UNIVERSITY OF TORONTO

Protein-Ligand Docking Using AutoDock 4

Shirin Shahsavand

Professor Boris Steipe

Department of Biochemistry

Faculty of Medicine, University of Toronto

December 2, 2011

Outline

- **Introduction to Docking**
  - Basic Principles
  - The Docking Problem
  - Applications
  - Scoring Functions
- **Rigid-Body Protein-Ligand Docking**
  - DOCK
  - The CLIX Approach
  - Geometric-Hashing-Based Method
- **Flexible Protein-Ligand Docking**
  - Docking by simulation
  - Genetic Algorithms
- **AutoDock 4**

Introduction

- **Nomenclature and Useful Definitions**
  - **Ligand**: Structure (usually a small molecule) that binds to the binding site.
  - **Receptor:** Structure (usually a protein) that contains the active binding site.
  - **Binding site:** Set of amino acid residues that physically interact with the ligand (usually at 6 Angstroms).
  - **ΔG (Free Enthalpy):** A thermodynamic potential that measures the "useful" or process-initiating work obtainable from an isothermal and isobaric thermodynamic system.

Introduction – The Problem

- **The molecular docking problem**
  - Given two molecules with 3D conformations in atomic detail, do the molecules bind to each other? If yes:
    - How strong is the binding affinity?
    - How does the molecule-molecule complex look like?
- **Docking problems in biochemistry**
  - Protein-Ligand docking
    - Rigid-body docking
    - Flexible docking
  - Protein-Protein docking
  - Protein-DNA docking
  - DNA-Ligand docking

Introduction – Basic Principles

- **The association of molecules is based on interactions**
  - hydrogen bonds, salt bridges, hydrophobic contacts
  - Electrostatics
  - very strong repulsive interactions (van der Waals) on short distances
- **The associative interactions are weak and short-range**

  (tight binding implies surface complementarity)
- **Most molecules are flexible**
- **The binding affinity is the energetic difference to the uncomplexed state**
- **The binding affinity describes an ensemble of complex structures**

