

ChemModeling

Compound Collection Analysis and Augmentation

Jon Swanson
jon@chemmodeling.com
(636) 329-0300
ChemModeling, LLC
April 27, 2009

Why Analyze and Augment a Collection?

- Compound collections are dynamic
 - Compounds deteriorate over time
 - New targets suggest new types of compounds to screen
- Understand overall quality of collection and improve it
 - Higher quality hits and follow-up SAR development
 - Increased hit-rate from high quality lead-like matter
 - Confidence in identity of compounds (not break-down products)
- Faster in-silico screening
 - Remove compounds medicinal chemists will reject anyway
 - Reduce duplication of pharmacologically similar compounds
 - Clustering and enrichment by target area
- Computational assessment to improve downstream success of lead and pre-clinical candidates

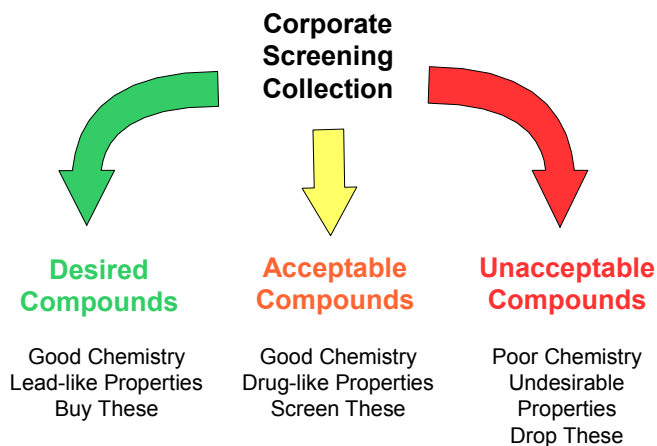
ChemModeling

2

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

We have found that companies are starting to think more seriously about their existing compound collections. Compounds that are being screened should be periodically reanalyzed for purity and composition. Compounds that fail this analysis should be removed and replaced with new compounds. In addition, as priorities change with respect to screening targets, libraries should be periodically augmented with additional compounds available from a number of vendors who sell screening libraries. Judicious selection can minimize cost and maximize return. We have experience, having completed a project of this type with a large biotech company.

Library Enhancement Strategy



Compounds can be further subdivided by target

ChemModeling

3

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

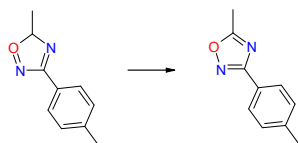
Typically the existing screening set is treated more gently than compounds that would be purchased, since the compounds are already in hand. However, even in this case, it may be desirable to weed out compounds that are unlikely to provide good hits. In the past the "acceptable compound" definition was used to filter the existing collection and the "desired compound" definition used to filter potential new purchases.

Normalization of a Compound Collection

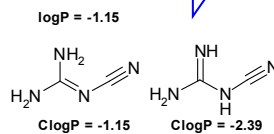
Initial Compound Collection

Correct Structure Entry Errors
Consistent Tautomeric Form
Consistent Functional Groups
Consistent Charge State

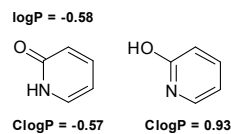
"Clean" Compound Collection



An example entry error from the PhysProp database corrected with fragment based rules



Examples of the effect of tautomeric form on ClogP corrected with ProtoPlex (ProtoPlex derived tautomers are on the left)



Yvonne Martin February 2007 CUP

ChemModeling

4

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

Our experience has been that compound databases often have a number of problems that our tools can correct. These include mis-entered structures, inconsistent representation of functional groups, multiple tautomeric forms, and inconsistent representation of salts. In the ~500K database we recently analyzed about 0.1% of the structures were mis-entered and about 2/3 could be corrected with fragment based rules. It was unclear what was intended for the other structures. About 2% of the compounds had multiple tautomeric forms. About 2.5% of the structures were duplicates (including multiple salt forms of the same structure) and 3% of the compounds were eliminated due to undesirable functional groups.

