CHEMINFORMATICS: LIBRARY & COMPOUND SELECTIONS

V. Feher & Y. Su
NCBR CADD Workshop Aug. 2, 2012
Top 10 Best Practices: Target-Focused Screening Libraries

1. Get familiar with your binding site
2. Scour the literature for known substrates, transition state (analogs), small molecule binders, natural products
3. Know your compound sources - NCI, Commercial, Natural Product, PubChem
4. Pre-filter your compound sources – REOS
5. Filter Known Properties – Flag PK Properties
6. Use multiple VS Methods - 2D & 3D similarity searches, docking and pharmacophore
7. Keep chemists in mind – diverse, not overly adorned scaffolds and SAR
8. Characterize hits with MS (NMR) – did you assay what you thought you bought
9. Reorder, retest and secondary assay
10. Analog searching for follow-up
What is Cheminformatics?

Cheminformatics is the process of amassing information about small molecules and using this information to make “better decisions faster” in the area of drug lead identification and optimization.”


Figure = NBCR Pipeline; http://www2.nbcrcr.net/wordpress2/?page_id=1175

Physical & Calculated properties
- MW
- # rotatable bonds
- octanol/water partition coefficient

Experimentally derived properties
- solubility
- microsome stability
- cell permeability
- toxicity

Chemistry knowledge
- tautomers
- ionizable groups
- chemical stability
How does Cheminformatics fit in with the CADD Pipeline?

Where ever we are dealing with small molecules.....
Top 10 Best Practices: Target-Focused Screening Libraries

1. Get familiar with your binding site
2. Scour the literature for known substrates, small molecule binders, natural products
3. Know your compound sources - NCI, Commercial, Natural Product, PubChem
4. Pre-filter your compound sources – REOS
5. Filter Known Properties – Flag PK Properties
6. Use multiple VS Methods - 2D & 3D similarity searches, docking and pharmacophore
7. Keep chemists in mind – diverse, not overly adorned scaffolds and SAR
8. Characterize hits with MS – did you assay what you thought you bought
9. Reorder, retest and secondary assay
10. Analog searching for follow-up
FILTERING
Why is filtering important?

1. Why dock & analyze compounds you’ll never test?

2. PAINS = “Pan-Assay Interference Compounds”

Problematic scaffolds – has cost their Institute time and $$

![Chemical structures](image)


See Table 11 in this reference. – It has functional groups that they have trouble with & see in the literature.
The Compound Library to Docking Pipeline

1. Rapid Elimination of Swill (REOS)\textsuperscript{1,2}

Filtering Tools:
- Open Eye FILTER
- Accelrys Pipeline Pilot
- CCG MOE “wash” & db tools
- Schrodinger Canvas/Qikprop

The Compound Library to Docking Pipeline

2. 2D Library to 3D Docking Library

2D to 3D conversion → valence check → tautomers → docking format → 3D

Docking Prep Tools:
- Open Eye OMEGA
- Accelrys Pipeline Pilot
- CCG MOE db tools
- Schrodinger LigPrep
- MGL Tools
Basic “Washing” -

- **Removing Salts & Unwanted Elements**
  - Filter out cationic atoms: Ca$^{2+}$, Na$^+$, etc
  - Filter out metals: Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd
  - Often the salt “filter” = keeping the largest molecule in the sdf entry

- ALLOWED_ELEMENTS H, C, N, O, F, P, S, Cl, Br, I
Basic “Washing” -

- **Proper Atom Types**
  - Filter adds hydrogens and checks if O, N, C valences make sense – sometimes sdf have corrupt entries
  - Checks formal charge
  - If it doesn’t make sense – it “fails” the compound

- **Ionization**
  - Filter uses a rule based method to add Hs & charge to particular groups for property calculations that occur later; it assumes pH = 7.4
Filter out: Reactives

Covalent Inhibitors

Electrophilic “suicide inhibitors” (serine, threonine, and cysteine proteases):

- α-Haloketones
- Boronic acids
- Aldehydes
- TFMKs
- 1,2-Dicarbonyls
- Saccharinoids

Reactive to protein functional groups

- Epoxides
- Aziridines
- Thioesters
- Sulfonate esters
- Phosphonate esters

Open Eye Filter Manual – has many examples
Filter out: Synthesis Intermediates, Chelators, other unwanteds

Blocking groups

Chelators – bind metals

Phosphates

Open Eye Filter Manual – has many examples
Filter out: Dyes

These may give false positives if you are using a photometric assay. Also these are of little pharmacological interest. Usually highly conjugated & flat aryl compounds. –NO₂ and –SO₃ groups add solubility to these flat conjugated systems.
Filter out: Aggregators & Promiscous Binders

Promiscous binders = compounds that give “positives” in many target assays.