Binding Affinity

Free Energy of Binding

$$\Delta G = \Delta H - T\,\Delta S$$

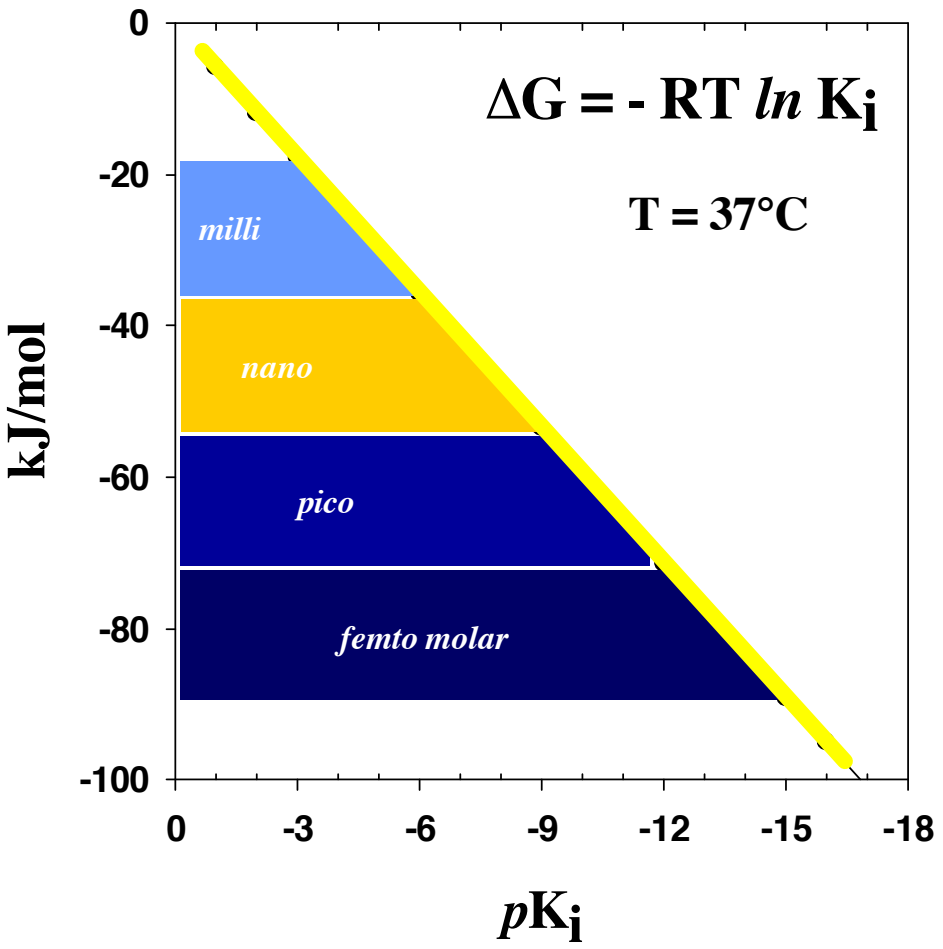
Equilibrium Constant

$$K_i = \frac{[P][L]}{[PL]}$$

~ 6 kJ/mol

≅ 1 order in K_i

"1 -2 hydrogen bonds"



Credit goes to **Matthias Rarey**
GMD - German National Research Center for Information Technology

Applications

- **Estimating the binding affinity**
  - Searching for lead structures for protein targets
  - Comparing a set of inhibitors
  - Estimating the influence of modifications in lead structures
  - De Novo Ligand Design
  - Design of targeted combinatorial libraries
- **Predicting the molecule complex**
  - Understanding the binding mode / principle
  - Optimizing lead structures

Scoring Function

- **Definition**
  - Fast approximate mathematical methods used to predict the strength of non-covalent interactions (aka binding affinity) between two molecules after they have been docked.
- **Input**
  - 3D structure of a protein-ligand complex
- **Output**
  - Estimated binding energy ΔG.
- **Comments**
  - measured ΔG describes energetic difference between bound and unbound state based on a structure ensemble.
  - measured ΔG is dominated by a single structure of minimal energy
  - $\Delta G = \Delta H - T\,\Delta S$ is very difficult to approximate

Classes of Scoring Function

- **Force field**
  - Describes only enthalpic contributions ΔH, no estimate for ΔG
  - Conformation terms (bond lengths and angles) have a steep rise (sometimes not used in docking calculations)
  - Time consuming calculations (electrostatics)
- **Knowledge-based scoring**
  - Analysis of known low-energy complexes:
    - frequent occurrence ➔ energetically favorable

Classes of Scoring Functions

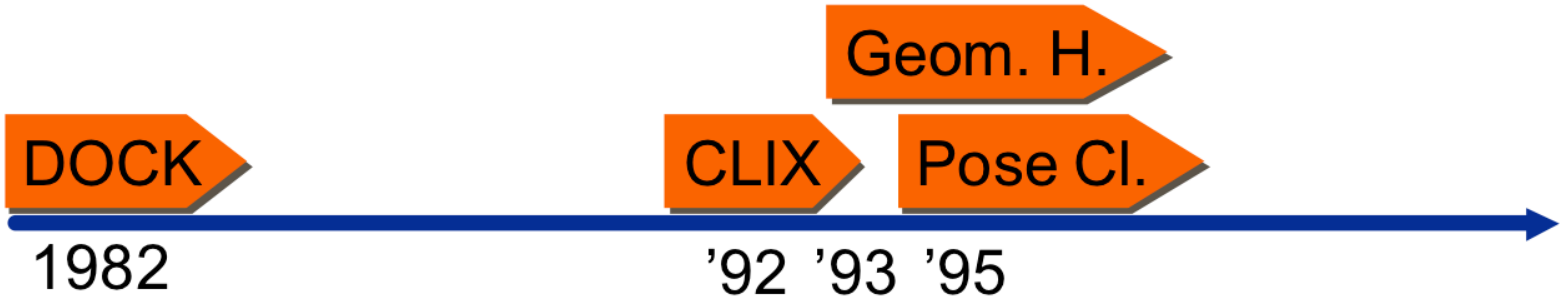
- **Empirical scoring functions**
  - Calibration of microscopic observations with measured macroscopic ΔG values
  - Data: set of protein-ligand complexes with known 3D structure and binding affinity ΔG
    - Example: Böhm-Function
      - Scoring function:

$$\Delta G = \Delta G_0 + \Delta G_{rot}\, N_{rot} + \Delta G_{hb} \sum_{\text{neutral H-bonds}} f(\Delta R) f(\Delta\alpha) +$$

$$\Delta G_{io} \sum_{\text{ionic interactions}} f(\Delta R) f(\Delta\alpha) + \Delta G_{lipo} \mid A_{lipo} \mid$$

