodo di oltre due miliardi di numeri casuali.**

Metodo middle-square

- Supponiamo di voler generare un numero casuale di 4 cifre
- Il metodo richiede come tutti i generatori di numeri casuali un valore iniziale, detto **seme**, dal quale vengono generati i successivi valori
- Ad es. a partire da 1234 avente 4 cifre
 - eleviamo al quadrato e otteniamo le 8 cifre 01522756
 - consideriamo solo le 4 cifre di mezzo 5227
 - ripetiamo il procedimento ottenendo 27321529 e 3215 e così via
- Ogni nuovo numero è determinato univocamente dal predecessore. Ogni successione di numeri generata da questo algoritmo si ripeterà prima o poi. Il numero di numeri della sequenza prima che intervenga una ripetizione è detta **periodo** della sequenza

N.B. Suggesto da John von Neumann nei primi anni dell'era dei computer (1946) per generare numeri pseudo-casuali distribuiti in modo uniforme.

Generatori più recenti: il Cristoforo Colombo

I generatori di numeri casuali più recenti, come quello incluso nelle versioni di MATLAB sono una **combinazione di operazioni di spostamento di registri** e manipolazione sui bit che non richiedono nessuna operazione di moltiplicazione o divisione.

Questo nuovo approccio risulta estremamente veloce ed inoltre garantisce periodi di sequenza incredibilmente lunghi.

Nelle ultime versioni MATLAB, **il periodo è 2^{1492}** , che ad un milione di numeri casuali al secondo richiederebbe **10^{435}** anni (circa 10^{425} volte l'età dell'universo) prima di ripetersi!

Data la coincidenza dell'esponente con la data della scoperta dell'America (12 ott. **1492**), questo generatore viene comunemente chiamato **il generatore di “Cristoforo Colombo”**.

Controlli statistici

REQUISITI DEI GENERATORI DI NUMERI RANDOM

1) uniformità: gli n numeri random devono essere uniformemente distribuiti nell'intervallo $[0,1]$ e tale proprietà viene stimata attraverso i momenti della distribuzione, ossia:

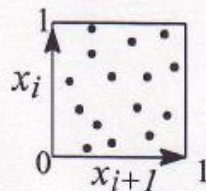
$$m_1 = \frac{1}{n} \sum_i^n x_i = 0.500$$

$$m_2 = \frac{1}{n} \sum_i^n x_i^2 = 0.333$$

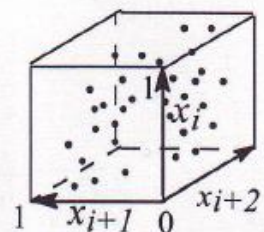
$$m_3 = \frac{1}{n} \sum_i^n x_i^3 = 0.250$$

2) assenza di correlazione: un test molto semplice di correlazione dei numeri della serie è quello di analizzare copie (x_i, x_{i+1}) e terne (x_i, x_{i+1}, x_{i+2}) nell'insieme di numeri random e vedere con quale frequenza si distribuiscono, rispettivamente, nel piano e nello spazio. Per un insieme non-correlato i punti devono risultare uniformemente distribuiti.

(x_i, x_{i+1})



(x_i, x_{i+1}, x_{i+2})



- a) Test di equidistribuzione (o di frequenza) per verificare la reale uniformità dei numeri generati. Essi vengono disposti in un istogramma di k canali e, si controlla se lo spettro così ottenuto proviene o meno da una distribuzione uniforme.
- b) Test seriale: generalizza il test precedente verificando l'uniformità della distribuzione di coppie di numeri casuali. Esse vengono disposte in una matrice bidimensionale di $(k \times k)$ elementi e si controlla se in tutti gli elementi vi è lo stesso numero di coppie.
- c) Test del poker; per scoprire eventuali correlazioni multidimensionali. In questo caso si considerano gruppi consecutivi di 5 numeri casuali e si determina quanti numeri uguali si trovano all'interno di ogni quintupla, secondo uno schema simile a quello del gioco del poker. La probabilità delle diverse configurazioni viene poi paragonata con quella prevista nel caso di una distribuzione di numeri puramente casuale, che viene facilmente ricavata partendo dalla distribuzione binomiale.

