Docking: Methods Overview

Vlad Kholodovych

University of Medicine & Dentistry
of New Jersey (UMDNJ)

kholodvl@UMDNJ.EDU
http://www2.umdnj.edu/~kholodvl

Docking

What:

Prediction of the optimal configuration and energy between two molecules

Why:

Assumption: components that dock well may bind to each other

How:

Change orientations of molecules to maximize their interaction

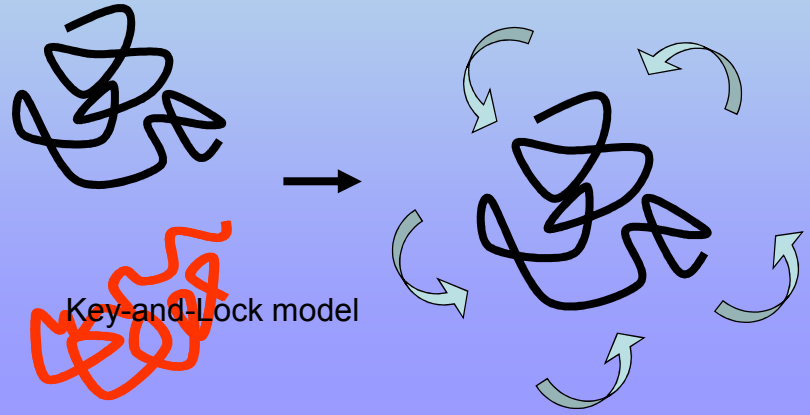
Minimize the total energy of interaction
(*The best binding pose has the minimal binding energy*)

RIGID DOCKING

Both molecules are rigid

Orientation change: translation, rotation

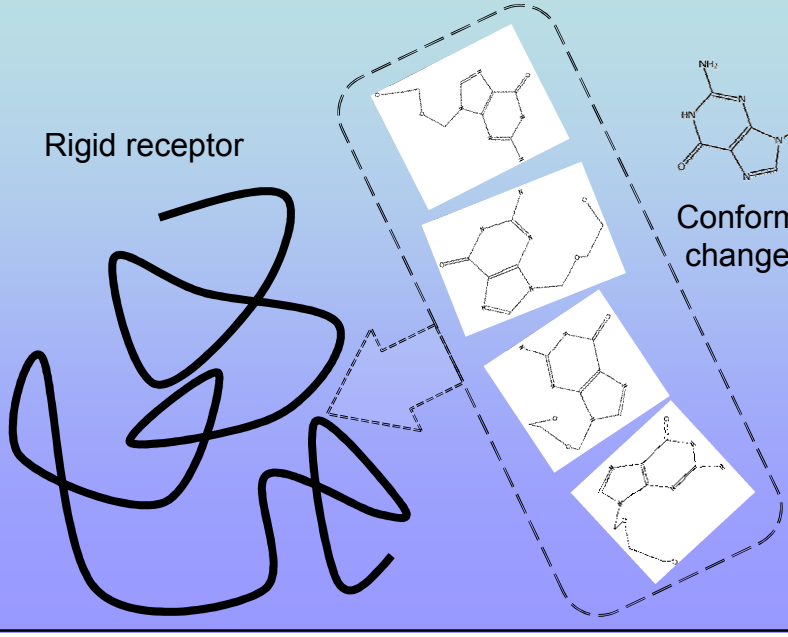
No change in conformation

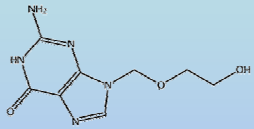


Key-and-Lock model

FLEXIBLE DOCKING

Rigid receptor





Conformational changes in ligand

Type of Docking Size Wise

Protein-Protein Docking

- Bound docking ("rigid" redocking):
  6 degrees of freedom: 3 for rotation, 3 for translation
- Unbound docking : side chain flexibility

Protein-Small Molecule Docking

- Rigid receptor, rigid ligand
- **Rigid receptor, flexible ligand**
- Flexible receptor, flexible ligand