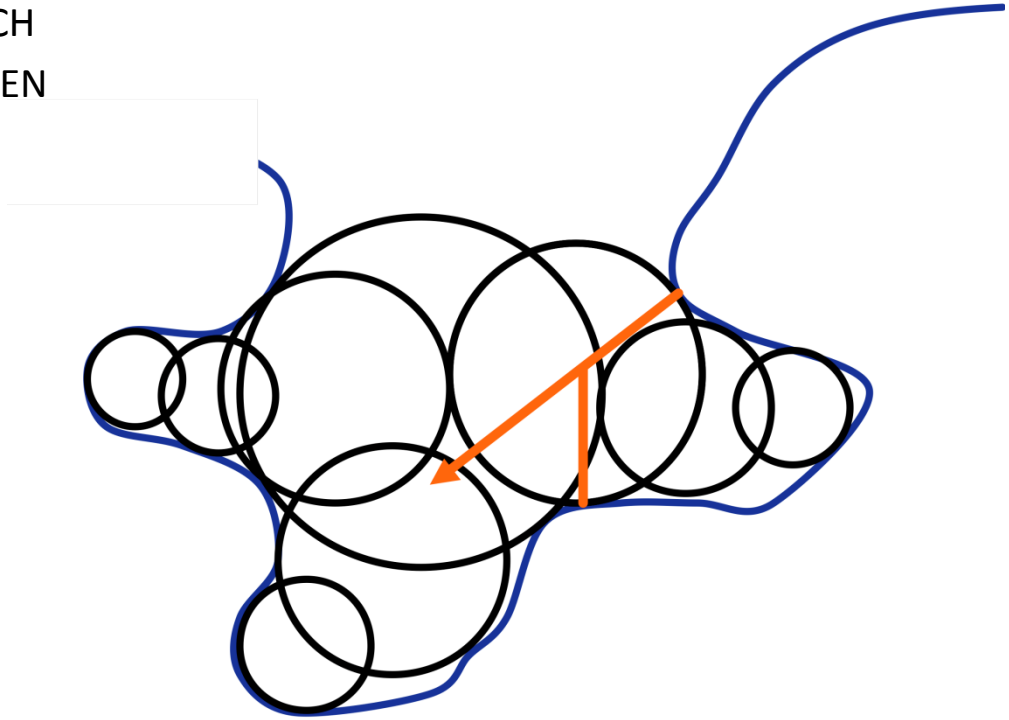
Rigid-Body P-L Docking

- **Main assumptions**
  - Protein is considered as rigid
  - Ligand is considered as rigid
- **Applications**
  - Docking of small or very rigid molecules
  - Docking of fragments (flexible docking, de novo design, combinatorial library design)
  - Docking of multi-conformer databases
- **History:**



DOCK

- **Basic Idea**
  - Represent active site by set of spheres, perform sphere matching
  - Uses 2 different algorithms
    - MATCH
    - SPHGEN



CLIX

- **Based on interaction maps calculated with GRID**
- **Algorithm**
  - Identification of *interaction target points* in the maps
  - Enumeration of all pairs of distance-compatible matches
  - superposition of two matching groups, sampling of rotation around common axis:
    - searching for additional matches
    - overlap test, scoring

Geometric Hashing

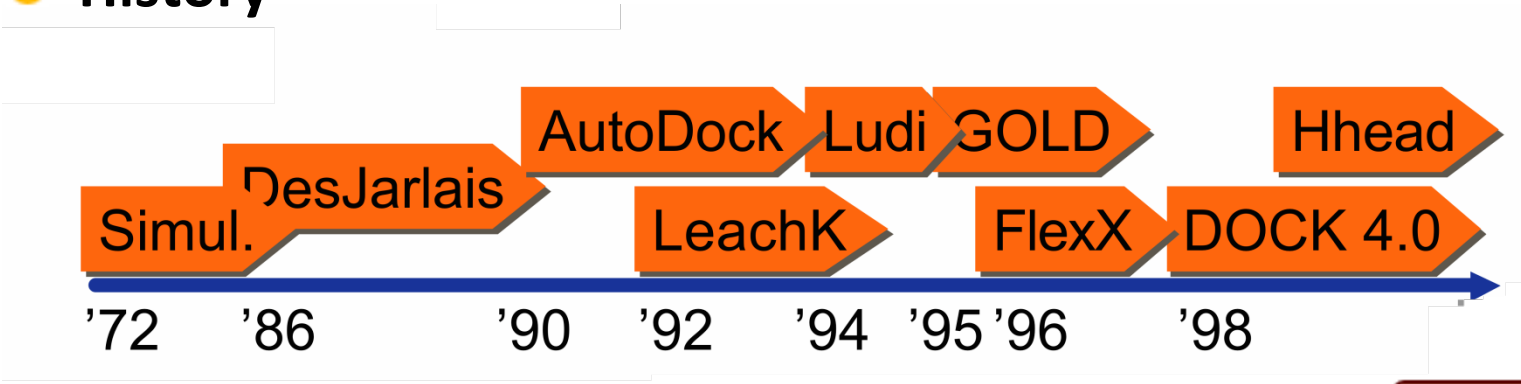
- **Key features**
  - Method from pattern recognition applied to docking
  - Based on the dock sphere representation
  - Allows direct application to database search
- **Constructing the hash table for ligand atom triplets (a,b,c):**
  - Entries have address based on atom-atom distances
  - Information stored: ligand id, basis (a,b)
- **Basic search algorithm**
  - Search for matching (two spheres, basis) allowing large number of third atom matches
  - Extension and evaluation of matches