Considerazioni

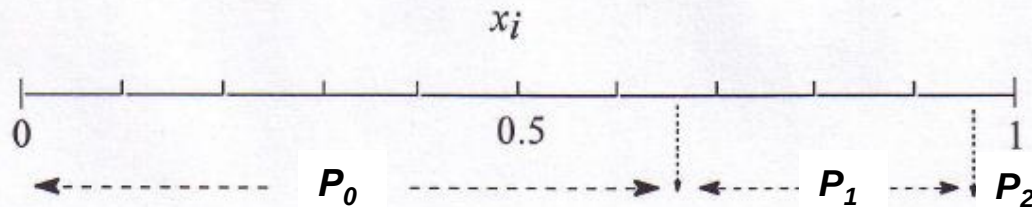
- Le caratteristiche che deve avere un generatore di sequenze pseudocasuali:
 - Ripetibilità, affinché sia possibile ripetere gli stessi esperimenti più volte
 - Soddisfacimento dei test: servono per verificare che il generatore sia quanto più simile ad un generatore di numeri casuali ideale.
 - Semplicità e rapidità: un simulatore che usa il generatore risulta quindi efficiente.
 - Periodicità lunga, in modo da poter effettuare simulazioni lunghe senza riutilizzare più volte la stessa sequenza.
 - Portabilità, cioè rendere l'implementazione del generatore indipendente dalla piattaforma.

ESEMPIO: Generazione di N numeri random x_i

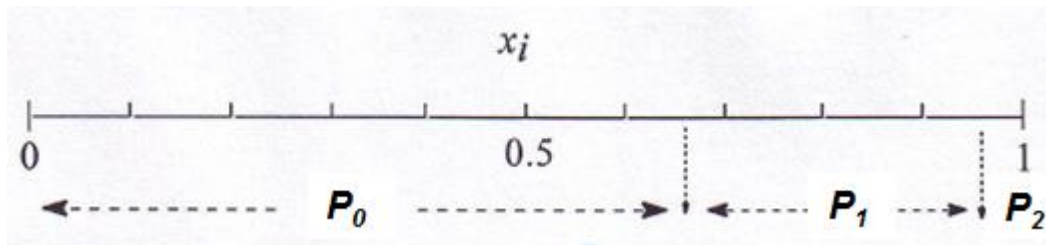
Stati energetici	$E_0=0$	Peso statistico	$P_0=0.66$
	$E_1=0.5$		$P_1=0.28$
	$E_2=1.5$		$P_2=0.06$

- la molecola adotterà la conformazione $E_0 = 0$ se $0 \leq x_i \leq P_0$;
- la molecola adotterà la conformazione $E_1 = 0.5$ se $P_0 \leq x_i \leq P_0 + P_1$;
- la molecola adotterà la conformazione $E_2 = 1.5$ se $P_0 + P_1 \leq x_i \leq P_0 + P_1 + P_2 = 1$

E' chiaro che i numeri random cadranno con maggiore frequenza in quella parte del segmento $[0, 1]$ più ampia.



Per $N=20$



n_i	x_i	conform.
1	.00002	E_0
2	.08503	E_0
3	.60135	E_0
4	.89161	E_1
5	.96796	E_2
6	.18969	E_0
7	.51498	E_0
8	.39801	E_0
9	.26291	E_0
10	.74351	E_1
11	.08955	E_0
12	.56039	E_0
13	.58223	E_0
14	.80957	E_1
15	.59192	E_0
16	.51171	E_0
17	.87663	E_1
18	.99508	E_2
19	.72621	E_1
20	.96661	E_2

$$R_g, \langle r \rangle^{1/2}$$



- **Diffusione della luce**
(Laser Light Scattering)
- **Viscometria**