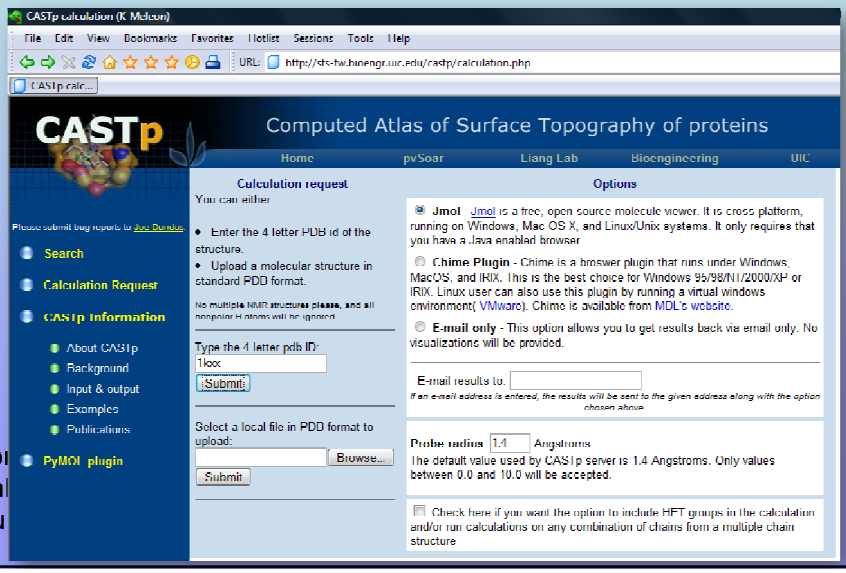
Receptor - Ligand Docking: Specifics

- Many possible sites for potential interactions
- Extremely large search space
  - Both molecules are flexible – hundreds to thousands of degrees of freedom (DOF)
  - Enormous number of possible docking poses

Receptor - Ligand Docking

Known protein.

No ligand, no information about site interactions

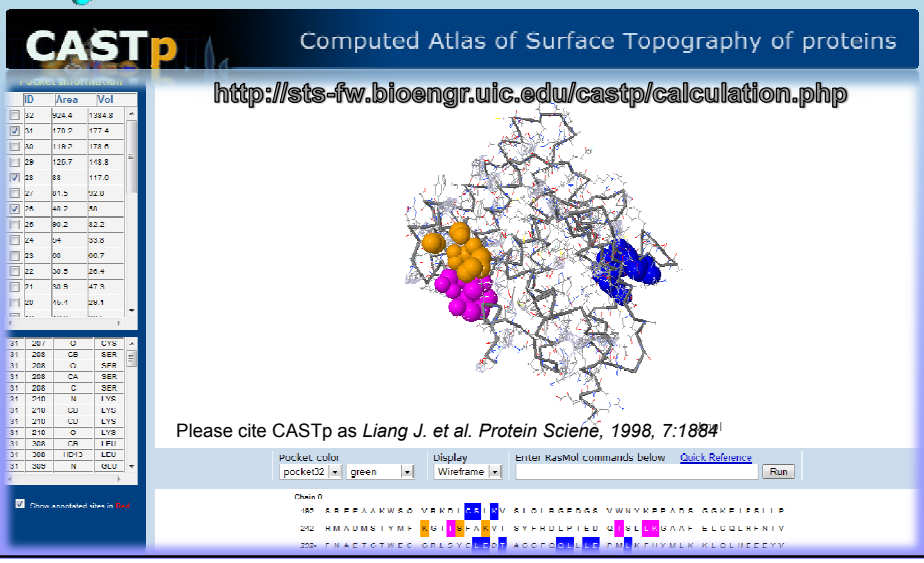


Hol
Onl
Mu

Receptor - Ligand Docking

Known protein.

No ligand, no information about site interactions



Protein - Ligand Docking

Rigid Docking

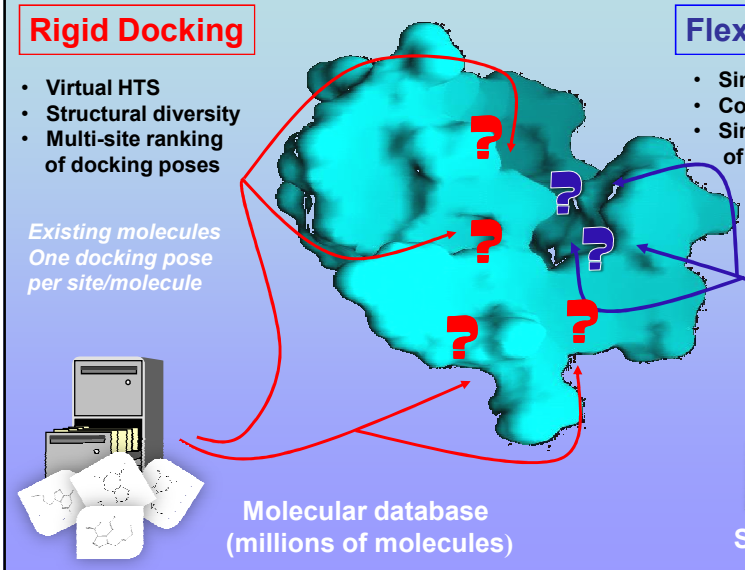
- Virtual HTS
- Structural diversity
- Multi-site ranking of docking poses

