g% Appunti Molecular Design % Saul Pierotti % May 9, 2019

Lex.1

Lex.2

- Le simulazioni di docking hanno una precisione sull'ordine degli Angstrom

Dati chimici, farmaceutici e biologici

- I dati di struttura possono essere archiviati in vari formati
  - pdb è tipico di macromolecole
  - mol2 è usato per piccole molecole
- 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, poichè 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 e non sperimentali
  - Può essere riportata la densità di carica elettronica dei vari atomi
    - Viene salvata per evitare di ricalcolarla
  - Può essere riportata la geometria molecolare, osssia lo stato di ibridazione

Formato PDB

- Riporta il tipo di atomo, l'aminoacido, le coordinate
- 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 perchè non visibili ai raggi X, ma possono essere inseriti su modelli virtuali

Metodi sperimentali

- Tramite cristallografia ai raggi X, specie per macromolecole
- Per piccole molecole si usa più NMR
  - Piu per ottenere la struttura, che la conformazione
  - Oggi sono usati spettrometri NMR anche per le proteine, anche se non è sempre applicabile
  - Non serve il cristallo (!)
- Per misurare precisamente gli H si usa la cristallogrfia a diffrazione neutronica

Metodi teorici

- L'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 risolve 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 quantisitici
- Inizio inserendo una struttura 2D al PC
  - Utilizzando il metodo prescelto, viene calcolata la struttura 3D

Cambridge Cristallographic Data Centre

- E' un DB a pagamento di strutture di piccole molecole
- Presenta molte più strutture di PDB

Uso delle SMILES

- Posso usarle per effettuare una ricerca in DB, di solito per sottstrutture o gruppi funionali
  - Questo permette di ricercare molecole con funzione biologica simile
  - La SMILES che uso è detta query

Calcolo delle strutture 3D

- Impiega metodi di minimizzazione energetica per calcolare la conformazione a partire dalla struttura 2D
  - Minimizzano l'energia interna della molecola
  - Il minimo di energia di solito corrisponde alla struttura cristallografica
- Si valuta come l'energia varia al variare della posizione, fino ad arrivare ad un minimo
- I metodi semiempirici e di meccanica molecolare sono troppo lenti
  - Si parla di secondi, ma se i composti sono milioni è un problema
  - E' stato inventato il metodo CONCORDE, che ha ridotto drasticament i tempi
  - Con la potenza di calcolo attuale non è più un problema

Lex.3

- La pKa è essenziale per determinare la permeabilità di molecole a livello delle membrane biologiche
- I farmaci tendono ad essere carichi a pH fisiologico
- Per convenzione si usa sempre la pKa, anche per le basi

Cosa determina la pKa

- L'elemento a cui H+ è legato
  - Più è elettronegativo meno è acido (?)
  - Più è grande meno è aacido poichè distribuisce la carica (?)
- Effetto induttivo
  - Altri atomi elettronegativi presenti nella molecola influenzano l'acidità
  - Mediato da legami σ
- Effetto di risonanza
  - Mediato da legami π
- Ibridizzazione
  - Più aaumenta il carattere s più l'acido è forte
- Effetto prossimtà
  - Vicinanza di un gruppo che forma legami H intramolecolari

Determinazione sperimentale della pKa

- Elettroforesi capillare, titolazione spettrofotometrica, titolazione potenziometrica
- Una pKa sperimentale è sempre preferibile ad una calcolata
- Non sempre è possibile la determinazione sperimentale per motivi tecnici o economici, e quando ho molte molecole

Metodi per determinare la pKa

- I metodi ab initio sono troppo poco precisi e richiedono molto tempo
- I metodi sperimentali richiedono di avere a disposizione un campione
- La QSAR è la tecnica più usata

