KNIME WORKFLOWS FOR CHEMOINFORMATICS
TASKS AT ANGELINI

“4th KNIME Users Group Meeting”, Zurich March 3rd 2011
Goal
Generate a representative subset of a proprietary library

Materials and Methods
- Fingerprints
- Similarity measure
- Subset generation methods

Process for subset selection

Subset characterization

Conclusions
Select a representative and structure-diverse sub-library of compounds of our proprietary library capable of:

- fully mimicking the profile of the entire library
- maximizing the structure diversity

The approach used is based on the "similar property principle":

**structurally similar molecules have similar physicochemical properties and possibly similar biological activities**

... then maximum coverage of the activity space should be achieved by selecting a structurally diverse set of compounds.
Tools

Knime

Schrodinger Knime Extensions (Canvas)

- Cheminformatics
- Fingerprint Based Tools
  - Fingerprint Generation
  - Generate Pairwise Matrix
  - Generate Pairwise Matrix (2 Inputs)
  - Similarity Matrix (from Molecules)
  - Dissimilarity Selection (from Matrix)
  - Build Report for Clustering (from Matrix)
  - Hierarchical Clustering (from Matrix)
- Filters and Mining Tools
  - Substructure Search
  - Maximum Common Substructure Search
  - REOS Filter
  - Structure filter
- Utilities and Converters
  - Principal Components
  - Multi-dimensional Scaling
  - Combine Fingerprints
  - Concatenate Bitvectors
  - Convert Fingerprint to Bitvector
  - Convert Fingerprint to Table
  - Convert Matrix to Table
  - Convert Table to Fingerprint
  - Convert Table to Matrix
  - Convert Bitvector to Fingerprint
KNIME workflow

Materials

“4th KNIME Users Group Meeting”, Zurich  March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
KNIME workflow

Materials

“4th KNIME Users Group Meeting”, Zurich  March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
Applications of fingerprint methods

- Fingerprints are binary string representations of the molecules as a sequence of “0” and “1”, encoding the absence or presence of a set of structural fragments (or chemical features).

- Clustering of molecules
- Diversity analysis and selection of molecules
- QSAR model building
- Similarity search of a database
- etc…
Fingerprints: descriptions of molecules

- **Structure keys**
  Each bit of a structure key maps to a specific, predefined substructure or molecular feature specified using a fragment dictionary.

- **Hashed fingerprints**
  Bits are enumerated from the molecules generating all possible fragments according to the fingerprint type.
  - May be applied to any type of molecule.
  - May map two distinct fragments to the same bit (collision).
Fingerprints in Canvas

- **Hashed fingerprints**
  - Linear
  - Dendritic
  - Radial (Circular fingerprints)
  - Molprint2D (Circular fingerprints)
  - Pairwise
  - Triplet
  - Torsion

- **Structure keys**
  - MACCS keys

Methods

Linear
Dendritic
Pairwise

Radial
Molprint2D
Atom Typing Schemes
1. All atoms equivalent; all bonds equivalent.
2. Atoms distinguished by HB acceptor/donor; all bonds equivalent.
3. Atoms distinguished by hybridization state; all bonds equivalent.
4. Atoms distinguished by functional type: \{H\}, \{C\}, \{F,Cl\}, \{Br,I\}, \{N,O\}, \{S\}, \{other\}; bonds by hybridization. (Radial)
5. Mol2 atom types; all bonds equivalent. (MOLPRINT2D)
6. Atoms distinguished by whether terminal, halogen, HB acceptor/donor; bonds distinguished by bond order.
7. Atomic number and bond order.
8. Atoms distinguished by ring size, aromaticity, HB acceptor/donor, ionization potential, whether terminal, whether halogen; bonds distinguished by bond order.
10. Daylight invariant atom types; bonds distinguished by bond order. (Linear, Dendritic, Torsion)

FP Scaling options
0. no scaling (default)
1. scale counts by feature size to unity
2. scale counts by feature size to feature size
3. scale counts by feature size to molecule size
4. scale squares of counts by feature size to unity
5. scale squares of counts by feature size to feature size
6. scale squares of counts by feature size to molecule size
7. scale sqrt of counts by feature size to unity
8. scale sqrt of counts by feature size to feature size
9. scale sqrt of counts by feature size to molecule size
10. use raw feature counts
11. use raw feature counts squared
12. use sqrt of raw feature counts

Metrics
Buser, Cosine, Dice, Hamann, Hamming, Dixon, Euclidean, Kulczynski, Matching Pattern Difference, Tanimoto, modifiedTan, Shape, Size, Simpson, Petke, Tversky, Yuel
Variance, Soergel, rogersT, Pearson, Minmax

25,344 combinations!
8 Fingerprints

Methods
Choice of fingerprints

Among the wide variety of possible 2D fingerprint methods there is no one that, a priori, consistently performed better than the others.

In the similarity searching based on 2D fingerprints it is rarely known which fingerprint is most suitable for the particular target and query.

Several investigations into the information content of fingerprints and atom-typing schemes revealed that no approach works better in all cases.

