Analyzing Exemplified and Markush Structures in Patents

Christopher E. Kibbey, Jacquelyn L. Klug-McLeod, Bruce A. Lefker, Mark A. Mitchell*, and Robert Owen

Pfizer Worldwide Research and Development, Worldwide Medicinal Chemistry
*Pfizer Worldwide Research and Development, Business Technology
Computational Sciences CoE

Competitive Intelligence and Drug Design

A thorough understanding of the competitive landscape is an important aspect of drug design.

• Influence the assessment of multiple series for early projects

• Assess the strengths and weaknesses of our chemical matter relative to our competitors

• Facilitate identification of unexplored areas of chemical space in a competitor’s IP

• Drive patent strategies to strike a balance between cost and maintaining our competitive advantage

• Protect Pfizer’s chemical equity

However, efficiently obtaining this information can be tricky and tedious, especially in crowded or rapidly changing environments.

• Access to electronic structures from patents and appropriate tools for analysis is critical to success
Patent structure databases in use at Pfizer and their application to drug design

Chemistry Landscape
- Substructure search
- Similarity search
- Exact-match search
- IP assessment

Project Datasets
Download structures for specific patents related to a medicinal chemistry project

Patent Ranking
- Rank order a series of patents by feature similarity to a query structure

Feature Analysis/Visualization
- Detailed analysis and visualization of substructure features within patent structures compared with those for a set of query structures

*IBM database of ~8 million exemplified structures from patents (nightly updates)
*GVK database of ~3 million exemplified structures from patents (quarterly updates)
**Thomson-Reuters database of 1.2 million Markush structures from patents (weekly updates)
Comparison of patent structures from commercial providers

Distribution of exemplified structures (by vendor) across five medicinal chemistry projects

<table>
<thead>
<tr>
<th>Project 1</th>
<th>Project 2</th>
<th>Project 3</th>
<th>Project 4</th>
<th>Project 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.3%</td>
<td>45.4%</td>
<td>44.9%</td>
<td>48.9%</td>
<td>67.3%</td>
</tr>
<tr>
<td>16.4%</td>
<td>8.7%</td>
<td>18.3%</td>
<td>19.6%</td>
<td>8.4%</td>
</tr>
<tr>
<td>6.4%</td>
<td>18.9%</td>
<td>6.8%</td>
<td>3.2%</td>
<td>24.8%</td>
</tr>
<tr>
<td>6.8%</td>
<td>8.8%</td>
<td>42%</td>
<td>10.8%</td>
<td></td>
</tr>
<tr>
<td>26.1%</td>
<td></td>
<td>32.6%</td>
<td>12.7%</td>
<td></td>
</tr>
</tbody>
</table>

Confirmation of exemplified structures from commercial providers

- CAS: 100.0%
- GVK: 98.1%
- IBM: 80.6%

Legend:
- Green: Confirmed exemplified
- Red: Confirmed reagent, intermediate, or unrelated substance
Types of structure variation encoded in exemplified and Markush structures in patents

- **Position (p-variation)**
  - ![Position example](image)

- **Substitution (s-variation)**
  - ![Substitution example](image)

- **Frequency (f-variation)**
  - ![Frequency example](image)

- **Homology (h-variation)**
  - ![Homology example](image)
Substructure and fingerprint-based similarity search results can be misleading when applied to structures in patents.

Substructure search cannot account for h-, s-, p-, and f-variation.

Furthermore, fingerprint-based similarity search algorithms may not take connectivity of fingerprint features into account.
Reduced graphs offer an alternative to fingerprint-based similarity search algorithms

Feature trees: A new molecular similarity measure based on tree matching
Feature Analysis – Reduced graph representation of a structure

Fragment Nodes

Reduced Graph

R  Ring
C  Chain
S  Substituent
### Feature Analysis - Encoding node fingerprints

<table>
<thead>
<tr>
<th>Substituents-S</th>
<th>Chains-C</th>
<th>Rings-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contains C</td>
<td>Contains C</td>
<td>Fused</td>
</tr>
<tr>
<td>Contains N</td>
<td>Contains N</td>
<td>Fused and Fully Aromatic</td>
</tr>
<tr>
<td>Contains O</td>
<td>Contains O</td>
<td>Contains C</td>
</tr>
<tr>
<td>Contains S</td>
<td>Contains S</td>
<td>Contains N</td>
</tr>
<tr>
<td>Has Single Bond</td>
<td>Has Single Bond</td>
<td>Contains O</td>
</tr>
<tr>
<td>Has Double Bond</td>
<td>Has Double Bond</td>
<td>Contains S</td>
</tr>
<tr>
<td>Has Carbonyl/Sulfonyl Bond</td>
<td>Has Carbonyl/Sulfonyl Bond</td>
<td>Has Single Bond</td>
</tr>
<tr>
<td>Has Triple Bond</td>
<td>Has Triple Bond</td>
<td>Has Double Bond</td>
</tr>
<tr>
<td>Is Branched</td>
<td>Is Branched</td>
<td>Has Carbonyl/Sulfonyl Bond</td>
</tr>
<tr>
<td>Has Charged Atoms</td>
<td></td>
<td>Has Triple Bond</td>
</tr>
</tbody>
</table>

![Molecules](image)  
S01010101000  
C01010100000  
R11111000000
Feature Analysis - Overlaying reduced graphs and scoring

*Score = weighted average similarity for matched nodes – penalty for unmatched nodes
Advantages of applying Feature Analysis to structures in patents

Exemplified structures can be represented by reduced graphs consisting of inter-connected substituent, chain, and ring nodes.

