The Scaffold Tree

An Analysis Method for Chemical Structure Data Sets
Ansgar Schuffenhauer
Contributors

- Novartis colleagues
  - Peter Ertl
  - Silvio Roggo
  - Nathan Brown
  - Paul Selzer
  - Jeremy Jenkins
  - Kamal Azzoui
  - Jacques Hamon
  - Edgar Jacoby

- MPI Dortmund
  - Stefan Wetzel
  - Marcus Koch
  - Steffen Renner
  - Herbert Waldmann
Chemical Classification

*Why is it needed?*

- Industrial and publicly funded research using increasingly large screening collections
  - An increasingly larger part of these compounds result from parallel synthesis

- HTS of these collections gives increasingly large hit sets

- It is important to identify hits belonging to a common chemical class
  - They can be explored in joint synthesis effort
  - This effort can be guided by SAR derived from the screening data

- Get the chemical classes right, map the biological response onto the “chemical map”
# Classification of Chemical Structures

## Clustering
- Classification derived from unsupervised machine-learning
- Information of complete dataset is required for classification
- No linear scaling with dataset size
- No incremental updates possible

## Rule-based
- Explicitly formulated rules encode “expert knowledge”
- Class assignment is derived for each structure independently
- Scales linearly with number of molecules in dataset
- Incremental updates possible
The Molecular Framework and its Generalizations

Bemis and Murcko  

Addition of a cyclic sidechain prevents recognition of common core

Prune terminal sidechains

Discard atom and bond type

Discard ring sizes and linkage lengths

Not well defined chemical entities

Original structure

Molecular framework

Topological framework

Reduced framework graph

5 | The Scaffold Tree | Ansgar Schuffenhauer | Sheffield June 2007
Are there Alternatives?

- Retain the molecular framework as classification element
  - Exoxygenic and “exolinker” double bonds are part of the molecular framework

- Instead removing atom & bond type and ring size information prune less important rings sidechains piecemeal
  - Do not disconnect the scaffold
  - Use prioritization rules to decide which ring to remove first
  - Small, generic set of rules, no “dictionary”

An introductory example

*Baccatin III*
Rule 3

Choose the parent scaffold with smallest number of acyclic linker bonds

- Scaffolds having the least number of acyclic linker atoms are likely to be more rigid
  - Rigid scaffolds are more likely to present their sidechains in a conserved orientation

- Acyclic linkers are strategic bonds
  - Likely to be formed late in a parallel synthesis effort
Rule 4

*Keep bridged and spiro rings and unusually fused ringsystems.*

- Such ring patterns are unusual and likely not to be formed unintentionally.
- Use difference between number of bonds being member > 1 ring ($n_{rrb}$) and number of rings ($n_R$) - 1
  \[ |\Delta| = |n_{rrb} - (n_R - 1)| \]
  - In most common linear ringfusion pattern $n_{rrb} = n_R - 1$

Case 1: bridged rings
Rule 4
continued

Case 2: non-linearly fused rings

Case 3: spiro rings
Rule 6
Remove rings of size 3, 5 and 6 first

- The majority of the commercially available building blocks are containing rings of size 3, 5 or 6.
- If rings of different sizes occur, they are likely to be built up intentionally to fulfill a dedicated purpose.
Rule 8

Remove rings with the least number of hetero-atoms first

- Hetero-atoms can make characteristic H-bonding interactions
- Hetero-atoms bound with an exocyclic double bond
Rule 9
If number of hetero-atoms is equal priority of hetero-atoms is $N > O > S$

- N-heterocycles play an important role in medicinal chemistry
- N and O atoms are capable of forming H-bonds

Rule 10
Keep larger ring with priority

Rule 11
Of mixed aromatic/non-aromatic ring systems retain non-aromatic rings with priority

- Avoid mapping to benzene in cases there are alternatives
Rule 13
Use canonical smiles as tiebreaking rule

- Keep scaffold which has the canonical smiles with alphabetical sort precedence
- The next pruning step will prune the ring which did “win” in the tie breaking
The introductory example revisited

**Baccatin III**

Which rule is used how often? See poster T19
Classification of a public HTS data set

*NCGC Pyruvate Kinase screen*

- HTS run at the NCGC, a NIH roadmap screening institute
  - Data can be downloaded from PubChem

- 602 active and ~50 000 inactive molecules

- Scaffolds shown in tree
  - Have at least 5% actives
  - Represent at least 0.02% (10 compounds) of the whole data set.
Scaffold Tree Example for HTS results

PubChem Pyruvate Kinase Data Set

Color intensity by fraction of actives
Scaffold Tree Example for HTS results

PubChem Pyruvate Kinase Data Set

Color intensity by fraction of actives
Scaffold Tree Example for HTS results

*PubChem Pyruvate Kinase Data Set*

Color intensity by fraction of actives
Scaffold Tree Example for HTS results

PubChem Pyruvate Kinase Data Set

actives
Scaffold Tree doesn’t fit on a screen or a poster?

