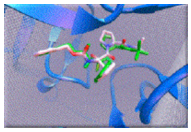
Using AutoDock 4 with ADT: A Tutorial

Dr. Ruth Huey
&
Dr. Garrett M. Morris

What is Docking?



"Predicting the best ways two molecules will interact."

1. Obtain the 3D structures of the two molecules.
2. Locate the best binding site.
3. Determine the best binding modes.

What is Docking?

*"Predicting the **best** ways two molecules will interact."*

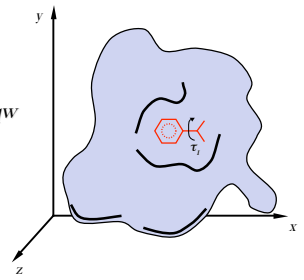
- We need to *quantify* or *rank* solutions;
- We need a ***Scoring Function*** or force field.

*"Predicting the best **ways two molecules will interact**."*

- *(ways—plural)* The experimentally observed structure may be amongst one of *several predicted solutions*.
- We need a ***Search Method***.

Defining a Docking

- **Position**
  - *x*, *y*, *z*
- **Orientation**
  - *qx*, *qy*, *qz*, *qw*
- **Torsions**
  - τ_1, τ_2, ... τ_n



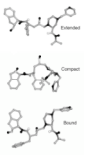
Key aspects of docking...

- Scoring Functions
  - *What are they?*
- Search Methods
  - *How do they work?*
  - *Which search method should I use?*
- Dimensionality
  - *What is it?*
  - *Why is it important?*

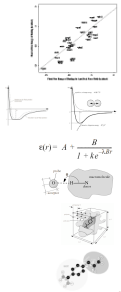
Scoring Function in AutoDock 4: Motivation

- To improve scoring function
  - improved hydrogen bonding
  - new desolvation energy term & internal desolvation energy
  - larger training set and new weights
- To permit protein sidechain, loop or domain flexibility (new DPF keyword, "flexres")
  - treats protein's moving atoms as part of the non-translating, non-reorienting part of the torsion tree
- To simulate the unbound state of the ligand & protein
  - *extended*, *compact* and *crystallographic* ligand conformations

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



AutoDock 4 Scoring Function Terms



$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta g_{tors}$$

- $\Delta G_{vdW} = \Delta G_{vdW}$
  12-6 Lennard-Jones potential (with 0.5 Å smoothing)
- ΔG_{elec}
  with Solmajer & Mehler distance-dependent dielectric
- ΔG_{hbond}
  12-10 H-bonding Potential with Goodford Directionality
- ΔG_{desolv}
  Charge-dependent variant of Stouten Pairwise Atomic Solvation Parameters
- ΔG_{tors}
  Number of rotatable bonds

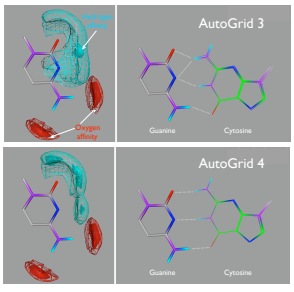
http://autodock.scripps.edu/science/equations
http://autodock.scripps.edu/science/autodock-4-desolvation-free-energy/

Pairwise terms in AutoDock 4

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

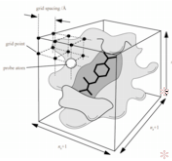
- Desolvation includes terms for all atom types
  - Favorable term for C, *A* (*aliphatic* and *aromatic* carbons)
  - Unfavorable term for O, N
  - Proportional to the absolute value of the charge on the atom
  - Computes the intramolecular desolvation energy for moving atoms
- Calibrated with 188 complexes from LPDB, K_i's from PDB-Bind
  Standard error (in Kcal/mol):
  - 2.62 (*extended*)
  - 2.72 (*compact*)
  - 2.52 (*bound*)
  - 2.63 (*AutoDock 3, bound*)
- Improved H-bond directionality

Improved H-bond Directionality



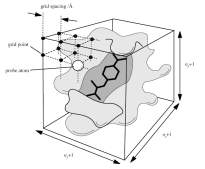
Huey, Goodsell, Morris, and Olson (2004) *Lett. Drug Des. & Disc.* **1**: 178-183

Why Use Grid Maps?



