Molecular Design

Saul Pierotti

May 14, 2019

Cosa studieremo

- Progettazione di molecole, non sintesi
- Come descrivere lo spazio biochimico in termini di strutture e reazioni
- I DBs utilizzati nella progettazione molecolare
- I metodi utilizzati per analizzare dati biochimici, ossia la chemiometria
- La relazione tra proprietà e struttura delle molecole (QSPR)
- Algoritmi sottostanti ai sistemi di previsione delle interazioni

Uso di modelli nelle scienze sperimentali e note introduttive

- I modelli sono impiegati in ogni settore scientifico, ma alcuni campi impiegano modelli hard ed altri soft
- Un esempio di modello hard è il modello ab initio, che permette di ottenere una proprietà del sistema in esame a partire da una sua configurazione
  - Uso modelli diversi per studiare proprietà diverse
- I risultati ottenuti tramite modelli hard sono approssimazioni della realtà
  - Le approssimazioni effettuate possono essere accettabili nella previsione di sistemi semplici (piccole molecole)
  - Questi modelli sono utili come previsione, o se non è possibile effettuare l'esperimento reale
  - Un modello hard usa dati solo calcolati
- Un modello soft è più accurato, ma necessita di poter ideare un esperimento adeguato in quanto necessita di dati sperimentali
- In questo corso utilizzeremo modelli che vanno dal soft al semi-hard
- La chemioinformatica, scienza che applica modelli informatici a sistemi molecolari, è stata fondata da Gasteiger, un chimico organico
  - Pur essendo un chimico organico ha sempre lavorato al pc e non ha mai fatto sintesi in laboratorio
  - Ha fondato la company Molecular Networks
  - Probabilmente verrà a fare una lezione a Perugia a fine aprile (!)
- In questo campo sono molto usati modelli basati sull'intelligenza artificiale
  - L'AI è una tecnica nata con l'informatica stessa
- In chemioinformatica sfruttiamo modelli per predire le interazioni tra molecole, non la loro struttura
- Una cosa si può definire compresa, conosciuta, se ne esiste un modello sufficientemente accurato
- Lo scopo della sintesi non è la produzione di composti, ma di proprietà
- I farmaci subiscono un iter di approvazione di più di 15 anni
- Nel settore farmaceutico guadagni e perdite sono enormi
- Poter tagliare fuori candidati problematici ad uno stadio precoce ha un potenziale enorme

Descrivere sinteticamente una molecola

- Il nome di un composto è più utile che venga definito in base alla sua struttura, non alla sua origine
- Una nomenclatura efficiente migliora la produttività dei chimici

- La nomenclatura IUPAC permette una nomenclatura univoca e descrittiva dei composti chimici

Formato dei dati chimici, biologici e farmaceutici

- I dati sono sia input che output dei programmi di modelling
- Un dato chimico deve contenete la composizione chimica, i legami e la geometria molecolare
- Per piccole molecole il formato preferito è mol2
- Per le proteine ed altre macromolecole si utilizza il pdb
  - Il record è l'atomo, indicato in coordinate xyz
  - E' indicato anche l'aminoacido di appartenenza
  - Sono anche riportate le molecole d'acqua legata
  - La precisione è riportata tramite il B-factor
    - La vibrazione termica causa incertezza nella misura
  - ATOM indica tutti gli atomi che partecipano alla struttura di aminoacidi
  - HETATM indica atomi di solvente, piccole molecole, ecc.
  - Gli H non sono mai presenti perché non visibili ai raggi X, ma possono essere inseriti su modelli virtuali
- I dati sono memorizzati in databases come il PDB e il CCDC (piccole molecole)
- Il Cambridge Cristallographic Data Centre (CCDC) è un DB a pagamento di strutture di piccole molecole
- Presenta molte più strutture di PDB
- Le banche dati pubbliche contengono sui ^{5} composti, quelle private farmaceutiche più di 10^6
- Tutti i formati riportano delle informazioni essenziali, e altre non essenziali
  - Il tipo di atomo
  - Le coordinate atomiche xyz
  - Non è necessario inserire i legami, poiché sono dedotti dalle distanze atomiche
    - In alcuni formati è comunque riportata una matrice di connettività
  - Spesso sono riportate informazioni sulla confidenza della posizione
    - La confidenza è assoluta per strutture calcolate
  - Può essere riportata la densità di carica elettronica dei vari atomi
    - Viene salvata per evitare di ricalcolarla
  - Può essere riportata la geometria molecolare, ossia lo stato di ibridazione dei vari atomi

