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

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

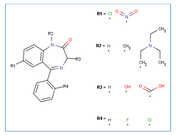
Introduction

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

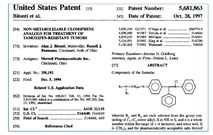
Markush structures describe a compound class by generic notation:
- Substitution variation (R-groups, atom and bond lists)
- Frequency variation (link nodes and repeating units)
- Position variation (variable point of attachment)
- Homology variation (e.g. alkyl, aryl)
- Conditions for generic features: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

They are used for the description of:



Combinatorial libraries

- Smaller libraries
- Usually simpler constructs:
  - R-groups
  - Link nodes
  - Atom lists



Patent claims

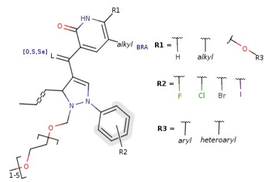
- The goal is as wide coverage as possible
- More sophisticated methods:
  - Homology variation (Alkyl, Aryl, etc)
  - Position variation
  - Etc.

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) *Cyclyl*
  - Carbocyclic
    - *Cycloalkyl (CYC)*
    - *Carboaryl (ARY)*
  - Heterocyclic
    - *Heteromonoalicyclyl (HET)*
    - *Heteromonoaryl (HEA)*
    - *Fused heterocyclyl (HEF)*
- *b) Acyclic carbon - carbon tree*
  - *Alkyl (CHK)*
  - *Alkenyl (CHE)*
  - *Alkynyl (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 tritiums, size of acyclic carbons, connecting atoms type
- Number of bonds by type

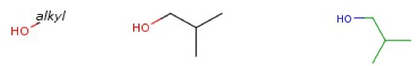
Searching homology groups

- 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 specific structures. From the Markush features homology groups are allowed on the target side.

e.g.: Query Target Hit (blue-specific, green homology)



Searching structural feature based groups

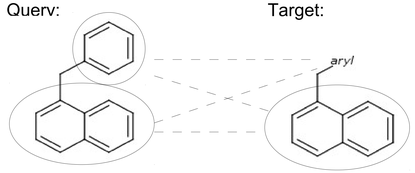
1. Initialization: Identification of all query fragments, that can possibly match the homology group.
2. Restricting the possible matchings: During the atom-by-atom matching algorithm the context may exclude some candidates. The remaining candidates are handled by the backtrack algorithm. Upon matching a query fragment to a target homology group, this fragment will be banned from matching other target atoms and other query fragments are banned from matching this homology group.

Example:

Dashed lines show possible matchings.

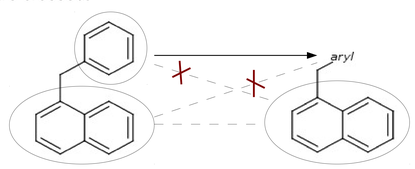
1. Initialization:

Querv: Target:



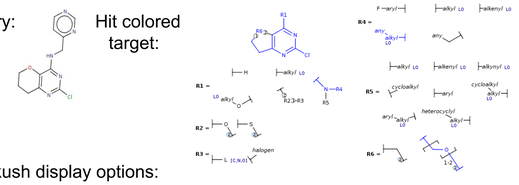
2. Restricting possible matchings:

Arrowed line shows fixed matching, banned matchings are crossed.

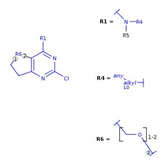
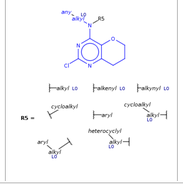
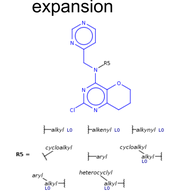


Hit Visualization Example

Query: Hit colored target:



Markush display options:

original Markush: (with unused rgroup definition removal option)	Markush reduction to hit	Markush reduction to hit plus homology expansion
		

Enumeration

- Sampling the Markush space
- Types: 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.

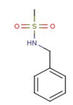
Validation

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:

	Number	Percentage
Target structures	8040	
Enumerated structures	14798	
Found	14545	98.29 %
Not found - from which known erroneous Markush structures: at least:	253 98	1.7% 0.66%

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. Database searches are expected to perform better.

Database execution time

Query: 

Processor Cores:	8
Markush structures:	695610
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 cluster
- 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.

ChemAxon