Pose-Clustering

- **A Method from pattern recognition applied to ligand orientation based on physico-chemical interactions.**
- **General Idea**
  - Each object leads to many correct sets of correspondences, each of which has (roughly) the same pose
  - Vote on pose. Use an accumulator array that represents pose space for each object

Flexible P-LDocking

- **Main Assumptions:**
  - Ligand flexibility is limited to torsion angles and ring conformations
  - Protein is considered as nearly rigid
  - Discrete models for conformations and interactions
  - ``Binding-pathway" is not considered.
- **Applications**
  - Analyzing complexes, searching for possible binding modes
  - Virtual screening of small molecule databases
- **History**



Docking by Simulation

- **Method**
  - Generate (random) start orientations
  - MD simulation / energy minimization for all start orientations
- **Pros/Cons**
  - Can handle protein flexibility to an arbitrary extent
  - Very time consuming
  - More a local minimization (large structural changes are difficult)

Genetic Algorithms

- **GOLD and Others**
- **Genetic Algorithms**
  - General purpose discrete optimization algorithm
  - Mimics the process of evolution
- **The overall model**
  - Possible solution (configuration) → Individual
  - Its representation chromosome → Chromosome
  - Object function fitness of individual → Fitness of individual
  - Modifying solutions (moves) → Genetic operators

GOLD

- **Molecule representation (N rotatable bonds)**
  - Conformation string (N bytes), one byte each coding a torsion angle
  - A matching string (integer), defines mapping between hydrogen bon donors/ acceptors: $M(k) = l$ if k^{th} interaction group of ligand forms interaction with l^{th} group of the protein.
- **Fitness evaluation of individual with chromosome c**
  - Build conformation according to c
  - Superimpose matched interacting groups
  - Calculate docking score: $-E_{hydrogen\ bond} - (E_{internal} + E_{complex})$
- **Population**
  - 5 sub-populations of 100 individuals each
  - About 20-50 runs, each up to 100000 genetic operations
- **Genetic Operators**
  - Crossover: two-point crossover between two parent individuals
  - Mutation: one-point mutation
  - Migration: one individual moves between sub-populations
- **operators are randomly selected**

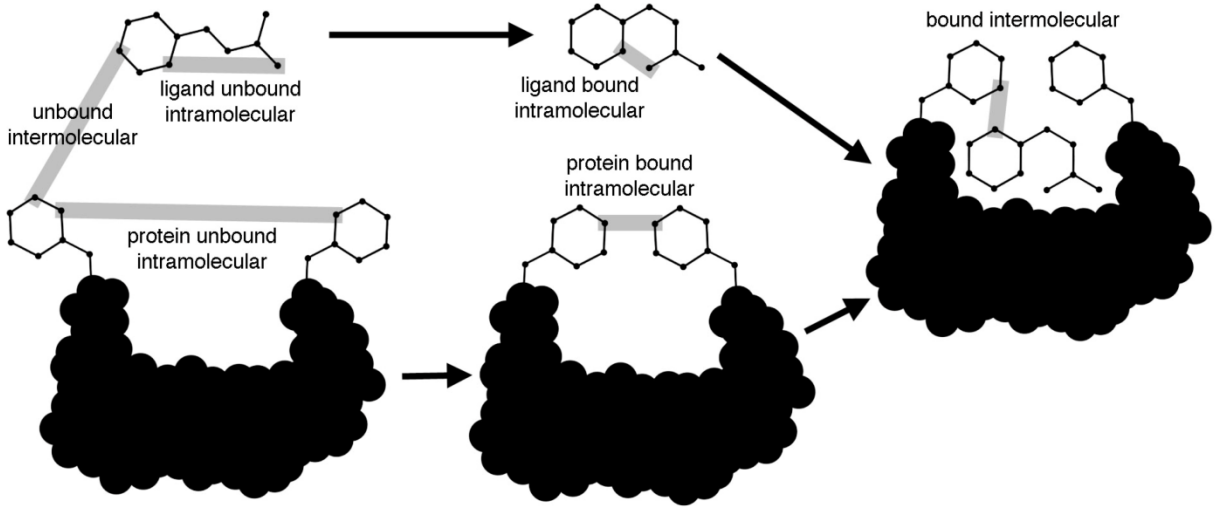
AutoDock(Finally!)