Notazione SMILES

- E' una rappresentazione che consente di convertire una molecola in una stringa
- E' il formato più usato in banche dati chimiche e farmaceutiche
- Trasformo la struttura in un grafo
  - Rimuovo gli idrogeni
  - Apro gli anelli ponendo un numero ad ogni rottura, che mi permette di identificare gli atomi separati
  - Il cicloesano può essere scritto C1CCCCCC1
  - Se un atomo chiude 2 anelli gli si assegnano 2 numeri consecutivi (es. C1CCCC2CCCCC12)
  - Se voglio indicare più di 9 cicli premetto il simbolo % (l'atomo che chiude il ciclo 12 è C%12)
- Indico i legami in modo standard
  - Scrivo 2 atomi consecutivamente per un legame semplice (es. CC)
  - In alcuni casi è necessario esplicitare il legame con - (es. 2 cicli aromatici collegati tra loro)
  - = per doppi legami
  - # per triplo legame
  - $ per quadruplo legame
  - . per un legame non esistente (es. $[Na^+].[Cl^-]$)
  - : per un legame aromatico con parziale carattere di doppio legame
- I composti aromatici possono essere rappresentati in vari modi
  - Con i doppi legami alternati (Kekulè) C1=CC=CC=C1

  - Con il simbolo (:) C:1:C:C:C:C:C1
  - Scrivendo i costituenti del ciclo in minuscolo c1ccccc1
- Le ramificazioni sono indicate con parentesi (es acido acetico CC(=O)O
- E' possibile indicare stereoisomeri
  - Per l'isomeria cis-trans indico con F/C=C/F (oppure F\C=C\F) l'isomero trans e F/C=C\F il cis (oppure F\C=C/F)
  - Gli stereoisomeri RS si indicano con @ se S e @@ se R (@ è una spirale antioraria!)
  - Il senso è quello rispetto al primo atomo elencato del centro chirale
  - L-Ala si indica N[C@@H](C)C(=O)O
- La codifica non è unica, una molecola può essere rappresentata in modi diversi
  - Questo crea problemi nel mining dei databases e per la ricerca di sottostrutture
- Canonical SMILES è invece univoco
  - Dipende da un algoritmo di canonicalizzazione
    * E' un problema complesso

Ottenere la struttura 3D

- La stringa SMILES viene convertita in rappresentazione 2D, che è poi usata per ottenere una 3D approssimata
- La struttura è migliorata con metodi di minimizzazione energetica o semiempirici
- I metodi semiempirici e di meccanica molecolare sono troppo lenti
  - Si parla di secondi, ma se i composti sono milioni è un problema
- Il software CONCORD riesce a convertire 1D in 3D in tempi ragionevoli
  - Spezza la molecola in frammenti a struttura nota, di cui ha un database interno
  - Ottiene una struttura 3D approssimata, che poi migliora con metodi di minimizzazione energetica
  - Si valuta come l'energia varia al variare della posizione, fino ad arrivare ad un minimo
  - Il minimo di energia di solito corrisponde alla struttura cristallografica

Metodi teorici e sperimentali

- I metodi teorici possono essere di varie tipologie
  - La predizione ab initio usa modelli hard basati esclusivamente su modelli quantistici, ossia risolve l'equazione di Schroedinger
    * Non sempre è computazionalmente possibile
  - I metodi semiempirici richiedono alcuni parametri sperimentali, e risolvono la funzione d'onda solo per gli elettroni di valenza
    * Sono più approssimati, ma più veloci
    * Usano modelli di meccanica classica per gli altri elettroni
  - Metodi di meccanica molecolare ignorano gli effetti quantistici
- I metodi sperimentali permettono di ottenere strutture e conformazioni
  - Uno dei metodi sperimentali più usato è la cristallografia ai raggi X, specie per macromolecole
  - Per piccole molecole si usa più NMR
    * E' usato per ottenere la struttura, più che la conformazione
    * Oggi sono usati spettrometri NMR anche per le proteine, anche se non è sempre applicabile
    * Nella spettrometria NMR non serve il cristallo (!)
  - Per misurare precisamente gli H si usa la cristallografia a diffrazione neutronica

Meccanica molecolare

- I sistemi che studiamo sono complessi, con molti elettroni e nuclei, e non possono essere predetti *ab initio*
- La meccanica molecolare (MM) è il metodo di predizione più veloce delle proprietà molecolari
- E' applicabile allo stato fondamentale ma non a quello eccitato
- Non permette di prevedere la distribuzione elettronica di una molecola