- *Saves time:*
  - Pre-computing the interactions on a grid is typically 100 times faster than traditional Molecular Mechanics methods
  - $O(N^2)$ calculation becomes $O(N)$
- AutoDock uses *trilinear interpolation*
  - to compute the score of a candidate docked ligand conformation
- 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*

Setting up the AutoGrid Box



- Macromolecule atoms in the rigid part
- Center:
  - center of ligand;
  - center of macromolecule;
  - a picked atom; **or**
  - typed-in x-, y- and z-coordinates.
- Grid point spacing:
  - default is 0.375Å (from 0.2Å to 1.0Å:).
- Number of grid points in each dimension:
  - only give *even numbers* (from 2 × 2 × 2 to 126 × 126 × 126).
  - AutoGrid adds one point to each dimension.
- Grid Maps depend on the orientation of the macromolecule.
- Make sure all the flexible parts of the macromolecule are inside the grid

To make a 'molecule' PDB file to show where the grid box is, use the script 'makebox':

- `% makebox mol.gpf > mol.gpf.box.pdb`

Relaxed Complex Method

Lin, J. H., Perryman, A. L., Schames, J. R., and McCammon, J. A. (2002). "**Computational drug design accommodating receptor flexibility: The relaxed complex scheme.**" *Journal of the American Chemical Society*, **124**: 5632-5633.

McCammon, J. (2005). "**Target flexibility in molecular recognition.**" *Biochimica et Biophysica Acta*, **1754**: 221-224.

Perryman, A. L. & McCammon, J. A. (2002). **AutoDocking dimolecoldes to the HIV-1 integrase core domain: Exploring possible binding sites for viral and genomic DNA.** *J Med Chem*, **45**: 3624-3627.

Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., and McCammon, J.A. (2004). "**Discovery of a novel binding trench in HIV integrase.**" *J Med Chem*, **47**(8): p. 1879-81.

Docking of the gCITEP inhibitor to snapshots of a 2 ns HIV-1 integrase MD trajectory indicated a previously uncharacterized trench adjacent to the active site that intermittently opens. Further docking studies of novel ligands with the potential to bind to both regions showed greater selective affinity when able to bind to the trench. Our ranking of ligands is open to experimental testing, and our approach suggests a new target for HIV-1 therapeutics.

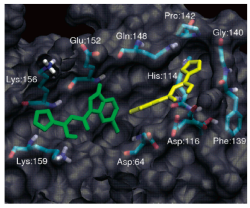


Figure 1. The two predominant docking conformations of 5CITEP to an open MD snapshot of integrase. The ligand in green shows 5CITEP in the orientation similar to the crystal structure of the complex. The ligand in yellow shows 5CITEP in its "flipped" orientation. Residues lining both ligand positions are highlighted.

Spectrum of Search: Breadth and Level-of-Detail

Search Breadth

- Local
  - Molecular Mechanics (MM)
- Intermediate
  - Monte Carlo Simulated Annealing (MC SA)
  - Brownian Dynamics
  - Molecular Dynamics (MD)
- Global
  - Docking

Level-of-Detail

- Atom types
- Bond stretching
- Bond-angle bending
- Rotational barrier potentials
- Implicit solvation
- Polarizability
- What's rigid and what's flexible?

Two Kinds of Search

Systematic

- Exhaustive
- Deterministic
- Outcome is dependent on granularity of sampling
- Feasible only for low-dimensional problems
- *e.g.* DOT (6D)

Stochastic

- Random
- Outcome varies
- Must repeat the search to improve chances of success
- Feasible for bigger problems
- *e.g.* AutoDock

Stochastic Search Methods

- Simulated Annealing (SA)*
- Evolutionary Algorithms (EA)
  - Genetic Algorithm (GA)*
- Others
  - Tabu Search (TS)
  - Particle Swarm Optimisation (PSO)
- Hybrid Global-Local Search Methods
  - Lamarckian GA (LGA)*

**Supported in AutoDock*

AutoDock has a Variety of Search Methods

- Global search algorithms:
  - Simulated Annealing (Goodsell *et al.* 1990)
  - Distributed SA (Morris *et al.* 1996)
  - 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)

How Simulated Annealing Works...

$$P(\Delta E) = e^{\left(-\frac{\Delta E}{k_B T}\right)}$$

the Metropolis criterion

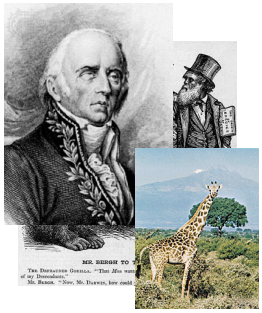
- Ligand starts at a random (or user-specified) position/orientation/conformation ('state')
- Constant-temperature annealing cycle:
  - Ligand's state undergoes a random change.
  - Compare the energy of the new position with that of the last position; if it is:
    - *lower*, the move is 'accepted';
    - *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, $0.85 < g < 1$
  - $T_i = g\,T_{i-1}$
- Rinse and repeat.
- Stops at the maximum number of cycles.