One mechanism for many of these is aggregation in solution. They are typically greasy & flat.

Another mechanism is general protein unfolding.

It is worth learning what these look like.


http://shoichetlab.compbio.ucsf.edu/aggregators.php
More examples of promiscuous binders

Filter out: Unwanted Functional Groups

These are removing cases where there are too many of a type of functional group.
You can CUSTOMIZE the rules depending on the goals for your library (OE rules).

RULE 2 alkyne
RULE 4 aniline
RULE 4 aryl_halide
RULE 2 carbamate
RULE 3 ester
RULE 5 ether
RULE 1 hydrazone
RULE 0 nonacylhydrazone
RULE 1 hydroxylamine
RULE 2 nitrile
RULE 2 sulfide
RULE 2 sulfone
RULE 2 sulfoxide
RULE 1 thiourea
RULE 1 thioamide
RULE 1 thiol
RULE 2 urea

RULE 6 alcohol
RULE 4 alkene
RULE 4 amide
RULE 4 amino_acid
RULE 2 amine
RULE 4 primary_amine
RULE 4 secondary_amine
RULE 4 tertiary_amine
RULE 2 carboxylic_acid
RULE 6 halide
RULE 0 iodine
RULE 2 ketone
RULE 4 phenol
RULE 1 imine
RULE 1 methyl_ketone
RULE 1 alkylaniline
RULE 1 oxime
RULE 0 isothiocyanate
RULE 0 isocyanate
RULE 3 lactone
RULE 3 lactam
RULE 1 thioester
RULE 1 carbonate
RULE 0 carbamic_acid
RULE 1 thiocarbamate
RULE 0 triazine
RULE 1 malonic
RULE 4 sulfonamide
RULE 1 sulfonylurea
OE’s FILTER: Physical Property Calculations & Default Cutoff’s

- Molecular Weight  \(130 \leq MW \leq 781\)
- Heavy Atom Count  \(9 \leq HVY \leq 55\)
- Carbon Count, Hetero Count, \(3 \leq \#C \leq 41\), \(1 \leq HETERO \leq 14\)
- Hetero/Carbon Ratio \(0.4 \leq HET/C \leq 4.0\)
- Chiral Centers Count \(0 \leq Chiral \leq 21\)
- H-bond Acceptors : \(0 \leq HBA \leq 13\)
- Halide Fraction: \(0 \leq \text{Halide Fraction} \leq 0.66\)
- MW of halide/MW of cmpd
- Formal Count \(0 \leq \# \text{Formal Charge} \leq 4\)
- Formal Sum: \(-2 \leq \text{Sum Formal Charge} \leq 2\)
- Unbranched chain: \# unbranched connected non-ring atoms
- Connected, non-ring \(0 \leq \text{keep} \leq 19\)
- Ring systems: \(0 \leq \text{keep} \leq 5\)
  - \# of contiguous rings
- Ring size \(0 \leq \text{keep} \leq 20\)
- Rotor Count : \(0 \leq \text{keep} \leq 16\)
  - \# of rotatable bonds
- Rigid Count: \(0 \leq \text{keep} \leq 55\)
  - \# of non-rotatable bonds

Note: some cutoffs change when more rules are added, eg: Lipinski, Verber, etc.

What’s most important?

I typically focus on:
cLogP
MW
# rotatable bonds
# halides: not too many –Br, -I
# rings
Calculated Properties: \( c\text{LogP} \) and \( c\text{LogD} \)

These are partition coefficients for small molecules between octanol and water.

There is a correlation between these values and a molecule’s solubility and ability to cross membranes.

\[
\log P_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{octanol}}}{[\text{solute}]_{\text{un–ionized}}_{\text{water}}} \right) 
\]

Note: \( c\text{LogP} \) ignores charges on molecules – it is invalid for compounds with charge.

\[
\log D_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{ionized}}_{\text{octanol}}}{[\text{solute}]_{\text{ionized}}_{\text{water}} + [\text{solute}]_{\text{neutral}}_{\text{water}}} \right)
\]

This is pH dependent and is usually reported for pH = 7.4.