Toward a Lead-like or Targeted Subset

Compounds in a screening set should have drug-like or lead-like properties

Lipinski's "Rule of 5" is the best known filtering criteria

Poor absorption or permeation of an orally administered drug is more likely to occur if any two of these criteria are violated:

- Molecular weight is greater than 500
- Lipophilicity is high (ClogP is greater than 5)
- Number of Hydrogen bond donors is greater than 5
- Number of Hydrogen bond acceptors is greater than 10

Properties of Oral Drugs Categorized by Gene Family

| | 90% MW | 90% ClogP | 90% HBD | 90% HBA | 90% Rbonds |
|---------------------------|-----------|--------------|------------|------------|---------------|
| Aminergic GPCRs | 460 | 5.6 | 2 | 6 | 8 |
| Ion Channels | 430 | 4.7 | 3 | 6 | 7 |
| Nuclear Hormone Receptors | 495 | 7.3 | 2 | 6 | 10 |
| Peptide GPCRs | 752 | 6.5 | 8 | 10 | 17 |
| Phospho-diesterases | 465 | 5.2 | 2 | 8 | 9 |
| Protein Kinases | 505 | 5.7 | 4 | 7 | 9 |
| Serine Proteases | 572 | 4.8 | 4 | 8 | 12 |

There are **MANY** others

=> **Rules need to be tailored to specific customers needs**

Hopkins, et al, Nature Biotechnology 2006, 7, 805-815

ChemModeling

5

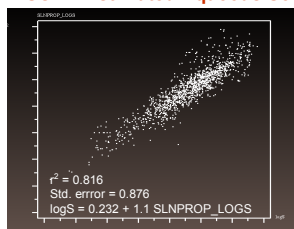
Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

The exact property filters depend on the customer and the use for which the collection is intended. Prior to setting filters we will provide the distribution of property values in the collection to allow one to estimate how much the collection will be reduced. We also typically analyze the diversity of the collection to determine how evenly distributed the compounds are in the chemistry space.

Other Factors are also Important

- Fragment Based Filters
 - Unwanted
 - Unstable
 - Toxic groups
- Similarity/Dissimilarity to Known Targets
- Custom Scoring Functions

ESOL – Estimated Aqueous Solubility



Plot of ESOL predicted solubility implemented in slnProperty versus the experimental logS values for compounds used as a training set for ALOGPS program from the ALOGPS website.

$$\log(S) = 0.16 - 0.63 * \text{ClogP} \\ - 0.0062 * \text{MW} + 0.066 * \text{RotBonds} \\ - 0.74 \text{ AromaticFraction}$$

AromaticFraction is fraction of heavy atoms in aromatic 6-membered rings

Delaney, J. S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000 – 1005.

ChemModeling

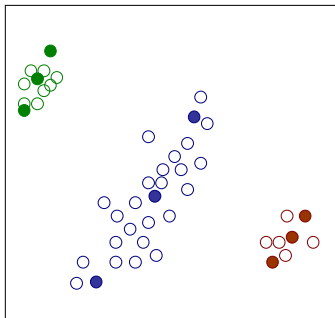
6

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

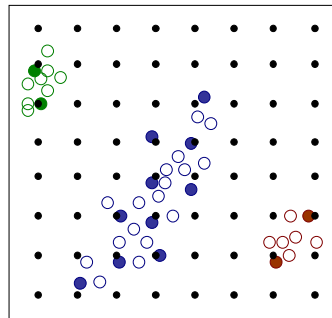
We have developed an extensive set of “undesirable” filters as a result of years of doing library design. We review these with the customer, who is free to add additional filters. We also have a fairly sophisticated scoring function built into our filtering program. It can take any combination of property or fragment filters and create a “score” value that can be filtered on. Either raw values (as in the ESOL implementation) or limits (as one might use to implement the Lipinski rules- i.e., at most two violations) can be encoded.

Toward a Representative Subset

Distance-based gridding of chemistry space allows representative selections



A typical cluster-based selection choosing 3 compounds per cluster



Selection based on equally-spaced grid points better samples clusters

Many different types of distances can be employed- 2D Tanimoto fingerprint similarities, topomer distances, SurFlex-Sim similarities, or others.