- Scaffold Tree Extracts meaningful chemical series
  - The tree allows visualization of scaffold hierarchy
  - If it would just fit onto the screen…

- Interactive visualization tool needed:
  - Manipulate resolution
  - Filter scaffolds

- See poster T30
Chemical Series and Biological Response

- Enrichment of actives in part of the series
  - Even in series with enriched activity the actives only enriched up to 20%
  - This suggests that the scaffolds potential for biological activity can be only materialized with the appropriate side chains
  - After all this is the rationale behind combinatorial chemistry
Chemical Series and Biological Response

- Chemical variation of library parallel to smooth change of biological response
  - Essential binding features ("privileged substructure") conserved in library variation
  - Attractive entry point for chemical exploration
  - Derivation of SAR possible

- Chemical variation of library orthogonal to smooth change of biological response
  - Chemical variation disrupts essential binding features
  - SAR appears to be "flat" or active compounds appear to be singletons

Chemical space

Desired biological activity
- Highly active
- Inactive

Library around scaffold projected into chemical space
Can we assess the relation between chemical class and biological activity in a more general manner?

- Can we compare clustering and rule-based classification?

Use in vitro data on a uniform assay panel to measure success

Two competing objectives

- As few partitions as possible
- Biological activity profile of compounds within partitions should be as similar as possible (low cluster spread SP)
- Evaluate by Pareto analysis

Pareto Analysis

One solution of the problem is superior to another solution only if it is superior in all objectives.

- A, B, C are all Pareto optimal solutions.
- B is superior to D.
- A-D are superior to E.

Random partitioning (lower end benchmark)
Profile based clustering (upper end benchmark)
Partitions by structure
Spread of Partitions in Biological Profile Space

- Distance of two compounds $i, j$ in profile space:

$$d(i, j) = \sqrt{\sum_{a}^{\text{assays}} \left[ \text{pIC}_{50}(i, a) - \text{pIC}_{50}(j, a) \right]^2}$$

- Average within partition $k$

$$sp_k = \frac{1}{2n_k(n_k - 1)} \sum_{i=1}^{n_k} \sum_{j=1}^{i<n_k, i<j} d(i, j)$$

- Average over all partitions weighted by partition size (to be minimized)

$$SP = \sum_{k=1}^{n_{\text{cluster}}} \frac{n_k}{n_{\text{total}}} sp_k$$
Pharmacology Safety Profile Data Set

- Safety-Pharmacology Profile
  - 1006 compounds
  - $IC_{50}/EC_{50}$ values in 27 assays (mostly aminergic GPCR and Ion Channels)
  - No missing values
    - Spread values well defined
Classification of Safety Data Set

Scaffold Tree Pharmacology Profile

- pIC50_Kmeans
- pIC50_noise_Kmeans
- Random
- Scaffold Tree

- 1 ring
- 2 rings
- 3 rings

SP vs. NPartitions graph with data points for different ring counts.
Classification of Safety Data Set
*Clustering with FCFP_4 Fingerprints*

![Classification of Safety Data Set Diagram](image)
Classification of Safety Data Set

Simple 1D descriptors (MW, AlogP, PSA, ..)

