Virtual Screening and Docking

- Screening of chemical libraries gets more expensive and the hit rates decrease, as the libraries get larger
  - This problem can be avoided by not screening the whole libraries, but only a small subset
  - Compounds that should be included in the screening, is determined by virtual screening (VS)
Virtual Screening and Docking

- Virtual screening is a method that selects the most promising compounds from an electronic database.
- Selection is carried out by searching databases for molecules fitting user-defined constraints such as:
  - Similarity to a set of known actives
  - Pharmacophore
  - Three-dimensional structure of the macromolecular target
This Lecture

- This lecture will focus on a description of a protein-based VS
- The VS is build on four essential steps
  - Preparation
  - Docking
  - Scoring
  - Postfiltering
Preparation of the Compound Library

- Commercially available screening collection's
  - Physically available compounds
    - Enamine
      - 857000
    - Akos screening samples
      - size: 700000
  - Web databases
    - ChemNavigator
      - 24000000
    - ChemMine
      - 5800000
    - ChemDB
      - 4100000
Preparation of the Compound Library

- First step: removing unsuitable molecules. There are generally two ways of doing this
  - The first method is to use a series of different filters, each one excluding compounds with certain properties
  - Second method is build on a knowledge based on binary classification system
Series of different filters

- First types of filters, remove highly reactive and toxic compounds, according to reactive moieties
- Lipinski Rule-of-Five remove all compounds that fulfil two or more of following conditions
  - Poor absorption or permeation are more likely when the molecular weight is over 500 Da
  - Octanol-water partition coefficient clogP is higher than 5
  - More than 10 hydrogen bond acceptor's
  - More than 5 hydrogen bonds donor's
Series of different filters

- Other types of more elaborate filters
  - Absorption-distribution-metabolism-excretion (ADME)
  - Aqueous solubility
  - Membrane permeation
  - Metabolic clearance
Knowledge-based binary classification system

- Knowledge based on binary classification system as neural networks, genetic algorithms (GAs), decision trees distinguishes between appropriate and inappropriate compounds
  - E.g. investigation of drug compounds, it is desirable to eliminate promiscuous binders compounds as they can perturb the assay or detection method
Automated work flow

- Using the different rules a automated work flows can be set up to perform the preparation of the compound library
  - Reading raw data from various suppliers
  - Cleaning data: remove erroneous and complex structures, counter ions
  - Filtering according to predefined property rules (Compound)
  - Detecting duplicates
  - Converting 1D to 3D format with potential enumeration of several conformers
  - Ionizing physiological pH
  - Computing physicochemical and topological descriptors
  - Storing in relational databases for future browsing purpose
Representation of Proteins and Ligands, Docking

- Conformational space accessible to a macromolecule is difficult to reproducing
- Approximation to reproduce the conformational space
  - Rigid body docking: both protein and ligand
  - Semiflexible docking: ligand is considered flexible
  - Fully flexible docking: both protein and ligand
Protein Flexibility

- Protein are highly flexible molecules
  - Small scale fast motions: side chain movements, binding sites
  - Large scale slow domain motions: Hinge bending movement allow the induced fit of the ligand
  - Conformational changes of the protein upon ligand binding can thus extend from only few side chain movements to large hinge bending movements
- Protein should be treated flexible in ligand docking
  - Computationally expensively
Protein Flexibility

- Approach to handle protein flexibility
  - Exploring conformational space of protein side chains using a rotamer library
  - Ensemble of several possible conformations derived experimentally, X-ray, NMR or molecular simulations
  - The conformational ensemble is then translated into a composite grid and used to calculate for any probe atom the potential stored in the composite grid
Ligand Flexibility

- Drugs are smaller than macromolecules, ligand flexibility is computationally easier
  - One approach is to store multiple conformations of the ligands in a database, each conformation being regarded as rigid during the docking process
  - Another approach is to store only one conformation per ligand, but then treat the ligands as flexible
    - Another possibility is to encode flexibility around the dihedral angles (AutoDock)
  - The incremental construction method, divides the ligand into fragments and builds it in the receptor binding site (FlexX and Dock)
Docking Algorithms