Different \( c\text{LogP} \) calculations can vary up to 1 log unit. It’s good to remember that when applying Lipinski guidelines.
TPSA = topological polar surface area

Some TPSA programs only count O & N atoms

Of all the calculated properties – this is one of the best correlations

Practical: Using OpenEye’s FILTER

Input = sdf, smiles, smarts

In 2D format

```
bash-3.2$ /Applications/OpenEye/bin/filter --help
Filter 2.1.1, 20100816
  DEChem version 1.7.2, 20100810
  Platform: osx-10.5-g++4.0-x86
  OpenEye Scientific Software, Inc.

  Licensed for the exclusive use of The laboratory of Rommie Amaro.
  Licensed for use only in University of California, San Diego, CA.

Help functions:
  filter --help simple  : Get a list of simple parameters
  filter --help all    : Get a complete list of parameters
  filter --help defaults : List the defaults for all parameters
  filter --help <parameter> : Get detailed help on a parameter
  filter --help html   : Create an html help file for this program
```
Practical: Using OpenEye’s FILTER

```
bash-3.2]$ /Applications/OpenEye/bin/filter --help simple
Filter 2.1.1, 20100810
  OEChef version 1.7.2, 20100810
  Platform: osx-10.5-g++4.0-x86
  OpenEye Scientific Software, Inc.

Licensed for the exclusive use of The laboratory of Rommie Amaro.
Licensed for use only in University of California, San Diego, CA.

Simple parameter list
  filter
    -dots : Write a dot to the terminal for every 500 cpds processed
    -fail : Write failed molecules to file
    -filter : Filter to use in filtering
    -flagTableFailures : Flag failure values in table form with an asterix
    -in : Input filename
    -info : Write to specified info file (override -prefix)
    -interval : Update info file every N molecules
    -log : Write to specified log file (override -prefix)
    -newrule : File listing additional filter rules and SMARTS
    -normalize : SMIRKS file for normalization of molecules
    -out : Output filename
    -param : Filter control parameter file
    -pkanorn : Apply a neutral pH model
    -prefix : Prefix for generic output files
    -salt : Molecule file of salts to remove instead of calling strip salts
    -sdtag : Add all filter parameters as SD Tags
    -select : Smarts pattern to select for groups
    -table : Write complete output in table form
    -typecheck : Check for broken molecules (impossible valence states)
    -unique : Keep only unique molecules & load known uniques from file
```

```
/Applications/OpenEye/bin/filter --in file.sdf --filter lead --prefix clean --fail failed --out clean_out
```
Practical: OpenEye’s FILTER output

- clean_out = compounds for VS
- clean.info file
  
  ******************** Progress Update ********************
  Number of Molecules Processed = 42
  Number of Molecules Passed = 36
  Number of Molecules Filtered = 6
  Number of Warnings = 2
  Number of Errors = 0
  Average molecules/sec = 439.427094
  Elapsed Time = 0.095587
  ********************

- clean.log
  
  - Top of the logfile lists the parameter settings used
  - End of the log file has each compound listed + pass/fail, failure reason

```
PHARMACOPIA  true
MIN_ABS 0.500000
AGGREGATORS true
PRED_AGG  true

**************

CCC(=O)Nc1cc(cccc1NCC[NH+](CC2)C)C(=O)OC ,Pass
Cc1ccc(c1NC(=O)C(=O)Nc2cccs2)Cl ,Pass
CC0c1cc(cccc10CC(=O)N2CCOC2)C#N ,Pass
Cc1c(c2nc(c(n2n1)N)C#N)c3ccc(c(c3)OC)OC ,Pass
CCOC(=O)c1cccc(c1)Nc(=O)Cn2nc(nn2)c3cccs3 ,Pass
CC1CCCC[NH+][1Cc2nc3ccccc3n2C ,Pass
C0c1c(cn(n1)Cc2ccccc2)[N+](=O)[0-] ,Maximum nitro(0) exceeded: 1
Cc1cc(cccc10CC(=O)NC)S(=O)(=O)N2CCc3c2cccs3 ,Pass
CC[NH+]1CCN(CC1)C(=O)CN(CC)S(=O)(=O)c2cc(cccc2Cl)Cl ,Pass
CCN(CC)C(=O)Nc1ccc(cc1)OC ,Pass
CCS(=O)(=O)c1ncc(c(n1)C(=O)Nc2cc(ccc2)OC)Cl ,Pass
```
Practical: Using OpenEye’s FILTER