Existing molecules
One docking pose per site/molecule

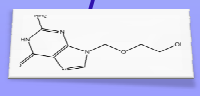
Flexible Docking

- Single/few molecule
- Coformational diversity
- Single site ranking of docking poses

Systematic scan
Multi-pose output



Molecular database
(millions of molecules)

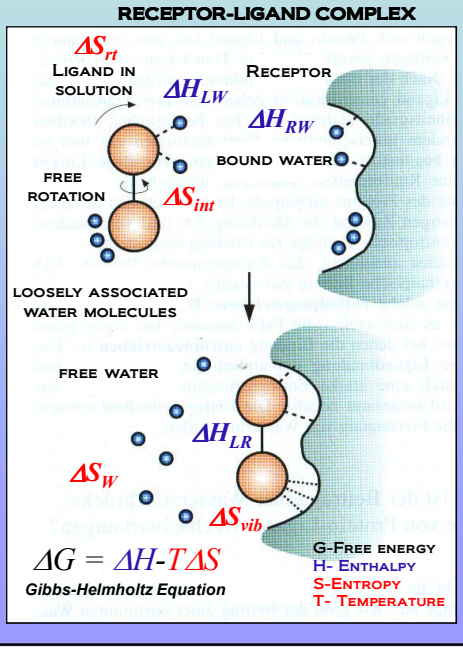


Single molecule

Known Target Protein: Screening

	Rigid	Flexible
Database search	***YES***	***NO***
Multi-conformational	***NO***	***YES***
Basis	***Volume/Surface complementarity***	***Interaction Energy***
Algorithm	***Clique matching (shape)***	***Simulated Annealing Genetic Algorithm Fragment Building***
Programs	***Dock GOLD GRAMM MOE***	***AutoDock, Dock FlexX GOLD MOE***

DOCKING: DIFFICULTIES



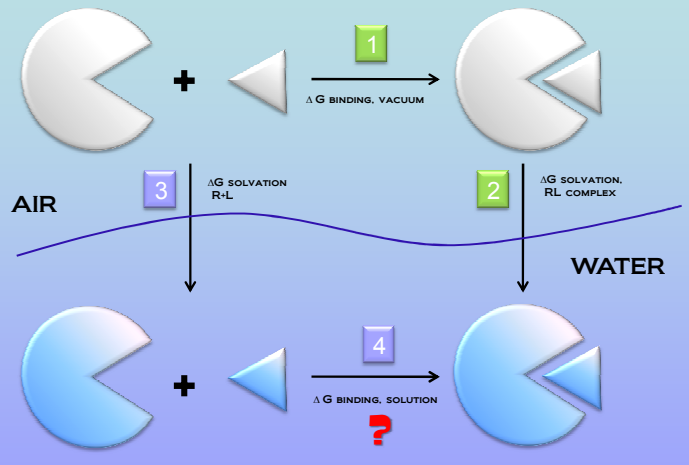
Multiple steps in the receptor – ligand interaction:

- **Approach**
- **Desolvation of the ligand and the binding site of a protein**
- **Penetration into the protein cavity**
- **Change of the ligand orientation**
- **Adoption of the correct "active" conformation**
- **Establishing of new H-bonds, electrostatic and hydrophobic contacts**

In general free energy function is formed from several terms

$$\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elect} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol}$$

SOLVATION BINDING



Hess's law of heat summation:

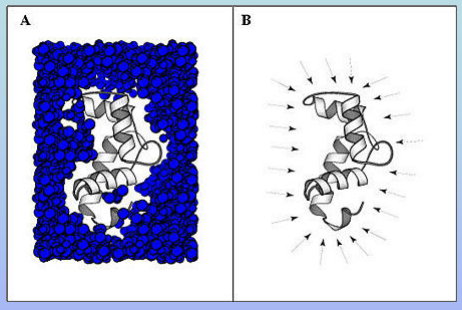
The change in free energy between two states will be the same, no matter what the path.