A Perugia è stato sviluppato MoKa

- Molto veloce, accurato, considera le molecole poliprotiche
- Indipendente dalla rappresentazione esplicita degli H
- Calcola il LogP (partizione EtOH-H2O) e LogD (LogP in funzione del pH)
- Può essere allenato con un modello
- Si basa sui campi GRID, dove un probe si muove e valuta le interazioni con la molecola
  - Si stanno ora sviluppando probes poliatomici
- Si creano dei livelli allontanandosi dall'atomo di riferimento
  - Il livello 0 è l'atomo stesso, e si pongono tutti i bit a 0 eccetto quello per N
  - La rappresentazione è non più binaria
  - Si crea un fingerprint concatenaando questi bit
- Si crea una tabella che correla il fingerprint con la pKa, usata come modello di training
  - Usa il metodo PLS
- Bisogna considerare i tuatomeri, poichè ognuno ha una sua pKa (!)

Lex.4

Proteine

- 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
- Sono note almneno 500000 sequenze proteinche
- Le strutture proteiche 3D note sono molte meno

Cristallografia

- E' importante avere un campione proteico puro
- E' difficile ottenere il cristallo
- Una volta era un processo artigianale, ora è automatizzato
  - Si usano sali, acqua, metalli pesanti, tensioattivi
- 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
- Dal reticolo di diffrazione non ottengo la posizione ma la densità elettronica, non discerno bene atomi da gruppi di atomi
- Una volta bisognava fare fitting della sequenza sulla densità elettronica
  - Ora si fa via software
- La strettezza della mappa di densità mi riflette la risoluzione della struttura che posso ottenere
- Si considera alta risoluzione 1.5Å, bassa risoluzione 5Å
  - Per poterci lavorare bene deve essere almeno 2.5Å
  - 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

Trovare le binding pockets

- Il primo step è trovare il sito o i siti di modulazione/legame
- Un singolo composto potrebbe interagire con più di una tasca nella stessa proteina
- E' possibile descrivere un pocketoma che raccolga le tasche note, e predica le interazioni di una molecola
  - E' rappresentato come network che misura la distanza di fitting di 2 tasche
- Si fanno simulazioni di fitting provando le varie tasche disponibili

Lex.5

Molecular interction fields (MIFs)

- Il prof ha scritto un libro che ci darà in pdf
- Goodford ha fondato la Wellcome Trust
- Il software **GRID** è gratis per l'accademia e a pagamento per i *for profit*
- Il target dei MIFs è l'insieme dell'interazione, non i singoli componenti
- La zona di interesse può essere l'intera molecola o una particolare porzione dello spazio
  - La zona di interesse è definita con una griglia
- In tale regione metto un **probe** chimico, con cui la scansiono muovendolo per righe, colonne e piani
- La sonda chimica da utilizzare può essere scelta in base alle necessità
  - Posso usare ioni, piccole molecole, ecc...
  - Posso anche usare parti fittizie di molecole, ad esempio un gruppo OH non legato a nulla
  - La scelta della sonda dipende dal tipo di interazioni con la proteina che voglio studiare
- In ogni punto calcolo l'interazione del probe con la proteina
- Per simulare l'interazione calcolo la sommatoria dell'interazione del probe
  - Valuto la Van der Waals, l'interazione di carica, i legami idrogeno ed il contributo entalpico
  - Una risultante negativa indica attrazione, una positiva repulsione del probe
  - L'interazione è valutata con tutta la proteina, anche nelle zone al di fuori dalla regione di interesse (!)
- La proteina ed il probe si muovono per minimizzare l'energia libera
  - Oltre a legami diretti col probe, si valutano anche tutti i legami indotti all'interno della proteina stessa
  - Questo può alterare la conformazione in zone distanti della proteina (!)
  - Spesso una molecola d'acqua può fare da ponte per trasmettere un legame H

Contributo di Van der Waals (Lendar-Jones (?))

