Molecular Design

Saul Pierotti

May 13, 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, oppure HETATM se non appartiene a nessuno
  - Sono anche riportate le molecole d'acqua legata
- I dati sono memorizzati in databases come il PDB e il CCDC (piccole molecole)
- Le banche dati pubbliche contengono sui ^{5} composti, quelle private farmaceutiche più di 10^6

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
- 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