Handing of homology variation in structure representation, patent Markush search, enumeration and visualization

R Wagner, S Creperti, N Máté, A Baharev, T Csizmazia and F Csizmadia; ChemAxon Ltd, Mármaros köz 3/a, 1037 Budapest, Hungary, rwagner@chemaxon.com

Introduction
Cheminformatics systems usually focus on handling specific molecules and reactions. However, generic (Markush) structures are also indispensable in various areas, like combinatorial library design or chemical patents for the description of compound classes.

What is a Markush structure
Markush structures describe a compound class by generic notation:
- Substitution variation (R-groups, atom and bond lists)
- Frequency variation (link nodes and repeating units)
- Position variation (variable point of attachment)
- Homology variation (e.g. alkyl, aryl)
- Conditions for generic features: occurrence lists, dependency, etc.

Please use the description of:

Cheminformatics systems usually focus on handling specific classes.

ChemAxon has been involved in research connected to Markush design or chemical patent classes.

The ChemAxon Markush project
ChemAxon has been involved in research connected to Markush structures for six years. ChemAxon tools enable drawing, visualization and enumeration of Markush structures as well as searching them in memory and database without enumerating the library members.

Current supported Markush features
The following generic features are available:
- R-groups (nesting, multiple attachments)
- Atom and bond lists
- Repeating units and link nodes
- Position variation
- Homology groups (alkyl, aryl, etc)

including conditions by properties

Collaboration with Thomson-Reuters
- Thomson-Reuters: content provider: Merged Markush Service (MMS) data, Derwent World Patent Index (DWPI) patent data, Derwent Chemistry Resource (DCR) Exemplified structures
- ChemAxon: Software provider

Classification of homology groups
Italics: groups handled with ChemAxon tools Parentheses: Thomson-Reuters name of groups.

1. Structural feature based
a) Cylics
  - Carbocyclic
    - Cycloalkyl (CVC)
    - Carboxylic (CAR)
    - Heterocyclic
      - Heterocylic (HET)
      - Heteromonoalicyclyc (HM)
    - Fused heterocyclic (FHET)
b) Acyclic carbon - carbon tree
  - Alkyl (CH)
  - Alkyn (CN)
  - Alkyl (CHY)

2. Defined groups: Can be expressed by a limited set of definitions (implemented as R-group definitions), the above homology groups can be used:
- Halogen (HAL)
  - Any (XX) – union of all other groups
- Protecting (PRT) – context sensitive definitions (nitro, alcohol, carboxy protecting groups.)
- Customization: Further groups may be specified by providing the R-group definitions. Context sensitive definitions: dependence on the context of the groups may be specified.

3. Matched by the given group only: Unknown (UNK), Fluorescent (DYE), Acyl (ACY)

Homology Properties
Additional homology properties refine the matching and enumeration behavior by restricting the represented structures:
- Monocyclic – fused
- Saturated – unsaturated
- Linear – branched
- Number of atoms: all together, ring atoms, by type, deuteriums
- Number of bonds by type

Searching homology groups
- Structural feature based groups: Any specific query fragment fulfilling the required criteria can match the given group provided that the structural context is appropriate
- Defined groups are searched based on their definitions similarly to R-groups.

Query-side support
Homology groups are supported on the query side for searching and chemical database tools to handle homology structures.

Enumeration
- Sampling the Markush space
- Typics: random, sequential
- Homology groups are enumerated using a sample set of substructures which both fulfill the properties of the group and are chemically reasonable. These sets were obtained by data mining a 100K drug-like database and taking the most frequently occurring fragments.

Validation
Enumerates of a given structure need to be found by the search process in the same Markush structure. Therefore searching the Markush with its enumerate is a suitable technique for measuring the correctness of searching. The following table shows results on Thomson-Reuters MMS patent Markush structures, with JChem version 5.5.

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target structures</td>
<td>8040</td>
<td></td>
</tr>
<tr>
<td>Enumerated structures</td>
<td>14798</td>
<td>18.4%</td>
</tr>
<tr>
<td>Found</td>
<td>14545</td>
<td>98.29%</td>
</tr>
<tr>
<td>Not found</td>
<td>253</td>
<td>1.7%</td>
</tr>
<tr>
<td>Markush structures: at least</td>
<td>98</td>
<td>0.66%</td>
</tr>
</tbody>
</table>

The average search time per record was 1.48 seconds. This is a search time for a memory search, which is executed on a single thread and includes import of the molecule files. Databases searches are expected to perform better.

Database execution time
- Query
  - Processor Core: 4
  - Markush structures: 14798
  - Hits: 34229
  - Execution time: 70 min

Future work
- Scale-up: to search the full patent Markush literature from Thomson-Reuters MMS efficiently: Speed-up search, database screening, computational clustering
- Improve correctness
- Further query features, for example full Markush-Markush search.
- Further visualization and analysis functions and tools for Markush Enumeration and Search

Summary
ChemAxon successfully extended its structure drawing, visualization and chemical database tools to handle homology structures. Work is in progress to speed-up searching and implement missing features.

Acknowledgments
We are grateful to our partners and clients for providing us valuable feedback and data sets for testing the programs.