What’s important and what’s not there?
Analyzing sets of compounds from patents

Greg Landrum, Ph.D.
KNIME AG
The problem
The problem

• You found an interesting patent but want to get an overview of the chemistry in the document
• Many of the structures are fragments or building blocks
• There’s no indication of which structures are particularly interesting or relevant
• What structures were left out of the patent?
What we’ll do

• Identify a cluster of compounds that is likely to contain the key compound
• Approximate the MCS for the patent by finding the “core” for those compounds
• Do an R-group decomposition to find side chains used in the patent
• Enumerate all possible combinations of those sidechains
Making it work

• We could automate all of this, but here we’ll show how to put a human in the loop by making it interactive

• We’ll do this using KNIME Analytics Platform and the RDKit

www.knime.com  www.rdkit.org
A classic approach

• Find the “most interesting” compounds in the patent by identifying those that have a large number of neighbors (=very similar compounds)

• Straightforward and well validated, but just gives you a (small) set of compounds.

A refinement

• Construct a network based on similarity and calculation the "hub score" of each node. Rank compounds by hub score.

• Hub score in these undirected networks: determined by the number of highly connected neighbors

Validation 1: Key compound

- Start with ChEMBL "marketed drugs" list. Filter out drugs violating Ro5 (ChEMBL label). Take 25 most recent (by "First Approval" field). 19 of these were useful.
- Pick oldest SureChEMBL patent containing each drug and download the structures.
- Success criterion: marketed drug is in first 10 compounds sorted by Hub Score.
- Results:
  - Success: 8
  - Failure: 11
- 16 of the 19 examples have the marketed drug in one of the first three clusters.

1 Note that this is stricter than in Hattori et al.
Can we do more than find the key/interesting compound?

• Approximate the Markush structure: take the whole cluster and do a fuzzy MCS
  – Matches at least 90% of the compounds
  – Ignore atom/bond types
  – Only complete rings
• Retrieve all compounds matching that substructure
• Do an R-group decomposition using the substructure as the core
Example "Markush" structures

- Brexiprazole: “PIPERAZINE-SUBSTITUTED BENZOTHIOPHENES FOR TREATMENT OF MENTAL DISORDERS”

- Markush from patent

- Fuzzy MCS from network
Example "Markush" structures

- Avibactam: “AZABICYCLIC COMPOUNDS, PREPARATION THEREOF AND USE AS MEDICINES, IN PARTICULAR AS ANTIBACTERIAL AGENTS”

Avibactam

Markush from patent

Fuzzy MCS from network
Validation 2: ”Markush” structure

• Start with ChEMBL ”marketed drugs” list. Filter out drugs violating Ro5 (ChEMBL label). Take 25 most recent (by “First Approval” field). 19 of these were useful
• Pick oldest SureChEMBL patent containing each drug and download the structures.
• Generate the network, pick the cluster with the highest hub score, and generate fuzzy MCS
• Check to see if this retrieves the marketed drug
• Results:
  – Success: 13
  – Failure: 6
What about structures that were left out of the patent?

- Combine core structure and sets of possible side chains to generate new structures
Let’s look at some patent data
The workflow, part 1
The workflow, part 1

Choose the file with patent structures
The workflow, part 1
The workflow, part 1

Filter molecules by property
Filter by property

Filter

<table>
<thead>
<tr>
<th>RowID</th>
<th>Image</th>
<th>schembl_id</th>
<th>molecular_weight</th>
<th>logp</th>
<th>ring_count</th>
<th>rotatable_bond_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row1</td>
<td><img src="image1.png" alt="Image" /></td>
<td>SCHEMBL101</td>
<td>262.2850036621094</td>
<td>5.106599807739258</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Row6</td>
<td><img src="image2.png" alt="Image" /></td>
<td>SCHEMBL1050</td>
<td>152.1470317382812</td>
<td>0.8956359956741333</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Filter by property

How do you create this?
Filter by property

How do you create this?
The workflow, part 1

View network and choose structures
View network and structures

Compound network
View network and structures

Compound network

<table>
<thead>
<tr>
<th>RowID</th>
<th>Image</th>
<th>Hub score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCHEMBL5212176</td>
<td><img src="image1.png" alt="Image" /></td>
<td>1</td>
</tr>
<tr>
<td>SCHEMBL874122</td>
<td><img src="image2.png" alt="Image" /></td>
<td>0.8885267423</td>
</tr>
<tr>
<td>SCHEMBL676234</td>
<td><img src="image3.png" alt="Image" /></td>
<td>0.8555384619</td>
</tr>
</tbody>
</table>
The workflow, part 2

Find the core with fuzzy MCS
Find all molecules with that “core”
Do an R group decomposition around the “core”
R-group decomposition results
The workflow, part 3

Combinatorially generate all possible sidechain combinations
The workflow, part 3

Generate RDKit molecules for each core + sidechain combination
Remove examples that were in the patent
Summary

• Network metrics are a helpful extension to the usual approach for identifying the key compound(s) in a patent

• Using the open-source KNIME Analytics Platform it’s easy to build a workflow to interactively explore and analyze these data
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States.