- Different docking tools differ in aspects such as
  - Description of molecular interactions
  - Algorithms used to generate ligand structures
  - Average run time per molecule
- Algorithms is divided into deterministic and stochastic approaches
  - Deterministic algorithms are reproducible
  - Stochastic algorithms include a random factor and are not fully reproducible
Incremental Construction Methods (Dock)

- Ligand is divided into single fragments and incrementally reconstructed inside the active site.
  - First step is the identification of points in the active site where ligand atoms may be located *sphere centers*, identified by generating a set of overlapping spheres that fill the site.
  - An anchor fragment is then selected from the rigid fragments. This anchor is orientated within the active site, independently of the rest of the ligand.
  - All possible anchor placements are scored in terms of their interactions with the protein, and the best ones are used for the ligand.
Incremental Construction Methods (FlexX)

- Treats the ligands as flexible and the protein as rigid. Divides the ligands along its rotational bonds into rigid fragments as Dock.

- Differs from Dock in the method used for determining the placement of the anchor fragment:
  - FlexX defines interaction sites for each possible interacting group of the active site and the ligand:
  - The anchor fragment is oriented by searching for placements where three interactions between the protein and the ligand can occur.
Genetic Algorithms (GA)

- GA mimics the process of evolution by manipulating a collection of data structures called *chromosomes*
  - Each chromosome encodes a possible protein-ligand complex conformation
  - A chromosome is assigned a score on the basis of the protein-ligand interactions
- Starting point is a randomly generated *parent* population, the GA repeatedly applies two genetic operators *crossover* and *mutation* resulting in *children*
  - *Crossover* requires two *parents* and produces two *children*
  - *Mutation* introduces random changes
Genetic Algorithms (GA)

- Parent chromosomes are randomly selected to introducing an evolutionary pressure into the algorithm
- This evolutionary pessure ensures that, over time, the population should move toward an optimal solution
- Main features of a genetic algorithm

1) A set operators (crossing over, mutation, etc.) is chosen. Each operator is assigned a weight
2) An initial population is randomly created and the fitness of all individuals computed
3) An operator is chosen using a roulette wheel selection, based on operator weights
4) The parents able to reproduce are chosen, using roulette wheel selection based on fitness scores
5) Genetic operator is applied, children chromosomes generated, and the fitness of each child determined
6) If not already present in the population, children replace the least fit individuals
7) Goto three unless the maximum number of genetic operations is reached
Tabu Search (TS)

- A TS is characterized by imposing restrictions to enable a search process to negotiate difficult regions (tabu list)
- At start a solution is initialized by randomizing the position of the ligand within a certain box around the active site
  - From this solution, a user defined number of moves is generated by a mutation procedure and a ranked according to a scoring function
  - The highest ranked is always accepted as the new current solution, and the previous solution is added to the tabu list
Tabu Search (TS)

Main features of a tabu search

1) Create an initial solution at random. Make it the current solution.
2) Evaluate the current solution. If best so far, record it as best solution.
3) Update the tabu list
   a) If Tabu list not full, add current solution to the list
   b) Else, replace the oldest member with the current solution
4) Generate and evaluate x possible moves e.g. 1000 from the current solution
5) Rank x moves in ascending order of interaction energy
6) Examine the moves in rank order
   a) If move has the lowest energy than the best solution so far, accept it and goto seven
   b) If move is not tabu, accept is and goto seven
   c) If no acceptable move, exit
7) If the maximum number of iterations is reached, exit with the best solution found
   a) If best solution has not changed for a number of iterations e.g 100 goto one
   b) Else, goto two
Simulated Annealing and Monte Carlo Simulations

- Simulated annealing is a molecular simulation, in which the system is cooled down at regular time intervals
  - During each constant temperature cycle, random changes are made to the ligand's current orientation and conformation
  - The new state is immediately accepted if its energy is lower than the preceding state
  - The configuration can also be accepted based on a Boltzmann equation
- Disadvantages: depends on the initial placement of the ligand and the algorithm does not explore the solution space exhaustively
Shape fitting Methods