- Default filters
  - lead
  - drug
  - blockbuster

The filters are text files – they can be found in the directory
/OpenEye/data/

```plaintext
# Copyright (C) 2000-2005, 2010 by OpenEye Scientific Software, Inc.
#
# This file defines the rules for filtering multi-structure files based on
# properties and substructure patterns.

MIN_MOLWT 150 "Minimum molecular weight"
MAX_MOLWT 440 "Maximum molecular weight"

MIN_NUM_HVY 10 "Minimum number of heavy atoms"
MAX_NUM_HVY 25 "Maximum number of heavy atoms"

MIN_RING_SYS 0 "Minimum number of ring systems"
MAX_RING_SYS 3 "Maximum number of ring systems"

MIN_RING_SIZE 0 "Minimum atoms in any ring system"
MAX_RING_SIZE 20 "Maximum atoms in any ring system"

MIN_CON_NON_RING 0 "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING 15 "Maximum number of connected non-ring atoms"

MIN_FCNGRP 0 "Minimum number of functional groups"
MAX_FCNGRP 12 "Maximum number of functional groups"

MIN_UNBRANCHED 0 "Minimum number of connected unbranched non-ring atoms"
MAX_UNBRANCHED 3 "Maximum number of connected unbranched non-ring atoms"

MIN_CARBONS 5 "Minimum number of carbons"
MAX_CARBONS 23 "Maximum number of carbons"

MIN_HETEROATOMS 2 "Minimum number of heteroatoms"
MAX_HETEROATOMS 12 "Maximum number of heteroatoms"
```
Filtering: Customizing FILTER

you can edit your own using the “lead” filter as a template or
add SMARTS strings to remove specific compounds in a separate text file & using “- newrule” in your command line

/Applications/OpenEye/bin/filter –in file.sdf –filter lead –newrule file.txt –out clean_out

SMARTS theory:

For easier conversion of smiles strings to SMARTS:

http://www.chemaxon.com/marvin/help/formats/smiles-doc.html#SMARTS
or
Open Eyes’ OMEGA
Top 10 Best Practices:
Target-Focused Screening Libraries

1. Get familiar with your binding site
2. Scour the literature for known substrates, small molecule binders, natural products
3. Know your compound sources - NCI, Commercial, Natural Product, PubChem
4. Pre-filter your compound sources – REOS
5. **Filter Known Properties – Flag PK Properties**
6. Use multiple VS Methods - 2D & 3D similarity searches, docking and pharmacophore
7. Keep chemists in mind – diverse, not overly adorned scaffolds and SAR
8. Characterize hits with MS – did you assay what you thought you bought
9. Reorder, retest and secondary assay
10. Analog searching for follow-up
Flagging vs. Removing

- Each user may have different screening goals
  - Flagging properties, dyes and/or potential aggregators may be useful to simply raise awareness of a potential compound hazard
    - eg. PAINS – includes azo dyes because they are looking for easy chemistry and cancer drugs with high toxicity tolerance but they remove many aggregators
    - If you are looking for compounds likely to pass the BBB, you may want to flag for low TPSA values to prioritize which hits to follow up on later

- Over-filtering may lead to low hit rate, but lack of filtering can lead to wasted resources & time
Example: Library for Neglected Diseases, Dundee Scotland

The Compound Library to Docking Pipeline

2. 2D Library to 3D Docking Library

Docking Prep Tools:
Open Eye OMEGA
Accelrys Pipeline Pilot
CCG MOE db tools
Schrodinger LigPrep
MGL Tools
Tautomerization

- Why is it important to VS?
  - Location of HBA and HBD
  - Structural conformation
  - Properties

Keto-enol tautomers

How does Cheminformatics fit in with the CADD Pipeline?

Where ever we are dealing with small molecules.....
Selecting Compounds: What to do about data overload!

- There are many ways to select your best docked compounds
  - Docking score – top 100(?)
  - Consensus score, other docking scores
  - Rescoring – MM-PBSA, MM-GBSA, TI, NN-Score

- You need to look at your selections – with large compound databases – perhaps multiple poses saved & RCS – this is more that can be reasonably reviewed visually
Compound Clustering

- Take the top scoring compounds (100 – 5000) and their docking scores – cluster them by chemical types (scaffolds, chemotypes)

- You can then look for scaffolds that give the best scores and view a few of them
  - Do they make reasonable interactions within the pocket?
  - Are the conformations of the compounds reasonable?
  - Are there particular functional groups on the chemotype skewing the score?
Scaffold Hunter

http://scaffoldhunter.sourceforge.net/index.html

Tripod: NIH Cheminformatics Scaffold – Activity Diagram

Another “freeware” option for scaffold binning

http://tripod.nih.gov/
Top 10 Best Practices:
Target-Focused Screening Libraries

1. Get familiar with your binding site
2. Scour the literature for known substrates, small molecule binders, natural products
3. Know your compound sources - NCI, Commercial, Natural Product, PubChem
4. Pre-filter your compound sources – REOS
5. Filter Known Properties – Flag PK Properties
6. Use multiple VS Methods - 2D & 3D similarity searches, docking and pharmacophore
7. Keep chemists in mind – diverse, not overly adorned scaffolds and SAR
8. Characterize hits with MS – did you assay what you thought you bought
9. Reorder, retest and secondary assay
10. Analog searching for follow-up
Scaffold Diagrams can assist with selecting sets for SAR

