Using AutoDock 4 for Virtual Screening

Garrett M. Morris
William Lindstrom
Ruth Huey
Christoph Weber

Outline

- Introduction to Virtual Screening
  - Definition of Virtual Screening
  - Why use virtual screening?
  - HTS *vs.* VS
  - Different types of libraries
  - Comparison of libraries
  - NCI Diversity Set
  - SMILES
  - Small molecule structures
  - Converting 1D & 2D into 3D
  - Example: AICAR Transformylase
  - Single Docking *versus* Virtual Screening
- Hands-on Tutorial
- Introduction to TSRI Supercomputers

Virtual Screening

- *Definition of Virtual Screening:*

  Use of high-performance computing to analyze large databases of chemical compounds in order to identify possible drug candidates.

  W.P. Walters, M.T. Stahl and M.A. Murcko, "Virtual Screening-An Overview", *Drug Discovery Today*, **3**, 160-178 (1998).

- Virtual Screening is also known as:
  - High-Throughput Docking
  - High-Throughput Virtual Screening

Why Use Virtual Screening?

- VS is a computational filter
  - reduces the size of a chemical library to be screened experimentally, ~10^6 to ~10^3 —Saves time & money
- May improve likelihood of finding a good compound
  - as opposed to random screening
  - enhanced "hit rates"
- VS can:
  - perform analysis before an assay is established
  - evaluate virtual combinatorial libraries before synthesized
- In the "post-genomic" era, *many* new targets will be discovered…

HTS *versus* VS

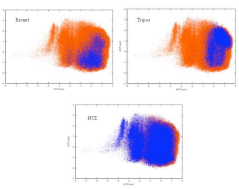
- High Throughput Screening (HTS):
  - Tests activity *in vitro.*
  - Assays are not infallible (false negatives).
  - Chemical synthesis & testing are expensive.
- Virtual Screening (VS):
  - Computes binding activity *in silico.*
  - VS is also known as "vHTS".
- HTS and VS are complementary:
  - Use VS to exclude compounds which are predicted not to bind, helping to "enrich" the library…
  - VS can also help to identify false-negatives in HTS

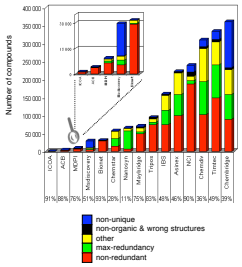
Different Types of Libraries

Which library you choose depends…

- Comprehensive (> ~500,000 compounds)
  - *search in the dark*
- Diversity-based to cover 'chemical space'
  - *efficient search in the dark*
- "Focused" or "Targeted" for *lead identification*
  - *e.g.* filtered by 2D or 3D pharmacophores
  - *search with a flashlight*
- "Focused" or "Targeted" for *lead optimization*
  - *focussing the spotlights*
- Combinatorial Libraries

Comparison of Libraries





Library metrics from "Preparation of a molecular database from a set of 2 million compounds for virtual screening applications: gathering, structural analysis and filtering", J-C Mozziconacci, *et al.*

www.univ-orleans.fr/SCIENCES/ICOA/eposter/eccc9/ECCC91.htm

NCI Diversity Set

~140,000 compounds

> 1.0 gram available

71,756 compounds

Diversity based on unique 3-point pharmacophores

1,990 compounds

How it was built…

Chem-X (Oxford Molecular Group) was used.

(1) Defined **3-point pharmacophores** based on hydrogen bond acceptor, hydrogen bond donor, positive charge, negative charge, aromatic, hydrophobic, acid, base and defined distance intervals.

(2) Generated a set of **~1,000,000 pharmacophores for all acceptable conformations of each structure**.

(3) A diverse subset was built up by comparing all pharmacophores for each acceptable conformation, **adding the structure to the set if it had 5 or more new pharmacophores**.

SMILES

- *Simplified Molecular Input Line Entry Specification*
- A string of letters, numbers and other characters that specify the atoms, their connectivity, bond orders, & chirality
- http://www.daylight.com/smiles/f_smiles.html

Depiction	SMILES	Name
H_2O	O	water
CH_4	C	methane
	CC(=O)O	acetic acid
	C1CCCCC1	cyclohexane
	c1ccccc1	benzene
	c1ccccc1[N+](=O)[O-]	nitrobenzene

Small Molecule Structures

- Sources of Small Molecule Structures:
  - *CCDC's Cambridge Structural Database*
    - the world repository of small molecule crystal structures
    - http://www.ccdc.cam.ac.uk/products/csd/
  - *NCI, National Cancer Institute*
    - http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html
  - *PubChem*
    - http://pubchem.ncbi.nlm.nih.gov
  - *ZINC, ZINC Is Not Commercial*
    - http://zinc.docking.org
- More information:
  - *Molecular Docking Web*
    - http://mgl.scripps.edu/people/gmm/index.html#SmallMolecules