- E' dovuto alla fluttuazione degli elettroni di valenza
- L'energia dell'interazione è 0 per distanze infinite, diviene negativa all'avvicinarsi dei nuclei, incontra un minimo e poi aumenta verso infinito per distanza 0
  - A è il termine attrattivo e dipende da r^-6, B è il termine repulsivo e dipende da r^-12

Contibuto Coulombiano

- L'interazione Coulombiana dipende da inversamente da r^2, direttamente dal prodotto delle cariche e dalla costante dielettrica del mezzo
- La costante dielettrica è un termine preponderante, in H2O l'interazione è 80 volte inferiore che nel vuoto
- In biologia si usa un'equazione modificata e più complessa
  - Modella l'azione dell'acqua nelle proteine, in cui la costante dielettrica non è più costante (!)

Contributo del legame H

- Il legame H è elettrostatico ma non è descritto adeguatamente dall'interazione Coulombiana
  - La direzionalità modifica l'interazione al punto da doverla descrivere separatamente
  - Dipende dalla geometria delle molecole interagenti e dei loro orbitali
- Ha una componente simile alla Lendar Jones (?), un contributo Coulombiano ed una componente geometrica
- Screenando DBs per la posizione di molecole d'acqua per un particolare gruppo chimico in diversi contesti trovo gli angoli di interazione più favorevoli
- Il carbonile forma legami H sopratutto in direzione dei lone pairs, ma un po' anche tra di essi
  - Tra i lone pairs è possibile per H2O formare 2 legami con entrambi, ma la sovrapposizione orbitalica è inferiore ## Progettare una molecola

- Tramite vari probes vedo dove certi gruppi sono favoriti

Lex.6

- Le interazioni deboli sono da 0 a 10 kcal/mol
  - La Leidar-Jones è di circa 1 kcal/mol
  - Il legame H è di circa 4 kcal/mol
  - Le interazioni ioniche possono raggiungere i 15 kcal/mol
  - La componente entropica è sulle 1.5 kcal/mol

Componente entropica

- Posso utilizare un probe idrofobico, che spiazza delle molecole d'acqua di solvatazione da una superficie apolare
  - Queste molecole d'acqua subiscono un guadagno energetico di circa 0.9 kcal/mol dovuto alle maggiori interazioni che possono formare quando non solvatano la superficie idrofobica
  - Si ha un ulteriore guadagno per interazioni di Leidar-Jones di circa 1 kcal/mol tra la superficie ed il probe
- Consideriamo ora lo stesso probe idrofobico su una superficie polare
  - Si hanno le stesse componenti di prima che determinano -1.9 kcal/mol, ossia guadagno entropico del solvente e interazioni Leidar-Jones
  - Si ha una perdita di energia dovuta alla rottura dell'interazione tra il gruppo polare e le molecole di acqua, di circa +2.5 kcal/mol
  - Il processo non è spontaneo in quanto ha un'energia di circa +0.6 kcal/mol
- Spesso è la parte più importante per il numero di interzioni che si formano
  - Un farmaco idrofobico è spesso più potente di uno idrofilico perchè la sua interazione col target è favorita
  - E' molto più tollerante ad imprecisioni nella previsione del legame, poichè è poco direzionale
- In biologia la selettività è data dalle interazioni polari, la potenza dell'interazione da quelle idrofobiche

Piccole moelcole

- Interrogando la zona con +0.2 kcal/mol posso studiare la superficie della molecola
  - Questo mi permette di calcolare il volume e l'ingombro sterico della molecola
- L'arginina è spesso usato come amminoacido idrofobico (!)
- Posso usare probes anfifilici per evidenziare dove avviene la transizione idrofobico-idrofilico
- Studiando con probe idrofobico, H2O e anfifilico il colesterolo posso predire come questo si posiziona sulle membrane
- Le strutture delocalizzate sono molto apolari poichè polarizzabili