- Simple 1D: MW, ALogP, Num_RotatableBonds, PSA, Num_H_Acceptors, Num_H_Donors

Graph showing the performance of different clustering methods (pIC50_Kmeans, pIC50_noise_Kmeans, Random, Scaffold Tree, FCFP_4_PPClust, FCFP_4_DivKM, FEPOPS_DivKM_maj, simple_1D_DivKM) across varying numbers of partitions (NPartitions).
Are the classifications overlapping

**Adjusted Rand index matrix – Pharmacology Profile**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pIC50_Kmeans (196)</td>
<td>0.96</td>
<td>0.84</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>pIC50_noise_Kmeans (196)</td>
<td>0.84</td>
<td>0.77</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Murcko (SCINS, 167)</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
<td>0.19</td>
<td>0.26</td>
<td>0.24</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.10</td>
<td>0.11</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Scaffold_Tree (level 2, 311)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.19</td>
<td>1.00</td>
<td>0.33</td>
<td>0.34</td>
<td>0.15</td>
<td>0.14</td>
<td>0.13</td>
<td>0.10</td>
<td>0.14</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>FCFP_4_PPClust (196)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.26</td>
<td>0.33</td>
<td>0.83</td>
<td>0.79</td>
<td>0.31</td>
<td>0.30</td>
<td>0.25</td>
<td>0.21</td>
<td>0.36</td>
<td>0.09</td>
<td>0.12</td>
<td>0.13</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>UNITY_FP_PPClust (196)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.24</td>
<td>0.34</td>
<td>0.76</td>
<td>0.61</td>
<td>0.32</td>
<td>0.28</td>
<td>0.24</td>
<td>0.21</td>
<td>0.30</td>
<td>0.09</td>
<td>0.11</td>
<td>0.12</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Unity_DivKM (196)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.16</td>
<td>0.15</td>
<td>0.31</td>
<td>0.32</td>
<td>1.00</td>
<td>0.34</td>
<td>0.26</td>
<td>0.22</td>
<td>0.16</td>
<td>0.09</td>
<td>0.11</td>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>FCFP_4_DivKM (196)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.16</td>
<td>0.14</td>
<td>0.30</td>
<td>0.26</td>
<td>0.34</td>
<td>0.28</td>
<td>0.26</td>
<td>0.24</td>
<td>0.18</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>Simillog_DivKM (196)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.16</td>
<td>0.13</td>
<td>0.25</td>
<td>0.24</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td>0.17</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>RDF_DivKM (196)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.10</td>
<td>0.10</td>
<td>0.21</td>
<td>0.21</td>
<td>0.22</td>
<td>0.19</td>
<td>0.17</td>
<td>1.00</td>
<td>0.26</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>RDF_SOM (157)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.11</td>
<td>0.14</td>
<td>0.36</td>
<td>0.36</td>
<td>0.16</td>
<td>0.14</td>
<td>0.12</td>
<td>0.26</td>
<td>0.58</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>FEPOPS_DivKM_dist (189)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.05</td>
<td>0.05</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.08</td>
<td>0.06</td>
<td>0.05</td>
<td>0.48</td>
<td>0.31</td>
<td>0.10</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>FEPOPS_DivKM_maj (196)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
<td>0.12</td>
<td>0.11</td>
<td>0.11</td>
<td>0.10</td>
<td>0.10</td>
<td>0.07</td>
<td>0.06</td>
<td>0.31</td>
<td>0.80</td>
<td>0.11</td>
<td>0.14</td>
<td>0.08</td>
</tr>
<tr>
<td>FEPOPS_PPClust_dist (189)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
<td>0.13</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.10</td>
<td>0.11</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>FEPOPS_PPClust_maj (189)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.09</td>
<td>0.08</td>
<td>0.15</td>
<td>0.14</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>0.09</td>
<td>0.14</td>
<td>0.16</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>PhysChem_DivKM (196)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.09</td>
<td>0.06</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.15</td>
<td>0.11</td>
<td>0.07</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>1.00</td>
</tr>
<tr>
<td>PhysChem_PCA (182)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>Random (196)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

For adjusted rand
Index:
Hubert and Arabie
*J. Classif.* **1985**, 2, 193
Summary Pareto Analysis

- No partitioning method is generally superior to all others: trade-off between local precision and scaffold hopping potential

- The results of all methods is still more close to random than to the ideal biological clustering
  - Not all biological activity clusters are covered by single chemical class
    - Especially not all inactives
Summary

- We can with Scaffold Tree detect chemically meaningful series
  - Scaffold Tree allows us to detect series with enriched activity

- However chemical structure classes are not equivalent with biological activity classes
  - Within a chemical series biological activity can vary
  - A biological activity class can be spread over several structural classes
    - This is especially true for the “inactive” class.
  - This applies for a wide range of structural classifications

- However, continuous changes in biologic activity with chemical variation of a chemical series are actually desirable
  - That biological activity varies smoothly in a chemical class indicates optimization potential
  - Initial SAR derived from the series screening results may guide this optimization