CHALLENGES AND SUCCESSES IN MACHINE INTERPRETATION OF MARKUSH DESCRIPTIONS

Daniel Lowe, John Mayfield and Roger Sayle
NextMove Software (and MineSoft)
Cambridge, UK
TOPICS

• Specific compounds described by Markush structures
  – Simplifying the problem.

• Progress towards capturing Markush structures
  – Tackling the problem head-on.

• Performing generic structural searches (OntoGrep)
  – Finessing the problem.
INTRODUCTION: TRIVIAL MARKUSH

1. A compound of formula (I)

wherein

R¹ is selected from CONH₂ or CONHCH₃; and
R² is selected from H, F, Cl or Br;

or a pharmaceutically acceptable salt thereof.

Markush core

Definitions

US20150376170
**Approach 1: Simplify**

- Be careful what you ask for...
- Relatively few organizations or databases support the ability to capture process Markush records; Markush registration and duplicate checking, and searching Markush vs. Markush, Compound vs. Markush, and Markush vs. Compound.
- Lower hanging fruit is to export regular compounds from fully exemplified Markush examples, such as those commonly found in R-Group tables.
R-GROUP TABLES (EXAMPLE 1)

<table>
<thead>
<tr>
<th>No.</th>
<th>$X_n$</th>
<th>$R^1$</th>
<th>$R^2$</th>
<th>$R^3$</th>
<th>$A$</th>
<th>LogP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6-Cl</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td></td>
<td>$[b]_{2,42}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$[c]_{2,41}$</td>
</tr>
</tbody>
</table>

254th ACS National Meeting, Washington DC, USA, Tuesday 22nd August 2017
# R-GROUP TABLES (EXAMPLE 2)

**TABLE 1**

<table>
<thead>
<tr>
<th>No.</th>
<th>Ar</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-1</td>
<td>3-chlorophenyl</td>
<td><img src="image1" alt="Chemical Structure" /></td>
<td><img src="image2" alt="Chemical Structure" /></td>
</tr>
<tr>
<td>I-2</td>
<td>3-chlorophenyl</td>
<td><img src="image3" alt="Chemical Structure" /></td>
<td><img src="image4" alt="Chemical Structure" /></td>
</tr>
<tr>
<td>I-3</td>
<td>3-chlorophenyl</td>
<td><img src="image5" alt="Chemical Structure" /></td>
<td><img src="image6" alt="Chemical Structure" /></td>
</tr>
</tbody>
</table>

---

**US20160304465A1**
R-GROUP TABLES (EXAMPLE 3)

Additional compounds of Formula (1) are shown:

![Chemical Structure](image)

wherein non-limiting examples of $R^a$, $R^b$, $R^c$ and $R^d$ are further described herein below in Table.

<table>
<thead>
<tr>
<th>No</th>
<th>$R^i$</th>
<th>$R^6$</th>
<th>$R^8a$</th>
<th>$R^8b$</th>
<th>$R^9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>559</td>
<td>isochlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>560</td>
<td>isochlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>561</td>
<td>isochlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>562</td>
<td>isochlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>563</td>
<td>isochlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
<tr>
<td>564</td>
<td>isochlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
<tr>
<td>565</td>
<td>4-chlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>566</td>
<td>4-chlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>567</td>
<td>4-chlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>568</td>
<td>4-chlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>569</td>
<td>4-chlorophenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
<tr>
<td>570</td>
<td>4-chlorophenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
<tr>
<td>571</td>
<td>4-methylphenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>572</td>
<td>4-methylphenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
<td>-H</td>
</tr>
<tr>
<td>573</td>
<td>4-methylphenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>574</td>
<td>4-methylphenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-H</td>
</tr>
<tr>
<td>575</td>
<td>4-methylphenyl</td>
<td>-H</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
<tr>
<td>576</td>
<td>4-methylphenyl</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
<td>-CH$_3$</td>
</tr>
</tbody>
</table>
# RESULTS

(2001–JUNE 2016 PATENT APPLICATIONS)