How a Genetic Algorithm Works...

- 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

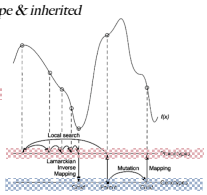
Lamarck

- Jean-Baptiste-Pierre-Antoinede Monet, Chevalier de Lamarck
- pioneer French biologist who is best known for his idea that acquired traits are inheritable, an idea known as Lamarckism, which is controverted by Darwinian theory.



How a Lamarckian GA works

- **Lamarckian**:
  - *phenotypic adaptations of an individual to its environment can be mapped to its genotype & inherited by its offspring.*
- **Phenotype** - Atomic coordinates
- **Genotype** - State variables
- (1) Local search (LS) modifies the *phenotype*
- (2) Inverse map *phenotype* to the *genotype*
  - Solis and Wets local search
  - advantage that it does not require gradient information in order to proceed
- **Rik Belew** (UCSD) & **William Hart** (Sandia).



Important Search Parameters

Simulated Annealing

- Initial temperature (K)
  - `rt0 61600`
- Temperature reduction factor (K^{-1} cycle)
  - `rtrf 0.95`
- Termination criteria:
  - accepted moves
    - `accs 25000`
  - rejected moves
    - `rejs 25000`
  - annealing cycles
    - `cycles 50`

Genetic Algorithm & Lamarckian GA

- Population size
  - `ga_pop_size 300`
- Crossover rate
  - `ga_crossover_rate 0.8`
- Mutation rate
  - `ga_mutation_rate 0.02`
- Solis & Wets local search (LGA only)
  - `sw_max_its 300`
- Termination criteria:
  - `ga_num_evals 250000 # short`
  - `ga_num_evals 2500000 # medium`
  - `ga_num_evals 25000000 # long`

  `ga_num_generations 27000`

Dimensionality of Molecular Docking

- **Degrees of Freedom (DOF)**
- **Position** / Translation (3)
  - *x,y,z*
- **Orientation** / Quaternion (3)
  - *qx, qy, qz, qw (normalized in 4D)*
- **Rotatable Bonds** / Torsions (*n*)
  - $\tau_1, \tau_2, \ldots \tau_n$
- **Dimensionality**, $D = 3 + 3 + n$

Multidimensional Treasure Hunt...

Dimensions	Landscape	Divide into 2	Treasure	Chances?
1				1/2
2				1/4
3				1/8

Sampling Hyperspace

- Say we are hunting in *D*-dimensional hyperspace...
- We want to evaluate each of the *D* dimensions *N* times.
- The number of "evals" needed, *n*, is: $n = N^D$
  - $\therefore N = n^{1/D}$
- For example, if $n = 10^6$ and...
  - $D=6$, $N = (10^6)^{1/6} = 10$ evaluations per dimension
  - $D=36$, $N = (10^6)^{1/36} = \sim 1.5$ evaluations per dimension
- Clearly, the more dimensions, the tougher it gets.

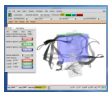
Next, AutoDock...

- Now for some specifics about AutoDock...

- More information can be found in the **User Guide**!

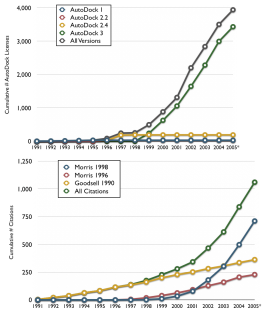
AutoDock / ADT

AutoDock & AutoGrid	**ADT**
1990	2000
Number crunching	Visualizing, set-up
Command-line. awk, shell & Python scripts. Text editors	Graphical User Interface. PMV ∴ Python GUI-less, self-logging & rescriptable
C & C++, compiled	Python, interpreted

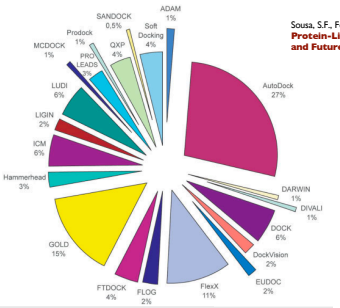


Community (1991 - mid 2005)