- Shape fitting methods estimated the steric and electrostatic interaction between the ligand conformations and the protein
- Uses a grid based representation of both ligand and protein with assignment of defined values
- Fast Fourier transforms are used to optimize ligand-protein complex
Scoring Functions

- Free energy Gibbs Helmholtz equation
- Relation to the binding constant $K_i$
- Different techniques available for predicting the binding free energy, differ in accuracy and speed
  - Free energy perturbation, vary accurate but time consuming
  - Scoring functions, much faster, but less accurate
- Scoring functions are used for two purposes
  - During the docking process, they serve as fitness function in the optimization placement of the ligand
  - Ranking of each solution in the database
Scoring Functions

- Scoring functions can be mainly categorized into three groups
  - Empirical Scoring Functions
    - FlexX, Hint, Ligscore
  - Force-field based Scoring Functions
    - AutoDock, Dock, Goldscore
  - Knowledge based Scoring Functions
    - Bleep, Pmf, SmoG
Empirical Scoring Functions

- Use several terms describing properties known to be important in drug binding
  - Construct a master equation for predicting binding affinity
  - Multilinear regression is used to optimize the coefficients to weight the computed terms using a training set
  - Terms describe polar interactions such as hydrogen bonds and ionic interactions apolar interactions
- Breakthrough!!! Boehm's function

- Disadvantage is the need of a training set to derive the weight factors
Force-field Scoring Function

- Based on nonbonded terms of classical molecular mechanics force field
  - Lennard-Jones potential describes van de Waals interactions
  - Coulomb energy describes the electrostatic components of the interactions
- Disadvantage of forcefield calculations is the exclusion of the entropic component of the binding free energy
  - Attention not to overestimate the larger and most polar molecules which get the highest enthalpy interaction scores
Knowledge scoring function

- Use a potential of mean force which encodes structural information gathered from protein ligand X-ray coordinates into Helmholtz free interaction energies of protein ligand atom pairs.

- Each interaction type between a protein atom of type $i$ and a ligand atom type $j$ at a distance $r_{ij}$ is then assigned a protein ligand interaction free energy $A(r)$.
Postfiltering

- Decrease the number of virtual hits
  - Docking 50000 compounds will generate hit list of 1-2000 molecules
- Scoring functions are far from able to predict binding free energies
  - Postdocking strategies to select virtual hits for experimental validation
  - Detect false positives and enhance true positive rates
Filtering by Topological Properties

- To eliminate protein ligand complexes with improper geometries
  - Different filters can be applied for evaluation
  - Three dimensional protein ligand complexes in terms of their steric complementarity
- Stahl et al. Developed a set of filters
  - Fraction of ligand volume buried inside the binding pocket
  - Solvent accessible surface of nonpolar parts
  - Close contacts between nonhydrogen bonded polar atoms
- Optimal, the buried ligand volume should be large and solvent accessible and close contacts as small as possible
Filtering by Consensus Mining Approaches

- An investigation of various conformations of the target for docking
  - Contribute to better handling of the binding site flexibility
- Consensus approaches usually require a substantial amount of knowledge (X-ray structure)
Filtering by Combining Computational Procedures

- Incorrect treatment of long range electrostatic effects and desolvation are clear drawbacks of fast scoring functions

- More powerful scoring methods
  - Molecular mechanics coupled to continuum solvation models (MM-PBSA, MM-GBSA)
  - Still limited do to computationally more demanding

- Bayesian statistics
  - A known binding mode of true actives is used to translate 3D protein ligand information into a bit vector with a number of bits coding, for either atom or each residue of a binding site
  - True actives share similar binding modes
Filtering by Chemical Diversity

- No clear definition of molecular diversity
  - Dependent on the molecular descriptors and on the metric used to measure diversity

- Focusing less on individual numerical values docking scores

- Focusing more on the distribution in chemical space
  - A false negatives may be recovered if they share a common diversity with true positives
Filtering by Visual Inspection

- Looking at the predicted 3D structure of the protein ligand hit
  - The ligand has been really docked in the expected binding site
  - The bound conformation of the ligand has a physicochemical meaning
  - The ligand interacts with key residues of the active site