- Permette di prevedere le proprietà cinetiche e termodinamiche e l'energia di una conformazione
- Sfrutta l'approssimazione di Born-Oppenheimer, considera i nuclei fermi nelle transizioni elettroniche
  - Assume quindi la distanza tra nuclei costante
- Considera gli atomi come sfere legati da forze elastiche, con carica netta o parziale
- Descrive le interazioni come potenziali, e determina l'energia di una conformazione in base a questi
- L'insieme dei parametri e delle funzioni potenziali è definito Force-Field (FF)
- Le forze intra- ed inter-molecolari sono definite da 4 contributi nei FF
  - $E = E_{stretching} + E_{bending} + E_{torsion} + E_{non-bonding}$
  - Possono essere considerati anche contributi aggiuntivi
  - L'energia totale non ha significato assoluto, ma le differenze energetiche sono significative
  - Lo scopo di un FF di MM è la predizione della struttura di una molecola
- L'energia di stretching è modellata come elastica (legge di Hooke) attorno ad una lunghezza di equilibrio, tipica del legame
  - $E_{stretching} = \sum k_b(r - r_0)^2$
  - k_b modella la rigidezza del legame
  - E' un'approssimazione, l'energia non ha un andamento parabolico ma di ordine superiore
- L'energia di bending considera la deformazione in modo elastico attorno all'angolo di equilibrio
  - In questo caso la costante è assegnata non ad un legame ma ad una tripletta di atomi
- L'energia torsionale è modellata da una funzione periodica
  - $E_{torsion} = \sum A[1 + cos(n\tau - \phi)]$
  - Il parametro A controlla l'ampiezza della curvatura, n la periodicità
  - La parametrizzazione coinvolge quartetti atomici
- L'energia di non legame considera le forze di London ed elettrostatiche
  - $E_{non-bonding} = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \frac{q_i q_j}{r_{ij}}$
  - E' il calcolo computazionalmente più impegnativo
  - La curva presenta una distanza optimum a cui l'energia è minima
  - A distanze inferiori si ha interazione repulsiva, a distanze superiori attrattiva
  - Il parametro A indica la polarizzabilità dell'atomo, B la durezza del guscio atomico
  - B viene determinato per cristallografia
  - Le cariche della componente elettrostatica sono pre-assegnate o calcolate
- Altri possibili contributi all'energia totale sono i legami idrogeno e l'effetto del solvente
- Nei FF comunemente usati sono necessari da 2000 a 50000 parametri misurati, per modellare 60 tipi di atomi unici

Cristallografia

- Per la cristallografia è importante avere un campione proteico puro
  - Un cristallo è un array periodico di molecole
  - I cristalli proteici contengono canali e buchi pieni di solvente
  - E' molto difficile ottenere il cristallo
  - Una volta era un processo artigianale, ora è automatizzato
    - Si modulano sali, acqua, metalli pesanti, tensioattivi
- Il cristallo viene colpito da raggi X, e l'analisi del pattern di diffrazione prodotto permette di generare una mappa di densità elettronica
  - Dal reticolo di diffrazione non ottengo la posizione ma la densità elettronica, non discerno bene atomi da gruppi di atomi
  - Fare il cristallo serve ad amplificare il segnale (!)
    - Più del 99% del raggio incidente non viene deviato
  - Nella zona centrale dell'immagine ho il fascio diretto, mentre attorno il pattern di diffrazione
- La sequenza della proteina viene adattata alla mappa di densità elettronica ruotando attorno ai vari legami, in un processo detto fitting
  - Più la mappa ha risoluzione elevata, meno ambiguità conformazionali vi sono
  - Si considera alta risoluzione 1.5Å, bassa risoluzione 5Å
    - Per poterci lavorare bene deve essere almeno 2.5Å

- Oggi la cristallografia si fa in pochi centri specializzati
  - In Europa si fa a Grenoble dove c'è un sincrotrone (European Synchrotron Radiation Facility, ESRF)
  - Contiene camere di analisi poste tangenzialmente all'anello del sincrotrone
- Oggi si sta iniziando ad usare la diffrazione a raggi X per medical imaging e microchirurgia
- Si conoscono più di 10^5 proteine, ma solo 10^4 strutture cristallografiche
- Vi sono grandi investimenti sullo studio delle proteine
- L'utlizzo di enzimi permette di effettuare reazioni chimiche estremamente selettive
- I detersivi per lavatrice hanno una grossa componente enzimatica
  - Le proteine sono stabilizzate con ponti SS e altri legami per farle resistere nelle condizioni di utilizzo

Molecular Interaction Fileds (MIF)

- Il target dei MIF può essere qualsiasi molecola a struttura 3D nota
- Si costruisce un reticolato di punti che circonda completamente il target
- Si posiziona un probe in ogni punto del grid, valutandone le interazioni col target
- I legami del target sono liberi di ruotare attorno agli assi consentiti, accomodando la presenza del probe