ChemModeling

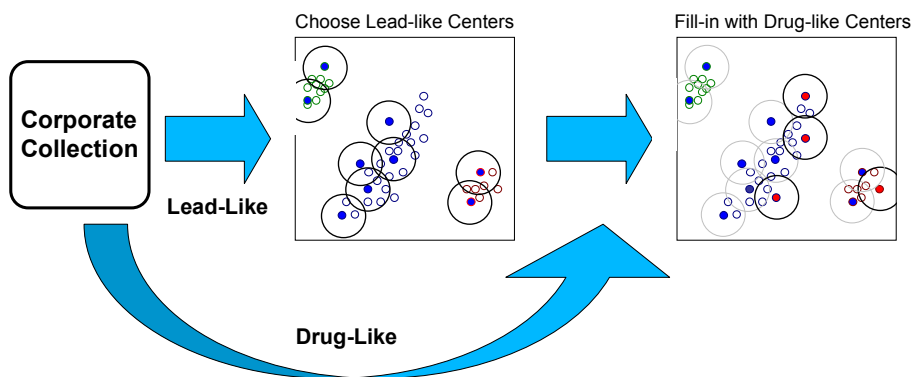
7

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

The neighborhood principle states that for a good metric, a small change in activity should correspond to a small change in the value of the metric. We choose metrics for our grid-based clustering that have good neighborhood behavior. This ensures that clusters should primarily consist of active (or inactive) compounds and that selecting a subset will not significantly reduce the chance of getting hits. In addition to the distance-based methods, we can also use low-D methods (BCUTS) or scaffold based (RECAP) clustering. We have added a k-means post-processor to the basic gridding algorithm that makes the centers chosen more central to the cluster.

Grid-based Approach Allows Flexibility

Select first from lead-like compounds and fill-in with drug-like compounds in chemistry space not covered by the lead-like selections.



Alternate approaches can be used, such as selecting based on similarity to existing targets and then filling in with lead-like matter.

ChemModeling

8

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

Our grid-based approach is fairly flexible. It accepts input seeds, so grid centers selected with a subset of the database can be used as input to select additional grid centers using a larger population. This allows sequentially selecting from several collections that would be ordered from most to least desirable. The method also successively refines the distances between centers. So that one can start with a larger distance between grid centers and later come back and cluster at a finer grid spacing. This can be useful when selecting sets of compounds to purchase.

Augmenting a Compound Collection

- Process vendor collection in same manner as corporate collection
- Produce a lead-like subset
- Compare corporate collection to vendor collection
 - Eliminate any vendor compounds that are within specified cut-off distance of corporate collection
- Cluster remaining lead-like, novel subset
 - Grid spacing for vendor collection often looser than for corporate collection
 - Can also fill-in clusters with low occupancy of corporate compounds
- Select compounds from clusters based on client preferences
 - Preferred vendors
 - Best properties
 - Best price
 - Purity

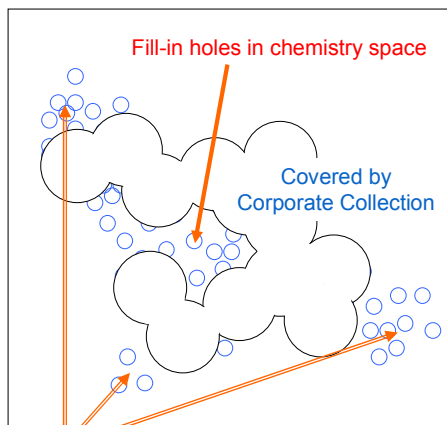
ChemModeling

9

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

We recently used a 3.5 million compound set provided by a customer to select compounds to augment their corporate collection. Roughly 40% of these compounds were unique with respect to their existing collection. The criteria used to filter this collection was much tighter than that used on the existing compounds.

Augmentation Can Be Tailored



- **Select sequentially**
 - Preferred Vendors
 - Preferred Targets
- **Select based on target**
 - Similarity to known actives
 - Privileged substructures
 - Meet pharmacophore model
 - Meet SAR model
- **Select based on properties**
 - Preferred vendors
 - Best properties
 - Best price
 - Purity

Include areas not covered by original collection

ChemModeling

10

Copyright 2009 ChemModeling All Rights Reserved, CONFIDENTIAL

In addition to selecting compounds, we can also analyze the various vendor datasets to determine how unique they are to each other, with respect to the corporate collection, or how many unique scaffolds are contained within the dataset. We can select compounds from an aggregate of vendor compounds or sequentially, allowing the maximum number of compounds to be selected from the most preferred vendors.