- AutoDock licenses
- Papers citing AutoDock (source: Science Citation Index Expanded)

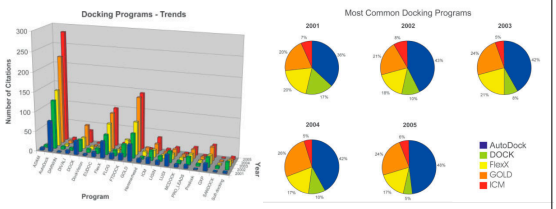


Number of Citations for Docking Programs —ISI Web of Science (2005)



Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2006) **Protein-Ligand Docking: Current Status and Future Challenges** *Proteins*, **65**:15-26

Trends in Citations of Docking Programs —ISI Web of Science (2005)



Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2006) **Protein-Ligand Docking: Current Status and Future Challenges** *Proteins*, **65**:15-26

Practical Considerations

- What problem does AutoDock solve?
  - *Flexible* ligands (4.0 *flexible* protein).
- 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 8-32 torsions.
- When is AutoDock not suitable?
  - No 3D-structures are available;
  - Modelled structure of poor quality;
  - Too many (32 torsions, 2048 atoms, 22 atom types);
  - Target protein too flexible.

Using AutoDock: Step-by-Step

- Set up ligand PDBQT—using ADT's "Ligand" menu
- *OPTIONAL:* Set up flexible receptor PDBQT—using ADT's "Flexible Residues" menu
- Set up macromolecule & grid maps—using ADT's "Grid" menu
- Pre-compute AutoGrid maps for all atom types in your set of ligands—using "`autogrid4`"
- Perform dockings of ligand to target—using "`autodock4`", and in parallel if possible.
- Visualize AutoDock results—using ADT's "Analyze" menu
- Cluster dockings—using "analysis" DPF command in "autodock4" or ADT's "Analyze" menu for parallel docking results.

AutoDock 4 File Formats

Prepare the Following Input Files

- Ligand PDBQT file
- Rigid Macromolecule PDBQT file
- Flexible Macromolecule PDBQT file ("Flexres")
- AutoGrid Parameter File (GPF)
  - GPF depends on atom types in:
    - Ligand PDBQT file
    - *Optional* flexible residue PDBQT files)
- AutoDock Parameter File (DPF)

Run AutoGrid 4

- Macromolecule PDBQT + GPF → Grid Maps, GLG

Run AutoDock 4

- Grid Maps + Ligand PDBQT + [Flexres PDBQT +] DPF → DLG (dockings & clustering)

Run ADT to Analyze DLG

Things you need to do before using AutoDock 4

Ligand:

- Add all hydrogens, compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- Ensure total charge corresponds to tautomeric state
- Choose torsion tree root & rotatable bonds

Macromolecule:

- Add all hydrogens, compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- Assign Stouten atomic solvation parameters
- Optionally, create a flexible residues PDBQT in addition to the rigid PDBQT file
- Compute AutoGrid maps

Preparing Ligands and Receptors

- AutoDock uses 'United Atom' model
  - Reduces number of atoms, speeds up docking
- Need to:
  - Add polar Hs. Remove non-polar Hs.
    - Both Ligand & Macromolecule
  - Replace missing atoms (disorder).
  - Fix hydrogens at chain breaks.
- Need to consider pH:
  - Acidic & Basic residues, Histidines.
  - http://molprobity.biochem.duke.edu/
- Other molecules in receptor:
  - Waters, Cofactors, Metal ions.
- Molecular Modelling elsewhere.

Atom Types in AutoDock 4

- One-letter or two-letter atom type codes
- More atom types than AD3:
  - **22**
- Same atom types in both ligand and receptor
- http://autodock.scripps.edu/wiki/NewFeatures
- http://autodock.scripps.edu/faqs-help/faq/how-do-i-add-new-atom-types-to-autodock-4
- http://autodock.scripps.edu/faqs-help/faq/where-do-i-set-the-autodock-4-force-field-parameters

Partial Atomic Charges are required for both Ligand and Receptor

- Partial Atomic Charges:
  - Peptides & Proteins, DNA & RNA
    - Gasteiger (PEOE) - *AD4 Force Field*
  - Organic compounds, Cofactors
    - Gasteiger (PEOE) - *AD4 Force Field*,
    - MOPAC (MNDO, AM1, PM3),
    - Gaussian (6-31G*).
- Integer total charge per residue.
- Non-polar hydrogens:
  - Always merge

Carbon Atoms can be either Aliphatic or Aromatic Atom Types

- Solvation Free Energy
  - Based on a partial-charge-dependent variant of Stouten method.
  - Treats aliphatic ('C') and aromatic ('A') carbons differently.
- Need to rename ligand aromatic 'C' to 'A'.
- ADT determines if ligand is a peptide:
  - If so, uses a look-up dictionary.
  - If not, inspects geometry of 'C's in rings. Renames 'C' to 'A' if flat enough.
  - Can adjust 'planarity' criterion (15° detects more rings than default 7.5°).