Lex.7

- Possiamo attuare le tecniche precedenti anche con piccole moelcole, ad esempio lipidi
- In una fosfatidilcolina sorprendentemmente l'N+ quaternario non è idrofilico (!)
  - Osserviamo una porzione idrofobica sulla coda alifatica, una idrofilica a livello del fosfato e una intermedia a livello dell'azoto
- I diacilgliceroli in acqua si dispongono con le code adiacenti per minimizzare la superficie idrofobica esposta
  - L'idrofobicità della molecola con le code appaiate è inferiore a quella della stessa con code separate (!)
- In un trigliceride similmente le code sono disposte aggrovigliate tra loro
- Minimizzare una struttra significa trovare la sua conformazione di minimo energietico

Uso dei MIF per produrre un farmaco antiinfluenzale

- In 7 anni, senza il cristallo della neuraminidasi, si è riuscito a sviluppare un farmaco capace di inibirla
  - Questo composto aveva poca affinità, con K di circa 1 umol
- A seguito della pubblicazione del cristallo del farmaco nella proteina si è cercato di aumentarne la potenza con poco successo
- Il prof Cruciani a d Oxford ha usato i MIF per migliorarlo
  - Ha visto che usando un paarticolare probe vi era una zona ad alta affinità nella tasca di legame
  - Si è quindi inserito tale gruppo in posizione consona, producendo lo Zanamivir
- La GSK ha comprato la ditta che lo produceva, rinominando il farmaco Relenza
- La Gilead ha copiato il farmaco aggiungendo un estere di un gruppo carbossilico, creando il Tamiflu
  - L'estere viene idrolizzato a livello gastrico riproducendo il gruppo carbossilico
  - Lo ha fatto per evadere il brevetto e per aumentarne la permeabilità di membrana
  - Ha anche sostituito una porzione idrofilica con una idrofobica, usando una previsione fatta con GRID
- La Roche ha comprato il Tamiflu a 100 milioni di dollari, ottenendoci guadagni enormi
- Il Tamiflu fino a qualche hanno fa aveva il 90% di share mentre il Relenza il 10%
- Adesso si sta sviluppando resistenza al Tamiflu, mentre non vi è ancora resistenza a Relenza
  - Questo perchè il Tamiflu è stato abusato mentre Relenza no

Flexible MIFs

- E' un MIF che tiene conto del movimento della proteina a seguito dell'interazione col probe
- Il risultato ottenuto è molto più accurato di un MIF statico se sto studiando una proteina con conformazione flessibile
- E' importante notare che le possibili interazioni predette sono in molti casi mutualmente esclusive (!)
- Questi campi sono usati per studiare come una certa regione possa alterare le proprie proprietà in virtù della conformazione

Preogettare lo scaffold

- GRID mi dice dove devo mettere le mie decorazioni
- Il chimico si preoccupa di creare uno scaffold per posizionare tali gruppi in modo appropriato
- Oggi esistono anche software in grado di suggerire lo scaffold appropriato
- Nota che tutti questi sitemi progettano ligandi, non farmaci, non è detto che le molecole che trovo abbiano attività biologica (!)

Lex.8

- Maybe missing (?)

Lex.9

- Non vi è correlazione tra momento idrofilico e forza delle interazioni idrofiliche (integy moment)
- E' importante disegnare descrittori non correlati tra loro, altrimenti descrivo 2 volte la stessa cosa (!)
- Se l'idrofilicità è diffusa anzichè localizzata è più facile che il composto passi la BBB
- Critical packing
  - E' un parametro usato da chi studia tensioattivi e micelle, inventato da un russo
  - E' un fattore geometrico che indica il fattore critico di impacchettamento
  - Alla concentrazione micellare critica una molecola anfifilica forma micelle anzichè stare in soluzione