Converting 1D & 2D to 3D

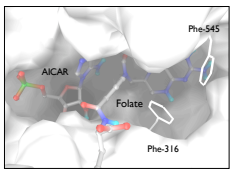
- Corina
  - 1000 structures for free
  - Specify input as SMILES or sketch using JME
  - http://www.molecular-networks.com/online_demos/corina_demo.html
- Dundee PRODRG2 Server
  - Specify input as PDB, MDL MOL or sketch using JME; returns PDBQ format
  - http://davapc1.bioch.dundee.ac.uk/programs/prodrg/
  - A. W. Schuettelkopf and D. M. F. van Aalten (2004). PRODRG - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica* **D60**, 1355-1363
- ZINC
  - Specify input as SMILES
  - http://zinc.docking.org/
  - Irwin and Shoichet (2005) *J. Chem. Inf. Model.* **45**(1), 177-82

Strategy

- Find the 3D structure and inhibition constant K_i of a complex of your desired target with an inhibitor ('*positive control*')
- Perform a "re-docking" on your positive control to verify your input files and parameters are reasonable.
- Note the predicted binding free energy (BFE) from AutoDock
- This energy, plus the standard deviation in the predicted BFE of the AutoDock force field, ~2.6 kcal/mol, forms the threshold above which we will be looking for "hits", molecules with better BFE than the positive control's BFE.
- Add the positive control inhibitor to your library before virtual screening

e.g. AICAR Transformylase

- AutoDock 3 was used to screen the NCI Diversity Set
- 1990 compounds, against AICAR transformylase,
  - an enzyme involved in the purine biosynthetic pathway
- AutoDock Parameters used:
  5 million evals per run
  100 runs per compound
- Took about 2 weeks using 32 nodes of The Scripps Research Institute's "redfish" Linux cluster (*circa* 2003)



Well-defined binding pocket of AICAR Transformylase

Chenlong Li

VS & Kinetic Inhibition Results

- *In silico:*
  44 top compounds, $E_{binding}$ <= -13.0 Kcal/mol
- *In vitro:*
  - 10 are insoluble in water
  - 18 precipitate in buffer solution
  - 8 out of 16 soluble compounds bind
    (50% success)

Li, C., Xu, L., Wolan, D.W., Wilson, I.A., and Olson, A.J. (2004) **Virtual screening of human 5-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of AutoDock to identify novel nonfolate inhibitors**. *J Med Chem*, **47**(27): 6681-90.

Tyrosine Phophatase 1B (PTP1B)

- HTS (*in vitro)* of 400,000 compounds
  - 300 hits with IC_{50}< 300μM
  - 85 validated hits with IC_{50}< 100 μM
  - 0.021% hit rate (= 85 / 400,000)
  - many violate Lipinski rules
- VS (*in silico)* of 235,000 compounds (DOCK)
  - 365 high-scoring molecules
  - 127 validated hits with IC_{50}< 100 μM
  - 34.8% hit rate (= 127 / 365)
  - hits are more drug-like

T.N. Doman *et al.*, (2002) *J. Med.Chem.* **45**: 2213-2221

VS of DNA minor groove binders

- Evans D.A. & Neidle S. (2006) **Virtual screening of DNA minor groove binders** *J.Med.Chem.* **49**(14): 4232-8.
- Compared DOCK 5.1.1 and AutoDock 3.0.5 for docking libraries of compounds to DNA minor grooves. (109d, 127d, 129d, 166d, 1d30, 1d64, 1fmq, 1fms, 1ftd, 1lex, 1m6f, 1prp, 1qv4, 1qv8,1vzk, 227d, 289d, 298d, 2dbe, 302d, 311d, 328d, 360d, 442d, 443d, 447d, 448d, 453d)
- Success in finding the crystal structure to within 2.0 Å RMSD:
  - AutoDock: 57%
  - DOCK: 40%
- AutoDock also gave the best enrichment of known binding compounds in a screen of 9216 randomly chosen molecules from the ZINC database, with an enrichment value SE(f=1%) = 86%; this could improve if the ZINC mol2 files were available with AMS-HEX charges.
- Showed that accurate prediction of the docked conformation is correlated with enrichment.
- Post-docking scoring in DOCK using the GBSA scoring function in DOCK did not improve enrichment with DOCK over the standard DOCK energy score (except at low f).
- Using the sampling parameters for DOCK and AutoDock that produced maximal enrichment in their virtual screening comparisons, AutoDock also performed faster (8s on average for AutoDock, 40s on average for DOCK, on a 3.0 GHz Intel x86-64).