The choice should be validate case by case exploring different fingerprinting methods.

Chemical and molecular similarity are strongly dependent on how the molecular structure is measured and described.
Fingerprints in Canvas

- **Hashed fingerprints**
  - Linear
  - Dendritic
  - Radial (Circular fingerprints)
  - Molprint2D (Circular fingerprints)
  - Pairwise
  - Triplet
  - Torsion

- **Structure keys**
  - MACCS keys

- **Combined fingerprints**
  - Dendritic + Radial
  - Dendritic + Triplet

---

Methods

- **Combined fingerprints**
  - Dendritic + Radial
  - Dendritic + Triplet

---

“4th KNIME Users Group Meeting”, Zurich  March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
Atom Typing Schemes

1. All atoms equivalent; all bonds equivalent.
2. Atoms distinguished by HB acceptor/donor; all bonds equivalent.
3. Atoms distinguished by hybridization state; all bonds equivalent.
4. Atoms distinguished by functional type: \{H\}, \{C\}, \{F,Cl\}, \{Br,I\}, \{N,O\}, \{S\}, \{other\}; bonds by hybridization. (Radial)
5. Mol2 atom types; all bonds equivalent. (MOLPRINT2D)
6. Atoms distinguished by whether terminal, halogen, HB acceptor/donor; bonds distinguished by bond order.
7. Atomic number and bond order.
8. Atoms distinguished by ring size, aromaticity, HB acceptor/donor, ionization potential, whether terminal, whether halogen; bonds distinguished by bond order.
9. Carhart atom types (atom-pairs approach). (Triplet)
10. Daylight invariant atom types; bonds distinguished by bond order. (Linear, Dendritic, Torsion)

FP Scaling options

0 – no scaling (default)
1 - scale counts by feature size to unity
2 - scale counts by feature size to feature size
3 - scale counts by feature size to molecule size
4 - scale squares of counts by feature size to feature size
5 - scale squares of counts by feature size to molecule size
6 - scale squares of counts by feature size to feature size
7 - scale sqrt of counts by feature size to unity
8 - scale sqrt of counts by feature size to feature size
9 - scale sqrt of counts by feature size to molecule size
10 - use raw feature counts
11 - use raw feature counts squared
12 - use sqrt of raw feature counts

Metrics

Buser, Cosine, Dice, Hamann, Hamming, Dixon, Euclidean, Kulczynski, Matching, PatternDifference Tanimoto modifiedTan, Shape, Size, Simpson, Petke, Tversky, Yuel Variance, Soergel, rogersT, Pearson, Minmax, McConnaughey
Similarity measures are generally based on the presence and/or absence of features in two molecules, estimating the number of matches or overlap between them.

The *Tanimoto coefficient* is the most widely used coefficient for binary fingerprints:

\[
T = \frac{c}{a + b + c} \quad 0 \leq T \leq 1
\]

- **a**: number of bits=1 present in A and absent in B
- **b**: number of bits=1 present in B and absent in A
- **c**: number of bits=1 common to both
KNIME workflow
“Distance-based” methods:
- Cluster analysis
- Dissimilarity-based compound selection

A clustering algorithm divide a group of objects into clusters where molecules within each cluster are more closely related to one another than objects assigned to different clusters.

Hierarchical Clustering
It is a non-overlapping clustering method. Clusters increase in size, starting from a situation of a single compound per cluster.

Agglomerative methods start at the bottom and merge similar clusters.

Ward’s method
Clusters are formed to minimize the variance

A representative subset was selected by taking one compound (centroid) from each cluster.
KNIME workflow

“4th KNIME Users Group Meeting”, Zurich March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
KNIME workflow

Hierarchical Clustering (from Matrix)

Flow Variables
- Linkage Type: ward
- Number of Clusters: 400

Column containing input: CanvasMatrix

OK  Apply  Cancel

Additional nodes and connections are present in the workflow diagram, including nodes for similarity matrix generation, molecular descriptors, and other operations relevant to chemoinformatics tasks.
KNIME workflow

Materials

KNIME workflows for chemoinformatics tasks at Angelini
Attempt to identify a diverse set of compounds directly.

1) The desired size of the final subset was decided.
2) A **selection method** to choose the first compound of the subset was selected.
3) Dissimilarity was calculated with every other compounds in the subset according to a **dissimilarity algorithm**.
4) The next compound was chosen as the most dissimilar to those in the subset.

**Most Dissimilar**

takes the compound with the smallest sum of similarities to the other molecules.

**MaxMin**

chooses the compound with the maximum distance to its closest neighbour in the subset.

**“Distance-based” methods:**

- **Cluster analysis**

- **Dissimilarity-based compound selection**
KNIME workflows for chemoinformatics tasks at Angelini
1st step: generation of several subsets

Angelini Library

Fingerprints calculation

• dendritic
• radial
• triplet
• molprint2D
• maccs
• radial-dendritic
• triplet-dendritic

Pairwise similarity measure

Subsets generation

Cluster analysis: 7 subsets
Dissimilarity-based method: 7 subsets
2nd step: subset selection by self-similarity

Self-similarity matrix calculation and graphical representation

The smaller the mean Tanimoto coefficient, the greater the subset structure diversity.