  - E' dato da un equazione che considera lunghezza lipofila della molecola, volume lipofilo e superficie idrofilica
- Equazione di Stokes-Einstein
  - Predice la diffusività in acqua usando solo parametri fisici
- Un modello del prof di machine learning usa i vari descrittori per predire la diffusività
- La diffusività è poi trattata come un altro descrittore
- E' in accordo con i dati sperimentali
- Elongation è un altro descrittore
  - E' la lunghezza più probabile della molecola
  - In una molecola flessibile è diversa dalla lunghezza della molecola disegnata (!)
  - Può essere calcolata con un'equazione
- LogP è il parametro più misurato sperimentalmente che si conosca
  - E' il logaritmo di una partizione
  - E' la partizione acqua/n-ottanolo
  - Una volta si usava olio di oliva ma poi si è standardizzato con n-ottanolo
  - n ottanolo è poco miscibile in acqua
  - $P = [n - ott]/[H_2O]$
  - Se LogP è 0 significa che P=1 e quindi la molecola ha la stessa solubilità in acqua e n-ottanolo
  - Se è 1 è 10 volte più solubile in n-ottanolo, se -1 viceversa
  - Approssima il comportamento sulle membrane
- Calcolare il LogP
  - Riesco a correlare bene con la conformazione, ma vi sono outliers
  - Gli outliers sono molecole molto flessibili che non hanno una conformazione ben definita
  - Quale conformazione scelgo per la previsione?
- pH fluidi biologici 7.4
- Posso dire al software di modificare la molecola per adattarla al pH di lavoro
  - Modifica lo stato di protonazione in base al pH
  - Usa l'algoritmo di MoKa
  - Il LogP varia molto con il pH (!)

Lex.10

- Metabolismo
- Farmacocinetica
- Organoidi usati per studiare metabolismo di xenobiotici

Lex.11

- Profarmaci
- Il Fluoro è isosterico all'idrogeno
- Previsione dei siti di metabolismo con MetaSite
- Il fluoro è spesso utilizzato per gestire il metabolismo dei farmaci
  - Il legame F-C è molto forte e difficile da rompere
- Devo valutare l'ingombro sterico per capire se la molecola entra nella tasca enzimatica
- Devo capire quale parte della molecola sarà vicina al sito attivo
- Posso giocare sulla termodinamica della reazione e sulla cinetica
- Diversi CYP hanno tasche di dimensioni e caratteristiche diverse

Lex.12

- Simulazione 3d di attività di un cyp
- La piridina si ossida su N a dare un N-ossido, molto polare

- La probabilità di un sito di metabolismo è data dal prodotto tra l'esposizione al sito attivo e la reattività della posizione
- Il radicale benzilico è un toluene senza un elettrone
  - E' uno dei radicali più stabili, e quindi uno di quelli che si formano più facilmente
  - Per questo una volta si pensava fosse sempre uno dei siti più reattivi per le ossidazioni CYP, che sono radicaliche
  - In realtà comunque dipende molto dall'esposizione del sito
- Se voglio una prova del metabolismo posso fare una LC-MS/MS prima e dopo metabolismo con microsomi
  - La differenza di massa mi da un'idea di che modifiche sono avvenute
  - Non sempre è univoco

Lex.13

- Intelligenza artificiale
- Alan Turing ha dato le basi per lo sviluppo dell'AI
- ENIAC, uno dei primi computer
- Kasparov e Deep Blue
- Perchè oggi vi è l'esplosione di AI?
  - Disponibili grandi quantità di dati
- Gestire l'AI