<table>
<thead>
<tr>
<th>Data type</th>
<th>Unique Compounds</th>
<th>Not in PubChem</th>
<th>Not in PubChem (SureChEMBL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemplified compound R-group tables</td>
<td>621,140</td>
<td>496,831 (80.0%)</td>
<td>532,166 (85.7%)</td>
</tr>
<tr>
<td>Text</td>
<td>4,759,009</td>
<td>564,886 (11.9%)</td>
<td>911,976 (19.2%)</td>
</tr>
<tr>
<td>Sketches</td>
<td>4,479,113</td>
<td>886,991 (19.8%)</td>
<td>1,179,229 (26.3%)</td>
</tr>
</tbody>
</table>

Structural identity checks performed using StdInChI
APPROACH 2A: ENCODE SKETCHES

• Interpretation of Markush core sketch
  – Substituent variation
  – Position variation
  – Frequency variation

• Features captured as ChemAxon extended SMILES (CXSMILES)
GENERIC FEATURES

Generic structure contain some type of variation

- substituent
- positional
- frequency
- homology (*least conventions*)

Complexity from conventions, e.g. label for \( \textbf{f-var: } (\text{CH}_2)_n \)

Complexity from combinations e.g. \( \textbf{s+p+f} \)
**POSITIONAL VARIATION**

Original

**.C=1C=C(C=CC1)C(C2=C(N(N=C2*)C3=CC=CC=C3)*)=O.**

Naïve Export

As CXSMILES

US20020016333A1-20020207-C00031
EXPANSION OF REPEATED GROUPS
R-GROUP SKETCH INTERPRETATION

- ChemDraw files from US patents used
- Substituent attachment points detected:

  ![Chemical structures](image)

  - Always interpreted as attachments points
  - Interpreted as attachment point when in R-group table
## Formula (Re)Interpretation

<table>
<thead>
<tr>
<th>Input</th>
<th>ChemDraw 15</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>HATU</td>
<td><img src="image" alt="HATU" /></td>
<td><img src="image" alt="HATU" /></td>
</tr>
<tr>
<td>$\text{C}_4\text{F}_9$</td>
<td><img src="image" alt="C4F9" /></td>
<td><img src="image" alt="C4F9" /></td>
</tr>
<tr>
<td>$\text{H}_3\text{PO}_4$</td>
<td><img src="image" alt="H3PO4" /></td>
<td><img src="image" alt="H3PO4" /></td>
</tr>
<tr>
<td>CON(cHex)2</td>
<td>No result</td>
<td><img src="image" alt="CON(cHex)2" /></td>
</tr>
<tr>
<td>III-2</td>
<td><img src="image" alt="III-2" /></td>
<td>No result</td>
</tr>
</tbody>
</table>
**COMPARSED TO PREVIOUS EFFORTS**

**SCRIPDB**
CDX and Molfile \w Open Babel
No (apparent) correction
Fragments split

**SureChEMBL**
Images \w CLiDE
Some correction
Partial recognition (e.g. 2 of 4 in collection)
Approach 2B: “Full” Markush

• Traditionally, the “full” Markush problem is to capture definition lists, R-groups, which can be generic (homology groups) and recursive.

• Definitions can now be generic
  – Lists of possible substituents
  – Substituents may be homology groups e.g. $C_{1-6}$ alkyl, heteroaryl
  – Ranges of values for repeated linkers
  – Constraints e.g. $R1$ is not $x$ if $R2$ is $y$
MARKUSH DEFINITION CAPTURE

• Map all variants of common Markush phrase definitions to one of the following:

<table>
<thead>
<tr>
<th>Phrase Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>rgroupAssignment*</td>
<td>“R1 is” “R5 selected from”</td>
</tr>
<tr>
<td>numberAssignment</td>
<td>“m is an integer from 1 to 4”</td>
</tr>
<tr>
<td>precedingGroupSubstituted</td>
<td>“optionally substituted by”</td>
</tr>
<tr>
<td>compoundVariants</td>
<td>“and stereoisomers”</td>
</tr>
<tr>
<td>markushGroup</td>
<td>“not present”, “double bond”</td>
</tr>
</tbody>
</table>

*Can be qualified to indicate that the R-groups are combined, e.g. to form a ring
wherein:

- $n$ is 1 or 2;
- $R^1$ is $-\text{C}(=\text{O})R^4$, $-\text{S}(=\text{O})_2R^4$, or $-\text{O}(\text{C}(=\text{O})_2\text{CH}_2\text{OH}$;
- $R^4$ is methyl or morpholine;
- $R^2$ is $\text{H}$ or $\text{F}$;
- $R^3$ is chloro or $\text{C}(\text{CH}_2)_2R^3$;
- $R^3$ is methyl, cyan, or hydroxymethyl;
- $X$ is $\text{CH}$, $\text{CH}_2$, or $\text{N}$;
- $X^2$ is $\text{CH}$ or $\text{N}$;
- $X^3$ is $\text{CH}$ or $\text{N}$; and
- $Y$ is $\text{CH}$ or $\text{O}$;

Name-to-structure: $\text{OC}^*$

Formula parsing: $\text{O(C}(\text{*})(\text{*)CO})[\text{*:1}]$ |$;$;R1&#39;_p;R1&#39;_p;;;_AP1$|
NAME-TO-STRUCTURE OF GENERIC CHEMICAL NAMES

• Modified version of, the chemical name-structure program, OPSIN supporting:
  – Positional variation
  – Homology groups
POSITIONAL VARIATION

Ar² is tetrazolyl, triazolyl, oxadiazolyl, thiadiazolyl, pyrazolyl, imidazolyl, oxazolyl, thiazolyl, isoxazolyl, isothiazolyl, furanyl, thienyl, pyrrolyl, pyrimidinyl, pyrazinyl, pyridinyl, hydroxypyridinyl, quinolinyl, isoquinolinyl, or indolyl

O*N1=C(C=CC=C1)*

|$;;;;;;;;_AP1$,m:1:4.5.6.7|
approach 3: ontogrep

• Form follows function... OntoGrep: Name=Search
• Classical approaches to handling Markush start by defining an intermediate mathematical/graph representation, and decomposing the problem in (1) matching these representations and (2) entering/capturing these representations.
• An artificial intelligence inspired approach is for the semantics of the search to be defined by natural language (text representations).
MOTIVATING EXAMPLES

• “Alleged” PubChem examples
  – zinc compounds
  – boronic acids
  – C6H12O
• Frequently text-mined entities
  – alkane
  – heterocycle
  – inorganic acid
  – solvent
Radioactive Smarts
Example OntoGrep Queries

- nitrogen containing heterocycles
- cationic ring systems
- cyclic alkanes
- branched acyclic alkanes
- transuranic elements
- zinc compounds
- binary compounds
- Lewis acids
- atropisomers
- polyspiro ring systems
- polyatomic elements
- inorganic salts

- radioactive compounds
- radioactive elements
- monocyclic ring systems
- neutral compounds
- uncharged compounds
- zwitterionic compounds
- carbon containing inorganics
- zinc oxides
- metal chlorides
- iron halides
- inorganic salts
PATENT TEXT MINING

• The classic example of pharmaceutical patent busting is the 2009 Bayer patent for Vardenafil (Levitra), entitled "2-phenyl substituted imidazotriazinones as phosphodiesterase inhibitors", US7696206B2.

• How much information can/could be mined from the title alone?
2-PHENYL SUBSTITUTED TRIAZINONES
IMIDAZOTRIAZINES

All 14 imidazotriazines
PISTACHIO: SIRI FOR CHEMISTS

30-40% yield Merck LiOH deprotections in the last six months indazole substructure
CONCLUSIONS

• Specific compounds can be reconstructed from a generic structure and specific R-group definitions
  – Immediately useful for key-compound extraction and compound-property relationship extraction

• Simple generic structural definitions can be captured, but many cases still too complex

• “Ontogrep” technology brings implicit Markush structure querying to high school student level.
ACKNOWLEDGEMENTS

• This work was made possible by funding from:
  – GlaxoSmithKline
  – Minesoft
  – NCBI, NIH
  – Novartis
  – Vertex Pharmaceuticals
Thank you for your time!

http://nextmovesoftware.com
http://nextmovesoftware.com/blog
roger@nextmovesoftware.com