- **Designed to compromise**
  - Between accuracy and time
- **Methods**
  - Rapid grid-base energy evaluation
  - Efficient seacrh of torsional freedom
- **Steps**
  1. Coordinate file preparation
  2. AutoGrid calculation
  3. Docking using AutoDock
  4. Analysis using AutoDock Tools (ADT)

AutoDock 4 – Theory

- **Free Energy Function**
  - Semi-empirical free energy force field
  - The force field is parameterized using a large number of protein-inhibitor complexes for which both structure and inhibition constants are known.
- **Force Field Evaluation**
  - The ligand and protein start in an unbound conformation

  1. *Intramolecular* energies are estimated for the transition from these unbound states to the conformation of the P-L in bound state.
  2. *Intermolecular* energies of combining the ligand and protein in their bound conformation is evaluated.



AutoDock 4 - Theory

- The force field includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf})$$

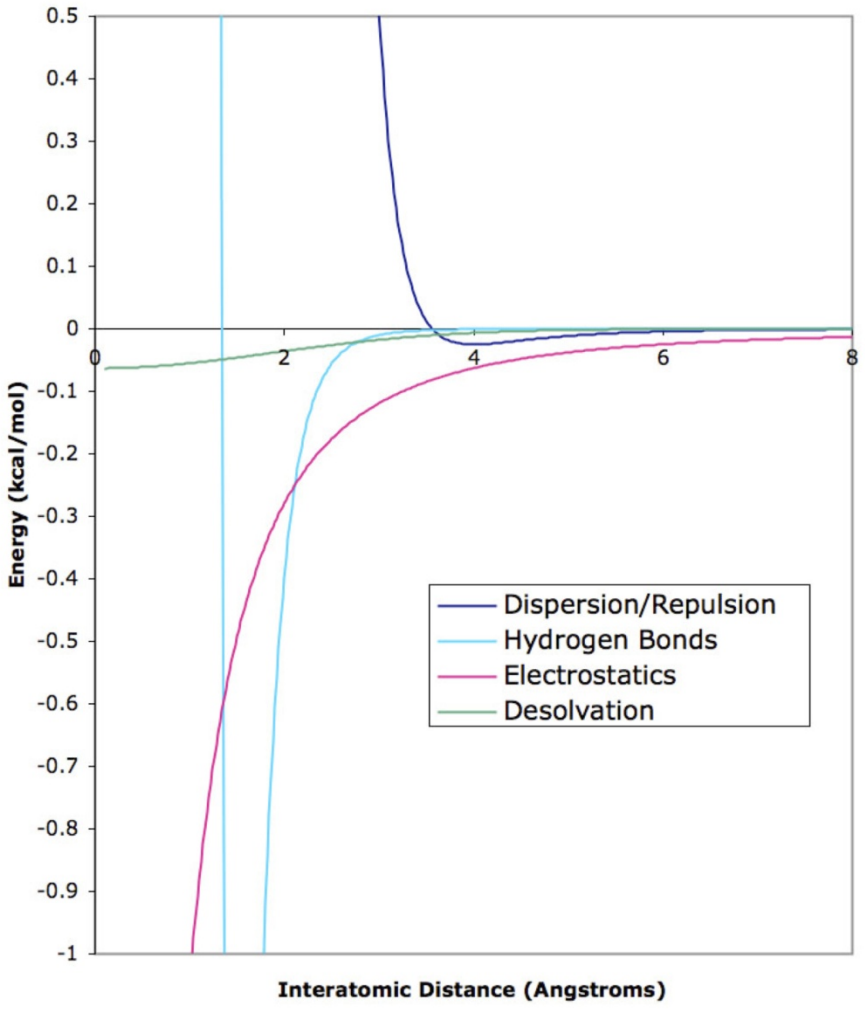
Where *L* refers to the ligand and *P* refers to the protein in a ligand protein docking calculation.