Tricky math
1+2= 3+4
4=1+2-3

The energy of binding in solvent can be calculated as followed:

$$\Delta G_{binding,solution}^{4} = \Delta G_{binding,vacuum}^{1} + \Delta G_{solvation(RL)}^{2} - \Delta G_{solvation(R+L)}^{3}$$

SOLVATION: EXPLICIT VS IMPLICIT



the N terminal domain of Troponin

*A A view of the inside of the initial box for a simulation which includes water molecules explicitly; the whole system consits of **1229** atoms in the protein and **6009** atoms for water.*

B In a implicit solvent simulation, the effect of water is included in a potential of mean force, W(X), vizualized as arrows. In this case, the simulation only includes the 1229 protein atoms.

Generalized Born / Surface Area - GB/SA model

$$G_{solv} = G_{cav} + G_{vdW} + G_{pol}$$

$$G_{cav} + G_{vdW} = \sum_{k=1}^{N} ASP(k)ASA(k)$$

$$G_{pol} = -166.0\left(1 - \frac{1}{\epsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp\left(-\frac{r_{ij}^2}{4 b_i b_j}\right)}}$$

$$G_{Born} = -\frac{q^2}{2b}\left(1 - \frac{1}{\epsilon}\right)$$ Born Equation

Born radius ***b*** is the effective radius of the atom defined by Born equation

the total solvation free energy G_{solv} is given as the sum of a solvent-solvent term (G_{cav}), a solute-solvent van der Waals term (G_{vdW}), and a solute-solvent electrostatics polarization (G_{pol})

Gcav + GvdW are computed together as a function of the solvent-accessible surface areas. where **ASA(k)** is the total solvent accessible surface area of atom k and **ASP(k)** is an empirical atomic solvation parameter.

where the double sum runs over all pairs of atoms (i and j). $\mathbf{q_i}$ and $\mathbf{q_j}$ are the partial charges of i and j, respectively, and $\mathbf{r_{ij}}$ is the i,jth atom pair separation. $\boldsymbol{\varepsilon}$ is the dielectric constant of the medium. $\mathbf{b_i}$ and $\mathbf{b_j}$ are the so-called Born radius of atom i and j, respectively .

Take Home Message

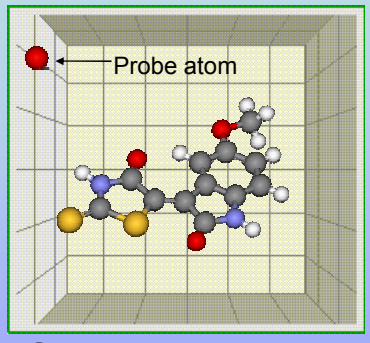
- **Binding energy** is a multicomponent function that is not easy to calculate
- Solvation component of binding energy is computationally demanding to be solved **explicitly**
- Instead, the **implicit** solvation term is included in a potential of mean force field, while only essential solvent molecules mediated protein-ligand interactions in the binding pocket are included explicitly
- GB/SA method is the most popular way to calculate solvation
- Docking programs use simple and fast methods to estimate binding interactions. Solvation factor is not included in the scoring function.

Two Important Questions in Docking

1. How to align (place) the ligand molecule into the binding pocket?
2. How to alter a ligand structure?

Docking alignment: grid

- Grid-based docking (AutoDock)
- Ligand-protein interaction energies are pre-calculated and then used as a look-up table during simulation

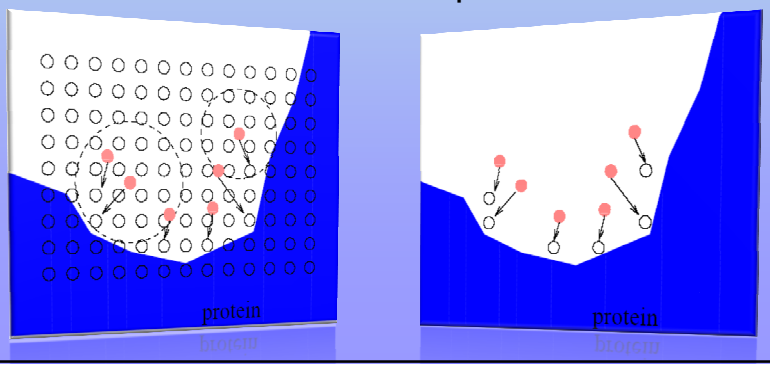


- Grid maps are constructed for each atom type presented in the ligand

Docking alignment: grid

Each ligand atom is matched with the grid point with the lowest energy within its neighborhood