VS of DNA minor groove binders (cont-d)

- Evans & Neidle used scripts in VMD to compute the RMSD values for only the heavy atoms, for both DOCK and AutoDock dockings. Only the best-scoring docked conformation was considered.
- For AutoDock, they used desolvation parameters for phosphorus based on a recent study that used AutoDock to examine RNA-ligand interactions
  - Detering *et al.* (2004) *J.Med.Chem.*, **47**:4188
- They also commented that,
  - "It is interesting that the AutoDock scoring function, which was parametrized with experimental protein-ligand inhibition constants, performs better than the DOCK scoring function, which is more closely matched to the original AMBER94 force field. It would thus appear that the parametrization is transferable from proteins to DNA."
- They also compared a variety of charge models in AutoDock. They concluded that AMS-HEX charges (*i.e.* using AMSOL with the AM1-CM2 Hamiltonian for non-polar organic solvent) gave the best performance for accuracy of x-ray structural prediction.

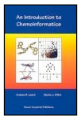
Single Docking *v.* Library Screen

- Use GUI
- Data in one directory
- Prepare input files:
  - Ligand PDBQT
  - Receptor PDBQT
  - GPF
  - DPF
- One AutoGrid calculation
- One AutoDock calculation
- Analyze Results

- Use scripts
- Data in tree structure
- Prepare input files:
  - Library of Ligand PDBQT files
  - Receptor PDBQT
  - GPF
  - Library of DPFs
- One AutoGrid calculation
- Submit AutoDock jobs to cluster
- Rank Results; Analyze best

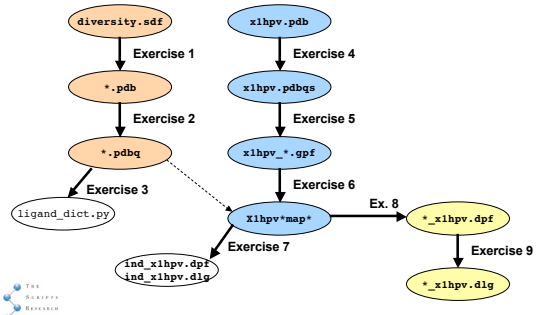
Recommended Reading

- Leach, A. R., Gillet, V. J.
  "An Introduction to Chemoinformatics",
  Kluwer Academic Press, 2003.

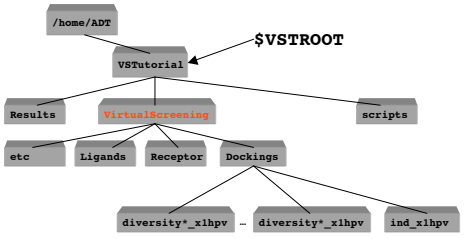


- Gasteiger, J. (ed), Engel, T. (ed)
  "Chemoinformatics: A Textbook",
  John Wiley & Sons, 2003.

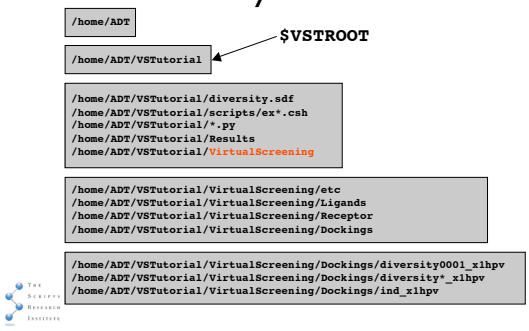
Virtual Screening Tutorial Map



Virtual Screening Tutorial Directory Structure



Virtual Screening Tutorial Directory Structure



General Comments

- use `pwd` and `ls` often

  If you are unfamiliar with the Unix command line and/or navigating around a hierarchical file system, use the `pwd` (print name of current/working directory) and `ls` (list directory contents) shell commands as much as you need to stay oriented in the file system. It's always helpful to draw a quick picture.

- use `man` often

  Use the `man` command copiously. Unix has a very useful on-line manual that you can read with the `man` command. For example, if you can't remember how to use the `ls` command to list a directory contents with file modification dates, type `man ls`. This will display the on-line manual page which describes the `ls` command and allow you quickly learn how do it. `man -k` is a useful option when you can't remember the name of the command you want to read about. (Look up `man -k` in the on-line manual by typing `man man`).

Unix Shell Commands Used

- **ls**
- **pwd**
- **cd**
- **mkdir**
- **..**
- **man**
- **setenv**
- **echo**
- **foreach**
- **>**
- **|**
- **cp**
- **ln -s**
- **cat**
- **more**
- **head**
- **tail**
- **wc**
- **grep**
- **sort**
- **awk**
- **sed**
- **vi or emacs**