- Each of the pair-wise energetic terms includes evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation:

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{(-r_{ij}^2 / 2\sigma^2)}$$

AutoDock 4 - Theory

- **Force Field**



AutoDock 4 – Search Methods

- **AutoDock 4 search method is stochastic.**

- **Methods**
  - Global search algorithms:
    - Simulated Annealing (Goodsell et al. 1990)
    - Genetic Algorithm (Morris et al. 1998)
  - Local search algorithm:
    - Solis & Wets (Morris et al. 1998)
  - Hybrid global-local search algorithm:
    - Lamarckian GA (Morris et al. 1998)

AutoDock 4 – Search Methods

Simulated Annealing

- Ligand starts at a random (or user-specified) position/orientation/ conformation ('state')
- Constant-temperature annealing cycle:
  - Ligand's state undergoes a random change.
  - Compare the energy of the new position with that of the last position; if it is:
    - lower, the move is 'accepted';
    - higher, the move is accepted if $e(-\Delta E/kT) > 0$;
    - otherwise the current move is 'rejected'.
  - Cycle ends when we exceed either the number of accepted or rejected moves.
- Annealing temperature is reduced, $0.85 < g < 1$
  - $T_i = g\, T_{i-1}$
- Repeat.
- Stops at the maximum number of cycles.

AutoDock 4 – Search Methods

- **Genetic Algorithm**
  - Start with a random population (50-300)
  - **Genes** correspond to **state variables**
  - Perform genetic operations
    - **Crossover**
      - 1-point crossover, ABCD + abcd → Abcd + aBCD
      - 2-point crossover, ABCD + abcd → AbCD + aBcd
      - uniform crossover, ABCD + abcd → AbCd + aBcD
      - arithmetic crossover, ABCD + abcd → [α ABCD + (1- α) abcd] + [(1- α) ABCD + α abcd] where: $0 < \alpha < 1$
    - **Mutation**
      - add or subtract a random amount from randomly selected genes, A→ A'
  - Compute the **fitness** of individuals (energy evaluation)
  - Proportional Selection & Elitism
  - If total energy evaluations or maximum generations reached, stop

AutoDock 4 – Search Methods

- **Lamarckian GA**
  - phenotypic adaptations of an individual to its environment can be mapped to its genotype & inherited by its offspring.
  - Typical GA evolves the **Genotype** (state variables)—the **Phenotype** (atomic coordinates) is used to evaluate the **fitness** of each individual
  - Lamarkian GA allows local optimization of the Phenotype, which is then applied to the Genotype
  - Advantage that it does not require gradient information in order to proceed.

AutoDock 4 – Search Methods

- **What range of problems is feasible?**
  - Depends on the search method:
    - LGA > GA >> SA >> LS
    - SA : can output trajectories, D < about 8 torsions.
    - LGA : D < about 8-16 torsions.
- **When is AutoDock not suitable?**
  - Modeled structure of poor quality;
  - Too many torsions (32 max);
  - Target protein too flexible.

AutoDock 4 – Step 1

Preparing the coordinates

- Prepare the ligand and receptor coordinate file for AutoGrid and AutoDock
  - **Known as PDBQT files, includes**
    1. Polar hydrogen atoms;
    2. Partial charges;
    3. Atom types;
    4. Information on the articulation of flexible molecules.
  - **Can use ADT to create the file using the following steps**

    (Edit menu)

    1. Add hydrogen atoms to the molecule. (Edit menu)
    2. Add partial charges.

    (Read the molecule into AutoDockTools using the Ligand (for the ligand) or Grid (for the

    receptor) menus, and create the PDBQT file)

    3. Delete non-polar hydrogens and merge their charges with the carbon atoms.
    4. Assign atom types, defining hydrogen bond acceptors and donors and aromatic and aliphatic carbon atoms.
    5. Choose a root atom that will act as the root for the torsion tree description of flexibility.
    6. Define rotatable bonds and build the torsion tree.

AutoDock 4 – Step 2

- **Running AutoGrid**
  - Requires pre-calculated grid maps, one for each atom type present in the ligand being docked.
- **Grid Maps**
  - Calculated by AutoGrid.
  - Consists of a three dimensional lattice of regularly spaced points, surrounding (either entirely or partly) and centered on some region of interest of the macromolecule under study. This could be a protein, enzyme, antibody, DNA, RNA or even a polymer or ionic crystal.
  - Typical grid point spacing varies from 0.2Å to 1.0Å, and the default is 0.375Å (roughly a quarter of the length of a carbon-carbon single bond). Each point within the grid map stores the potential energy of a 'probe' atom or functional group that is due to all the atoms in the macromolecule.