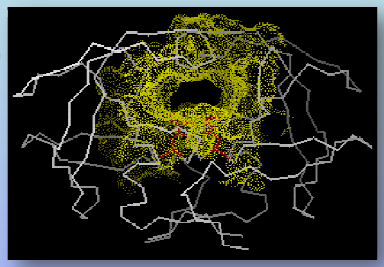
Total binding energy is estimated by summation from different “element” maps

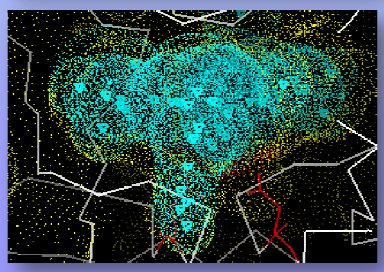


Docking alignment : spheres

Spheres-based docking (UCSF Dock)

The receptor cavities are filled up with spheres.
Sphere centers are then matched to the ligand atoms to determine possible orientations for the ligand.
Typically 10^4-10^5 orientations are generated for each ligand molecule.



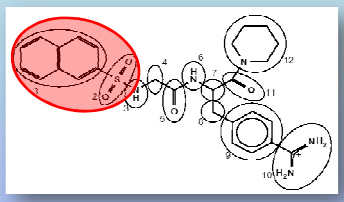


The sphere centers are identified by cyan triangles, and the sphere surfaces are shown in dots

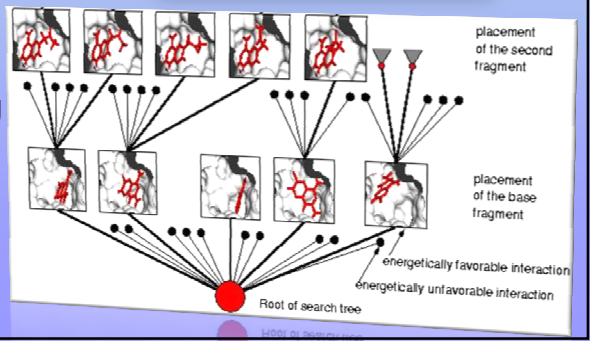
Docking alignment: fragment addition

FlexX and DOCK

- Split the ligand into fragments
- Determine the rigid core/root



- Incrementally add the fragments to the core/root frame during docking optimization



Ligand Modification Algorithms

Simulated Annealing

- Based on temperature effects
- Start with high temperature ⟶ Global search
- Lower temperature ⟶ Local search

Genetic Algorithm

- Charles Darwin's Theory of Evolution
  Genotype ⟶ Phenotype
- Lamarckian Algorithm (Jean –Baptiste de Lamarck)
  Phenotype ⟶ Genotype

Simulated Annealing

Simulated annealing is a global optimization technique based on the Monte Carlo method.

Generate small random changes in the current state (ligand coordinates)

Metropolis criterion :

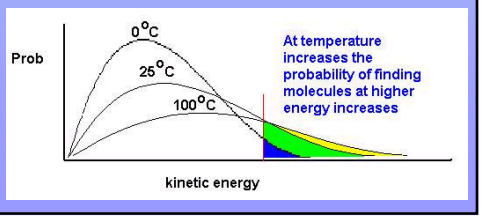
moves that decrease the energy of the system are always accepted

moves that increase the energy of the system are accepted only if probability *p* is larger than random chosen number between 0 and 1

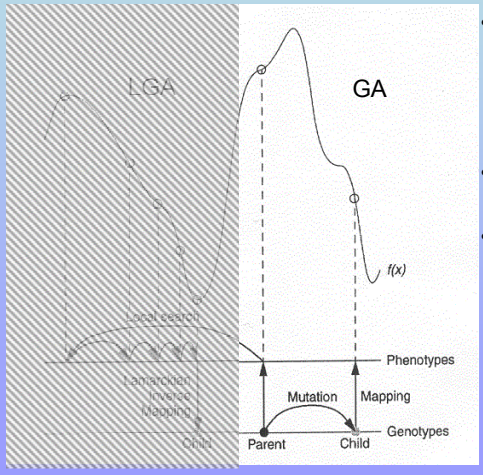
$$p = e^{\left(\frac{-\Delta E}{kT}\right)}$$ where $\Delta E = E_1 - E_0$