FA algorithm is compatible with f-, h-, p- and s-variation represented in patent structures (both exemplified and Markush)

- f-variation: allows for variation in chain, ring and substituent size by atom type
- h-variation: various homology group definitions can be encoded in node fingerprints
- p-variation: algorithm ignores attachment geometry among R, C, S nodes
- s-variation: algorithm accommodates differences in substitution pattern between pairs of reduced graphs being compared

The output of Feature Analysis provides a “similarity-like” score and a “substructure-like” match of ring, chain, and substituent features.
Visualizing the results of Feature Analysis

Select a query feature to display corresponding sub-structure features for exemplified compounds.

Per-atom differences highlighted in blue for 100% similar fragments.
Benzimidazole ring in the query structure maps to a variety of ring nodes within the target set.
Feature Analysis applied to ranking patents

<table>
<thead>
<tr>
<th>Id</th>
<th>FEATURES_ATOMPROP</th>
<th>BestScore</th>
<th>BestMatchAtom</th>
<th>PatentNumber</th>
<th>BioAssays</th>
<th>BioScores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td>US6251559</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.81</td>
<td></td>
<td>US2000619050</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.81</td>
<td></td>
<td>US20005152407</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76</td>
<td></td>
<td>US63827371</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.74</td>
<td></td>
<td>US5551666</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.74</td>
<td></td>
<td>US6159321</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.7</td>
<td></td>
<td>US20006311657</td>
<td>[0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]</td>
<td></td>
</tr>
</tbody>
</table>

Color-coding of query features to best match in patent

Distribution of scores for all structures in patent

Feature Score for best match in patent

Structure of best match in patent
Markush analysis in drug design

1. Influence the assessment of multiple series for early projects
   • What is the overlap of the project series with the patent Markush?
   • What regions of the project series exhibit potential novelty?

2. Facilitate identification of unexplored areas of chemical space
   • What regions of the Markush are under-represented by the exemplified structures in the patent?
   • Are these under-represented regions covered by granted claims?
Challenges to Markush analysis

Substructure and exact match searches against Markush structures can be problematic

- Use of specialized drawing rules to encode Markush structures
  - Normalized bonds (e.g. \( \text{as } \) and \( \text{as } \))
  - Limitations imposed by indexing system rules
    - May leave gaps in chemistry space covered by patent

- Inconsistent handling of stereochemistry in patents

No commercial solution exists for similarity search against Markush structures
Example: Substructure search fails due to Markush indexing

WO2010105179 – 24 Markush Structures

Exemplified Structure

Substructure Search

No Hits

1-12

13-24
Enumeration of virtual libraries from a Markush is an ineffective strategy for (similarity) analysis.

Random enumeration generates complete structures using randomly chosen Rgroup instances from the Markush.

Random enumeration yields structures with variation far-away from the core scaffold within an extremely large virtual space (e.g., \(~10^{15}\) ).
Enumeration of virtual libraries from a Markush is an ineffective strategy for (similarity) analysis.

Sequential enumeration generates structures using only the first instance defined at each Rgroup in the Markush.

Sequential enumeration generates smaller virtual libraries of close-in structures, but ignores much of the chemical space within the Markush.
Level enumeration generates structures through random enumeration of Rgroups up to a maximum specified nesting depth.

MMS allows up to 50 Rgroups with 4 levels of nesting per Markush structure.
Level 1
Library Size = 24

Level 2
Library Size = $10^{11}$

Level 3
Library Size = $10^{14}$

Level 4
Library Size = $10^{15}$

Level enumeration yields smaller libraries of close-in structures while maintaining representative Rgroup coverage.
Strategy for similarity search against Markush (Level enumeration + Feature Analysis)

Apply level enumeration within a Markush to generate a representative set of close-in structures

Encode the enumerated structures as a set of reduced-graphs
  • Examine Rgroup definitions to encode ring, chain, substituent nodes
  • Eliminate duplicate Rgroups that map to the same node-type

Compare the Markush reduced-graphs with that for a query structure
  • Score the best match (IP Assessment)
  • Encode target node mapping for the best match within the query structure
  • Return the corresponding structure from level enumeration as the best match for the patent Markush structure
Acknowledgements

Bonnie Bacon
Greg Bakken
Marudai Balasubramanian (Balu)
Markus Boehm
Rajiah Denny
Klaus Dress
Anton Fliri
Kevin Foje
Katelin Grover
Robert Goulet
Kazu Hattori
Steven Heck
Andrew Hopkins
Xinjun Hou
Greg Kauffman
Christopher Kibbey
Jacquelyn Klug-McLeod
Bruce Lefker
Jens Loesel
Scot Mente
Mike Miller
James Mills
Mark Mitchell
Israel Nissenbaum
Matthias Nolte
David O’Neill
Robert Owen
Martin Pettersson
Gena Poda
Gaia Paolini
Usa Datta Reilly
Vineet Sardar
Keith Schreiber
Meihua Tu
David Walsh
Simon Xi
Christoph Zapf