To run AutoGrid, the command is issued as follows:

```
% autogrid4 -p macro.gpf [-l macro.glg]
```

AutoDock 4 – Step 2

- **Why use GridMaps?**
  - Saves time:
    - Pre-computing the interactions on a grid is
    - typically 100 times faster than pairwise methods
    - $O(N^2)$ calculation becomes O(N)
  - AutoDock needs one map for each atom type in the ligand(s) and moving parts of receptor (if there are any)
  - Drawback: The receptor is conformationally rigid (although 'vdW softened')
  - Limits the search space

AhutoDock 4 – Step 2

Creating grid parameter files in AutoDockTools

- The tools available in "grid" menu of AutoDockTools may be used to create grid parameter files.
  - **`Grid>OpenGPF`**: gets parameters from an existing grid parameter file.
  - **`Grid>Macromolecule`**: has options for opening an existing PDBQT file or choosing a molecule that has been read using PMV.
  - **`Grid>SetMapTypes`**: tools to define the atom types for the grids that will be calculated. Grids must be calculated for each type of atom in the ligand, and if flexible sidechains are used in the receptor, their atom types must also be included. The option "Directly" allows the user to input the list of atom types directly. Other options allow the user to define the atom types based on a ligand or flexible residue that has been read by PMV, or to open ligand or flexible residue PDBQT anduse the atom types in these files.
  - **`Grid>SetMapTypes>SetUp CovalentMap`**: specifies parameters for creation of a covalent map, which may be used in specialized applications to favor binding of a given ligand atom in a single position. This is particularly useful for docking of covalent complexes between ligands and proteins. This will calculate a separate grid with atom type "Z" with a favorable Gaussian well at the coordinates given. The potential will have zero energy at the site, rising to the energy barrier height in surrounding areas.
  - **`Grid>GridBox`**: launches interactive commands for setting the grid dimensions and center. To enter numbers on the thumbwheel, place the cursor over the thumbwheel and type in the new value. Right clicking on the thumbwheel gives more options. IMPORTANT: when finished, use the "close saving current" option in the "File" menu on the Grid Options Panel. Options in the "Center" menu on the browser provide different methods to choose the center of the grid box.
  - **`Grid>OtherOptions`**: allows specification and editing of an existing parameter file.
  - **`Grid>Output`**: writes a new grid parameter file.
  - **`Grid>EditGPF`**: interactive editor for grid parameter files, which allows viewing of the lastestgrid parameter file written by AutoDockTools.

AutoDock 4 – Step 3

- **Docking with AutoDock**
  - Requires (i) grid maps calculated using AutoGrid, (ii) PDBQT file for the ligand, and (iii) a docking parameter file that specifies the files and parameters for the docking calculation.

An AutoDock calculation is processed after the execution of the following command:

```
% autodock4 [-k][-i][-u][-t] -p lig.dpf [-l lig.dlg]
```