Lex.14

- Primo software AI sviluppato a Stanford per risolvere gli spettri MS
- All'aumentare della massa il numero di molecole compatibili con un certo spettro MS aumenta esponenzialmente
- Fecero quind spettri MS/MS, che riducevano le ambiguità
- Svilupparono un software capace di riconoscere gruppi chimici dagli spettri, in questo modo affinando i risultati a poche o 1 molecola
  - E' considerato il primo knowledge-based system
  - E' basato sulla conoscenza di esperti nel campo, che hanno scritto il programma
- A Perugia si è riapplicato lo stesso concetto per identificare lipidi in studi di lipidomica
- Per determinare il passaggio tra 2 stati, non è necessario descrivere gli stati stessi
  - E' quello che viene fatto calcolando la strada per un posto
- I sistemi knowledge based sono applicabili solo a sistemi noti, non generano soluzioni nuove
  - Sono molto usati nella ricerca universitaria, meno in quella di nicchia ed applicata di frontiera
- Machine learning è il sistema più usato in AI
  - Estrae un pattern da dati raw
  - La sua efficacia dipende molto dal sistema di coordinate usato
  - Posso applicarlo per ottimizzare la resa di una reazione
    - Devo descrivere i reagenti e le condizioni

Vacanze di pasqua

Lex.15

- La parte difficile di AI è estrarre knowledge dai dati
- Pattern recognition
- I computer di solito lavorano con matrici
- Lavoriamo solo con matrici 2d perchè è possibile ricondurre un a matrice n-dim a 2d
- Analisi di immagine

    - Un'immagine può essere ricondotta ad una matrice
    - L'esperimeto è il pixel, le variabili sono la luminanza dei vari canal Analisi di immagine
        * Un'immagine può essere ricondotta ad una matrice
        * L'esperimeto è il pixel, le variabili sono la luminanza dei vari canali
- MALDI/TOF su tessuti??
- Per ricondurre una matrice a 2d si fa unfolding
    - accodo tutti i dati su una sola dimensione, ottengo un vettore
- AI lavora meglio con dati hard (quantitativi)
    - Quando possibile è meglio convertere dati soft in hard
- I dati continui sono più trattabili matematicamente di quelli discreti
    - Sono derivatizzabili
- Least squares
- Per semplificare conviene linearizzare le relazioni tra dati
- Un modello non linear è la superificie di risposta
- Il modello più semplice è più probabile, overfitting
- Le reti neurali fanno un po' overfitting
- Linear discriminant analisys
    - Prendo una retta casuale e vi proietto tutti i punti, e valuto quanto efficacemente li clusterizza
    - Prendo un'altra retta e faccio la stessa cosa
    - La retta di separazione è perpendicolare a quella di proiezione
    - Se lavoro con uno spazio 3d proietto su dei piani

Lex.16

- La scelta dei descrittori influenza la qualità della discriminazione
- LDA funziona bene con poche classi, fino a 5 circa
- LA PCA è un metodo unbiased che non richiede la conoscenza pregressa della presenza di classi
    - E' un metodo di riduzione della dimensionalità
    - La prima componente è la direzione spaziala che discrimina il maggior numero possibile di punti
    - E' la combinazione lineare delle varie dimensioni
    - Score plot con oggetti ripetto alle componenti principali
    - Loading plot riporta il cos delle componenti principali rispetto alle variabili originali

Lex.17

- slides su chemiome.chm.unipg.it/MolDes19/
- In PCA posso fare autoscaling per rapportare equamente le dimensioni delle variabili
    - Riporta tutte le variabili nel range 0-1 moltiplicando per una costante
- La sterling a corciano produce steroidi
- Projection to latent strcutures (PLS)
    - E' un metodo supervised
    - Cerca di relazionare una matrice di variabili con una matrice di risposte
    - Una volta si faceva multiple regression analysis
        * Funzionava solo con poche x
        * Non permette buchi nella matrice
        * Non è stabile con x correlate tra loro
    - Trovo PC1 nel mondo x e nel mondo y
    - Creo un plot con le 2 PC1 x e y
    - Posso fare la stessa cosa con PC2 ecc, ma di solito perde correlazione
    - In realtà modifica le PC per massimizzarne la relazione
- Per evitare overfitting faccio cross validation, all'aumentare delle variabili l'errore prima diminuisce e poi aumenta perché sto modellando il rumore

Lex.18

Lex.19

- In molti casi le proprietà chimiche di una molecola non sono sufficienti per prevedere le interazione con un suo recettore