E_1 is an energy of the new state, E_0 is an energy of the current state, T is the "temperature" of the simulation, and k is Boltzmann's constant

At high temperatures, many states are accepted, while at low temperatures, the majority of these probabilistic moves are rejected



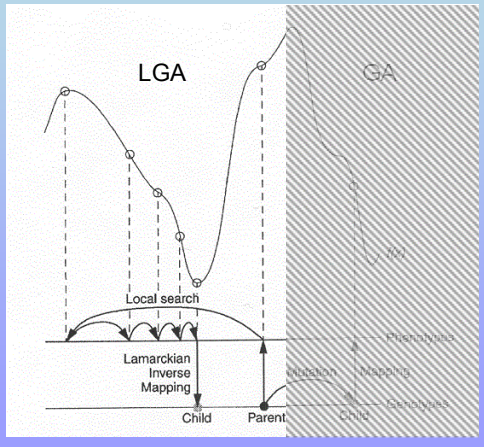
Genetic Algorithm



- In standard GA, the genotype (x,y,z coordinates plus rotation and any torsion angles) are mapped to the fitness function f(x)
- All new generation based only on parents genes
- Genotypes of parents with high f(x) values are mutated forming genotypes of children with lower f(x) values

Lamarckian Genetic Algorithm

- LGA finds lowest fitness function (energy) values first, then maps these values to their respective genotypes
- Each new child is allowed to create a new generation
- Genetic algorithm plus Solis and Wets local search
- Better performance than either simulated annealing or genetic algorithm alone



SCORING FUNCTIONS

Purpose: Evaluate the placement (and possibly a conformation) of the ligand-protein pair.

- **FORCE FILED FUNCTIONS:** use non-bonded energies of force fields (FlexiDock, Tripos, Amber, CHARMM)

- **EMPIRICAL SCORING FUNCTIONS:** derived from a set of protein-ligand complexes with measured binding affinity (more chemically appealing, however require significantly more statistical fitting than force fields functions (ChemScore, AutoDock)

- **KNOWLEDGE-BASED SCORING FUNCTIONS:** a statistical analysis of structures of protein-ligand complexes derived from databases, e.g. PDB. ("PMF", DrugScore).

Each scoring function has been derived from a different set of crystal structures and experimental data. Thus, it is reasonable to use multiple functions when evaluating docking results.

Consensus: structures which are considered as a good fit by multiple scoring functions should be tested first

C-SCORE

- **F-SCORE** FlexX scoring function (ref: Rarey, M.; Kramer, B.; Lengauer, T., Klebe, G. J. *J. Mol. Biol* (1996) **261**, 470-489

- **G-SCORE** Includes terms for hydrogen bonding, complex (ligand-protein) and internal (ligand-ligand) energies (ref : Jones, G. Willett, P.,Glen, R. Leach, A.R., and Taylor, R. *J. Mol. Biol*. (1997) **267**, 727)

- **PMF-SCORE** Potential Mean ForceDeveloped Helmholtz free energies of interactions for protein-ligand atom pairs (ref: Muegge, I., Martin, Y.C., *J. Med. Chem*. (1999) **42**, 791)

- **D-SCORE** Uses only charge and van der Waals interactions between the protein and the ligand (ref: Kuntz, I.D., Blaney, J.M., Oatley S.J., Langridge, R., Ferrin, T.E., *J. Mol. Biol.* (1982) **161**, 269)

- **CHEMSCORE** Includes terms for H-Bonding, metal-ligand interaction, lipophilic contact, and intercept term and rotational entropy (ref: Eldredge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V., and Mee, R.P. *J. Comp.-Aided Molec Des.*,(1997) **11**, 425)

- **C-SCORE** is the sum of "good" score marks, min 0 to max 5.