<table>
<thead>
<tr>
<th>Mol</th>
<th>AF₁</th>
<th>AF₂</th>
<th>AF₃</th>
<th>AFₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF₁</td>
<td>1</td>
<td>0.12</td>
<td>0.04</td>
<td>0.26</td>
</tr>
<tr>
<td>AF₂</td>
<td>0.12</td>
<td>1</td>
<td>0.35</td>
<td>0.23</td>
</tr>
<tr>
<td>AF₃</td>
<td>0.04</td>
<td>0.35</td>
<td>1</td>
<td>0.42</td>
</tr>
<tr>
<td>AFₙ</td>
<td>0.26</td>
<td>0.23</td>
<td>0.42</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mol</th>
<th>Max Tc</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF₁</td>
<td>0.26</td>
</tr>
<tr>
<td>AF₂</td>
<td>0.35</td>
</tr>
<tr>
<td>AF₃</td>
<td>0.42</td>
</tr>
<tr>
<td>AFₙ</td>
<td>0.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Similarity</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.26</td>
</tr>
<tr>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>..</td>
<td>0.19</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Process

Subset selection
KNIME workflow

Materials

"4th KNIME Users Group Meeting", Zurich  March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
KNIME workflow

Materials

“4th KNIME Users Group Meeting”, Zurich  March 3rd 2011 - KNIME workflows for chemoinformatics tasks at Angelini
KNIME workflow

Chemistry External Tool 0:1

Fingerprints type

Chemistry External Tool 0:1

Atom type

MetaNode 2:1

TableRow To Variable Loop Start

Injections Variables (Data)

Molecule Re...

Variable To TableColumn

Loop End

CSV Writer

MetaNode 2:1

workflows for chemoinformatics tasks at Angelini
Subsets of AF library by Cluster Analysis

Mean Tc
library = 0.75
0.2 < subsets < 0.4
Subsets of AF library by Dissimilarity

**Mean Tc**
- Library = 0.75
- Subsets = 0.1
Subsets of AF library by random selection

Random1-2: draw randomly
Random3: stratified sampling by name
Random4: linear sampling

Diverse-subset
Random-subsets
Library

Percentage of compounds vs. Tanimoto coefficient
3rd step: subset characterization

- Molecular Weight
- Log P
- Rotatable Bond
- Polar Surface Area

Subset physicochemical characterization
Subset physicochemical profile:

**Molecular weight**

<table>
<thead>
<tr>
<th>Subset</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of compounds</td>
<td>60%</td>
</tr>
<tr>
<td>Molecular weight</td>
<td>1000</td>
</tr>
</tbody>
</table>

**LogP**

<table>
<thead>
<tr>
<th>Subset</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of compounds</td>
<td>45%</td>
</tr>
<tr>
<td>LogP</td>
<td>4</td>
</tr>
</tbody>
</table>

**Rotatable bond**

<table>
<thead>
<tr>
<th>Subset</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of compounds</td>
<td>60%</td>
</tr>
<tr>
<td>Rotatable bond</td>
<td>18</td>
</tr>
</tbody>
</table>

**Polar Surface Area**

<table>
<thead>
<tr>
<th>Subset</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of compounds</td>
<td>60%</td>
</tr>
<tr>
<td>Polar surface area</td>
<td>270</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compounds</th>
<th>M.W</th>
<th>LogP</th>
<th>RB</th>
<th>PSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset</td>
<td>50-700</td>
<td>-2-8</td>
<td>0-14</td>
<td>0-180</td>
</tr>
<tr>
<td>Library</td>
<td>50-700</td>
<td>-2-8</td>
<td>0-14</td>
<td>0-180</td>
</tr>
<tr>
<td>Drug-like</td>
<td>200-500</td>
<td>-2-5</td>
<td>0-8</td>
<td>0-120</td>
</tr>
</tbody>
</table>
Subset chemical space profile compared to the entire library

**PCA**

- Library
- Subset

**MFS map**

- Library
- Subset

R²ₓ[1] = 0.422392            R²ₓ[2] = 0.29364             Ellipse: Hotelling T² (0.95)
Conclusion

- We demonstrated that in general no approach works better than the other but we identified the best exploring and validating different fingerprinting methods.

- Using knime allowed us to easily investigate different combinations of methods.

- In the particular case of our system the best choice, in terms of self-similarity index distribution, is the Diversity Analysis using a combination of fingerprints Dendritic-Radial with the specific atom typing scheme Daylight.
Acknowledgements

- R. Ombrato (Head of Molecular Modelling Lab)
- N. Cazzolla (Head of Chemistry Department)
- The group of the Medicinal Chemistry Dept
- E. Fioravanzo, A. Bassan (S-IN Soluzioni Informatiche)

And all of you for your attention!!