AutoDock 4 – Step 3

- **Protocols**

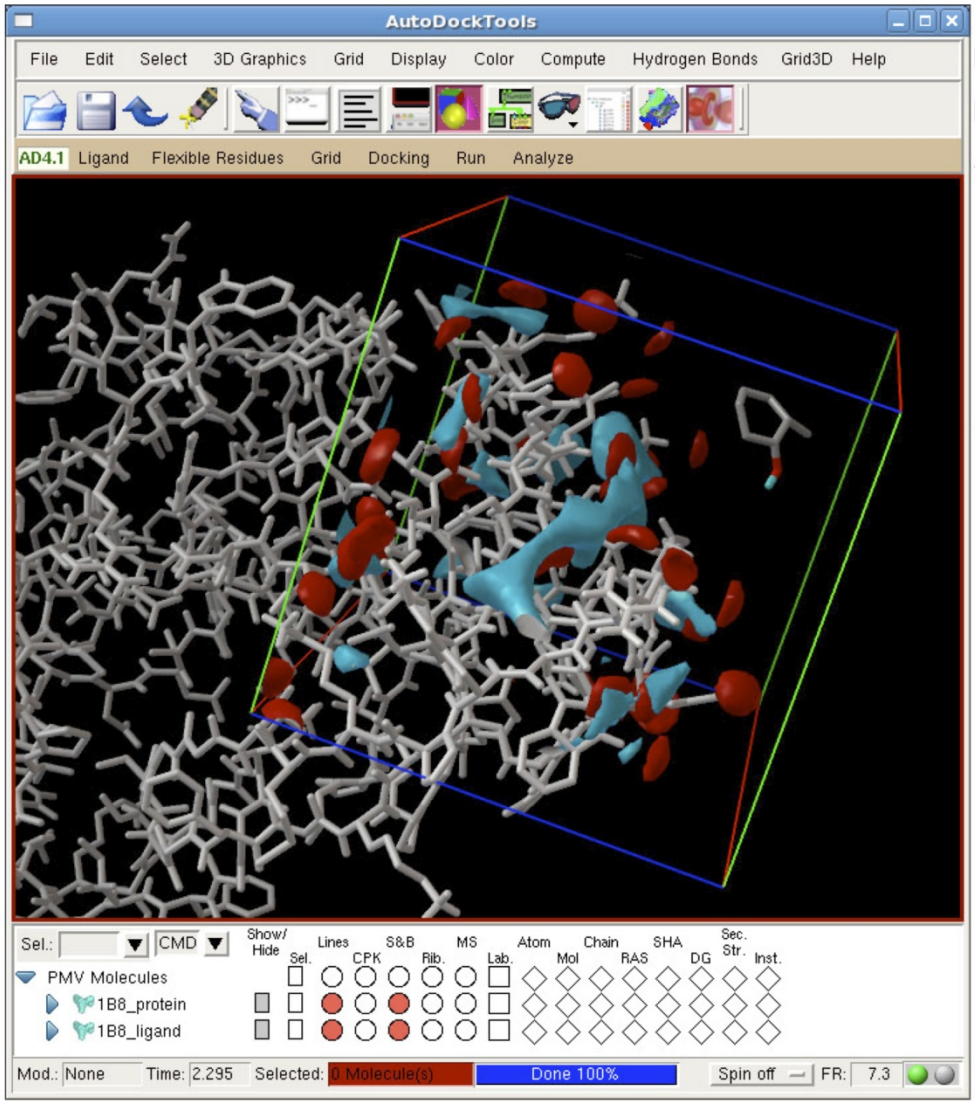
  1. ***Conformation Search***
     - Using mainly a Lamarckian Genetic Algorithm that provides the most efficient search for general applications.
     - It is typically effective for systems with about 10 rotatable bonds in the ligand.
  2. ***Number of Evaluations***
     - Each of the search methods include parameters for determining the amount of computational effort that will be used in the search. In the GA methods, this parameter is `ga_num_evals`, and in simulated annealing, this is `nacc and nrej`.
  3. ***Model for Unbound Ligand***
     - AutoDock assumptions.
  4. ***Special Cases***

AutoDock 4 – Step 4

- **Evaluate the results of Docking**

  1. ***Information in the Docking Log***
     - Contains the result of the analysis sing the `analysis` command as histograms and tables of RMSD values within each cluster.
     - AutoDock then writes the best conformation coordinates in each cluster.
     - The header for each conformation include the informationon predicted energy of binding.
  2. ***Analyzing Docking Results with ADT***
     - A Set of menu options in ADT can be used to analyze the docking results.
     - For in-depth information refer to page 21 – 23 of the AutoDock 4.2 userguide.

AutoDock Tools



PMV Menus

PMV Buttons

ADT Menus

Viewer

Dashboard

PMV Info

ADT

- For a list of useful menu options please refer to the provided user guide.

Autodock Applications

- **Blind Docking**
  Hetenyi et al. had already showed that AutoDock can be used to perform docking of peptides to their targets, without *a priori* knowledge of the location of the binding site. Now, they have shown that AutoDock can be used for drug-sized molecules (Hetenyi and van det Spoel (2006) "Blind docking of drug-sized compounds to proteins with up to a thousand residues" FEBS Letters 580: 1447-1450).

References

- Garrett M. Morris, et al. “User Guide for AutoDock Version 4.2: Automated Docking of Flexible Ligands to Flexible Receptors”
- Morris, G. M, et al. “Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function.” J. Comput. Chem., 1998, 19, 1639-1662.
- Huey, R., et al. “D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation.” J. Comput. Chem., 2007, 28, 1145-1652.
- Morris, G.M., et al. “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.” J. Comput. Chem.
- Rarey, M. “Protein-Ligand Docking” Presentation.
- http://www.cs.cmu.edu/~blmt/Seminar/SeminarMaterials/interactions.html

Thank You

Protein-Ligand Docking Using AutoDock 4

Shirin Shahsavand

shirin.shahsavand@utoronto.ca