Example

If a scaffold makes sense in the binding pocket
- purchase the best scorer +
- a variety of functional groups

This gives you and the chemist and idea of what changes might work.

http://scaffoldhunter.sourceforge.net/index.html

Scaffold Hopping

An approach to discover novel compounds with different central core structures from the known leads/hits.

Taken from
Drug Discovery Today
Classification of scaffold-hopping approaches
(2012) 17, 310–324
LIBRARY SOURCES
How does Cheminformatics fit in with the CADD Pipeline?

Where ever we are dealing with small molecules.....
Compiled Database of Commercial Compounds: ZINC

More about ZINC subsets is [here](http://zinc.docking.org/browse/subsets/). ZINC may be used free of charge for research by individuals and institutions. **Whereas you are free to share the results of a ZINC search or a screen of molecules from ZINC, you may not redistribute major portions of ZINC without the express written permission of John Irwin.**

<table>
<thead>
<tr>
<th>Standard</th>
<th>Lead-Like</th>
<th>Fragment-Like</th>
<th>Drug-Like</th>
<th>All Purchasable</th>
<th>Shards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>3,171,902</td>
<td>532,513</td>
<td>13,205,607</td>
<td>17,833,934</td>
<td>54,061</td>
</tr>
<tr>
<td>Updated</td>
<td>2012-07-27</td>
<td>2012-07-16</td>
<td>2012-03-13</td>
<td>2012-03-15</td>
<td>2012-03-20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clean</th>
<th>Lead-Like</th>
<th>Fragment-Like</th>
<th>Drug-Like</th>
<th>All Clean</th>
<th>Clean Shards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>2,842,477</td>
<td>354,399</td>
<td>11,116,836</td>
<td>12,171,291</td>
<td>35,318</td>
</tr>
<tr>
<td>Updated</td>
<td>2012-08-01</td>
<td>2012-07-20</td>
<td>2012-04-06</td>
<td>2012-04-06</td>
<td>2012-03-29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In Stock</th>
<th>Lead-Like</th>
<th>Fragment-Like</th>
<th>Drug-Like</th>
<th>All Now</th>
<th>Shards Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>2,011,486</td>
<td>440,891</td>
<td>6,059,029</td>
<td>8,168,850</td>
<td>49,488</td>
</tr>
<tr>
<td>Updated</td>
<td>2012-07-26</td>
<td>2012-07-16</td>
<td>2012-04-24</td>
<td>2012-04-23</td>
<td>2012-03-19</td>
</tr>
</tbody>
</table>

|------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|--------------------------|------------------------|

Great resource for docking collections – but you may want to select just a few vendors
Library Sources – Commercial Library Example, Asinex

Old collections are from Russian universities – collected in the late 1990s.

They typically synthesize their own libraries.

They use sophisticated chemistries.

Some libraries will be highly specialized for current popular drug targets in pharma & they will be expensive!

General collections are a good source - Individual compounds can be ordered on-line, 48 hr delivery - Larger orders, 4-6 weeks.

It’s cheaper to order from them directly than through eMolecules.

http://www.asinex.com/
Library Source - NCI

- **Benefits** - Great resource for free compounds!
- **Caveats** – The old adage: “You get what you pay for”
  - NCI = National Cancer Institute - Most cancer compounds are toxic
  - This database has lots of dyes, steroid-like & aggregators
  - Typically they are not good starting scaffolds for chemists
  - A few compounds are being used by many academic labs

Properties
- 5 pharm4 features – dissimilar to others
- ≤ 5 rot. Bonds
- Planar
- ≤ 1 chiral center
- No leaving groups
- No organometallics
- No polycyclic aromatic hydrocarbons

- Purity >90%
- MS/LC

Library Sources: Natural Product Library

Natural Product Sources:
- Collaborations
- ZINC

Natural product properties differ from synthetic libraries:
- More O atoms, Less N atoms
- Less rotatable bonds
- Less aromatic rings
- More fused rings
- More chiral centers

Note: Many NPs break common Lipinski guidelines and so do most antibacterials.
Reputable Sources

- Asinex
- Chembridge
- ChemDiv
- Specs
- LifeChemicals
- IBS
- Maybridge
- NCI*