Programs

Protein-Protein Docking

- **FTDock** algorithm maps both molecules onto orthogonal grids and performs a global scan of translational and rotational space. The scoring method is primarily a surface complementarity score between two grids.

http://www.bmm.icnet.uk/docking/ftdock.html

- **MultiDock** is a program for refinement of FTDock results. Allows generation of side chain conformations based on a rotamer libraries.

www.bmm.icnet.uk/docking/multidock.html

Protein-Protein Docking

GRAMM:
Global Range Molecular Matching

Main features of GRAMM:

- requires only the atomic coordinates of the molecules (no information about the binding sites is needed)
- performs an exhaustive 6-dimensional search through the relative translations and rotations of the molecules
- molecular pairs: two proteins, a protein and a smaller compound, two transmembrane helices, etc.

http://vakser.bioinformatics.ku.edu/main/resources_gramm1.03.php

Ligand-Protein Docking

NAME: AutoDock
Authors: Olson et al 1990, Morris et al 1998
Publisher: Scripps
Ligand Alignment:
Grid
Search Algorithms:
Simulated Annealing
Genetic Algorithm
Lamarckian GA (GA+LS hybrid)

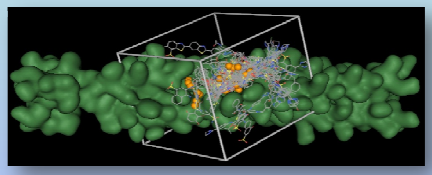
Unique features: Solvation term included in the scoring function, LGA

http://autodock.scripps.edu

The Scripps Research Institute

Ligand-Protein Docking

NAME: UCSF DOCK



Authors: Kuntz et al. 1982;
Ewing & Kuntz 2001

Publisher: UCSF

Ligand Alignment:
Spheres

Search Algorithms:
Fragment-Based

Unique features: Ligand is matched to spheres, grid scoring

http://dock.compbio.ucsf.edu/

Ligand-Protein Docking

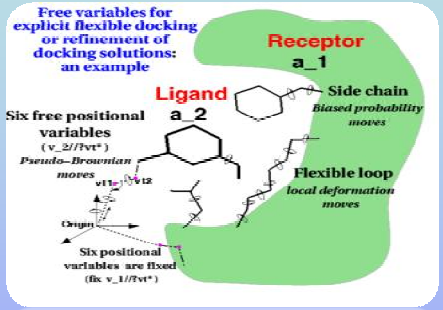
NAME: ICM Dock

Authors: Abagyan et al. 1994

Publisher: MolSoft

Ligand Alignment:
Grid

Search Algorithms:
Genetic Algorithm



Unique features: Protein/Protein, Protein/Peptide, Protein/Ligand docking; Partial flexibility of the protein

http://www.molsoft.com/docking.html

MolSoft
molecules in silico

Ligand-Protein Docking

Cambridge Crystallographic Data Centre

Genetic Optimisation for Ligand Docking

NAME: GOLD

Authors: Jones et al 1997

Publisher: Cambridge Crystallographic Data Center

Ligand Alignment:

Fitting Points

Search Algorithms:

Genetic Algorithm

Unique features: Automatic consideration of cavity bound water molecules; Partial protein flexibility (side chains and backbone for up to 10 residues)

Free on-line test run at:
http://gold.ccdc.cam.ac.uk/setup_and_run_docking.php

http://www.ccdc.cam.ac.uk/prods/gold/index.htm

Ligand-Protein Docking

NAME: MOE Dock

Authors:

Publisher: Chemical Computing Group

Ligand Alignment:

Spheres, pharmacophore triad

Search Algorithms:

Simulated Annealing
Genetic Algorithm

Unique features: Pharmacophore mapping/filtering, solvation term is included in ligand conformers generation

http://www.chemcomp.com/software-sbd.htm

Ligand-Protein Docking

NAME: FlexX
Authors: Rarey, Lengauer 1996
Publisher: BioSolveIt
Ligand Alignment:
Fragment Based
Search Algorithms:
Fragment-Based

Unique features: Pharmacophore based docking; docking into ensemble of protein structures; automatically assigned metal coordination geometry; high speed vHTS

http://www.biosolveit.de/FlexX/

Ligand-Protein Docking

NAME: Glide
Authors: Friesner et al 2004
Publisher: Schrodinger
Ligand Alignment:
Grid
Search Algorithms:
Simulated Annealing

Unique features:

http://www.schrodinger.com/ProductDescription.php

Evaluation of Docking

Root Mean Square Deviation (*RMSD*)

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

n - number of atoms

d_i - distance between two corresponding atoms i in two structures

Unit of RMSD: Ångstroms

Identical structures: *RMSD* = "0"

Similar structures: *RMSD* is small (1 – 2 Å)

Distant structures : *RMSD* > 3 Å

MOE and GOLD Docking Tutorial

http://www2.umdnj.edu/~kholodvl