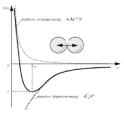
Defining Ligand Flexibility

- Set Root of Torsion Tree:
  - By interactively picking, *or*
  - Automatically.
    - Smallest 'largest sub-tree'.
- Interactively Pick Rotatable Bonds:
  - No 'leaves';
  - No bonds in rings;
  - Can freeze:
    - Peptide/amide/selected/all;
  - Can set the number of active torsions that move either the most or the fewest atoms

Setting Up Your Environment

- At TSRI:
  - Modify .cshrc
    - Change PATH & stacksize:
    - `setenv PATH (/mgl/prog/$archosv/bin:/tsri/python:$path)`
    - `% limit stacksize unlimited`
- ADT Tutorial, every time you open a Shell or Terminal, type:
  - `% source /tsri/python/share/bin/initadtcsh`
- To start AutoDockTools, type:
  - `% cd tutorial`
  - `% adt1`
- Web
  - `http://autodock.scripps.edu`
  - `http://mgltools.scripps.edu`

Choose the Docking Algorithm



- SA.dpf — Simulated Annealing
- GA.dpf — Genetic Algorithm
- LS.dpf — Local Search
  - Solis-Wets (SW)
  - Pseudo Solis-Wets (pSW)
- GALS.dpf — Genetic Algorithm with Local Search, *i.e.* Lamarckian GA

Run AutoGrid

- Check: Enough disk space?
  - Maps are ASCII, but can be ~2-8MB !
- Start AutoGrid from the Shell:

```
% autogrid4 -p mol.gpf -l mol.glg &
% autogrid4 -p mol.gpf -l mol.glg ; autodock4 -p mol.dpf -l mol.dlg
```

- Follow the log file using:

```
% tail -f mol.glg
```

  - Type <Ctrl>-C to break out of the '`tail -f`' command
- Wait for "Successful Completion" before starting AutoDock

Run AutoDock

- Do a test docking, ~ 25,000 evals
- Do a full docking, if test is OK, ~ 250,000 to 50,000,000 evals
- From the Shell:
  - `% autodock4 –p yourFile.dpf –l yourFile.dlg &`
- Expected time? Size of docking log?
- Distributed computation
  - At TSRI, Linux Clusters
  - `% submit.py stem 20`
    `% recluster.py stem 20 during 3.5`

Analyzing AutoDock Results

- In ADT, you can:
  - Read & view a single DLG, or
  - Read & view many DLG results files in a single directory
  - Re-cluster docking results by conformation & view these
- Outside ADT, you can re-cluster several DLGs
  - Useful in distributed docking
    - `% recluster.py stem 20 [during|end] 3.5`

Viewing Conformational Clusters by RMSD

- List the RMSD tolerances
  - Separated by spaces
- Histogram of conformational clusters
  - Number in cluster *versus* lowest energy in that cluster
- Picking a cluster
  - makes a list of the conformations in that cluster;
  - set these to be the current sequence for states player.

Advanced Topics

- Stochastic search methods rely on random numbers
- Random Number Generator, RNG

Random number generator

- RNG needs a seed or seeds.
  - Different seeds lead to different sequences of random numbers
- SA and GA use different RNGs
  - SA needs 1 seed
  - GA & LGA need 2 seeds
- A seed can be:
  - A long integer, say "3141529"; or
  - "time" = number of seconds since 1970 Jan 1; or
  - "pid" = UNIX process ID of this job

Acknowledgments

- Ruth Huey
- William Lindstrom
- David S. Goodsell
- Michel Sanner
- Sophie Coon
- Daniel Stoffler
- Michael Pique
- Art J. Olson
- Rik Belew (UCSD)
- Bill Hart (Sandia)
- Scott Halliday
- Chris Rosin
- Max Chang
- Flavio Grynszpan (TSRI)
- Many patient ADT users