Latest Developments in Markush Representation, Search, Analysis and Visualization

R Wagner, S Csepregi, N Máté, A Baharev, T Csizmazia, D Deng, A Allardyce and F Csizmadia; ChemAxon Ltd, Záhony u. 7, 1031 Budapest, Hungary

Introduction

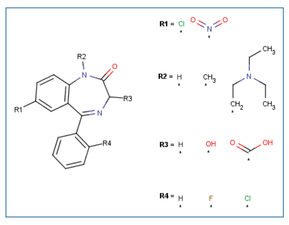
Cheminformatics systems usually focus on handling specific molecules and reactions. However, generic (Markush) structures are also indispensable in various areas, like combinatorial library design or chemical patents for the description of compound classes.

What is a Markush structure

Markush structures describe a compound class by generic notation:

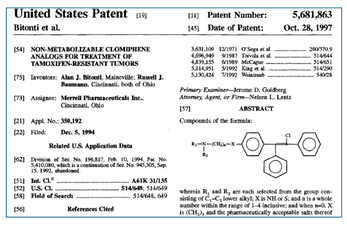
- Substitution variation (R-groups, atom and bond lists)
- Frequency variation (link nodes and repeating units)
- Position variation (variable point of attachment)
- Homology variation (e.g. alkyl, aryl)
- Conditions for generic features: occurrence lists, dependency, etc.

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.

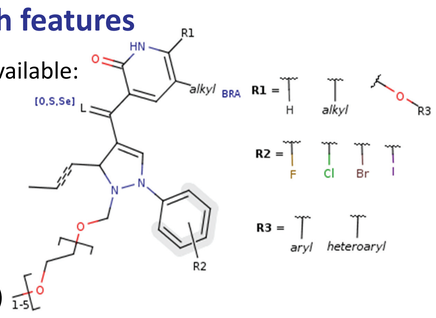
The ChemAxon Markush project

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

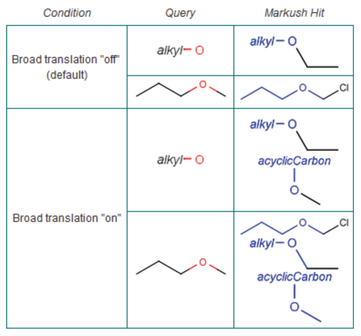
- Structural feature based groups: Any specific query fragment fulfilling the required criteria can match the given group provided that the structural context is appropriate
- Defined groups are searched based on their definitions similarly to R-groups.

Searching homology groups

Homology Broad Translation - As a search option, homology broad translation can be turned on or off;

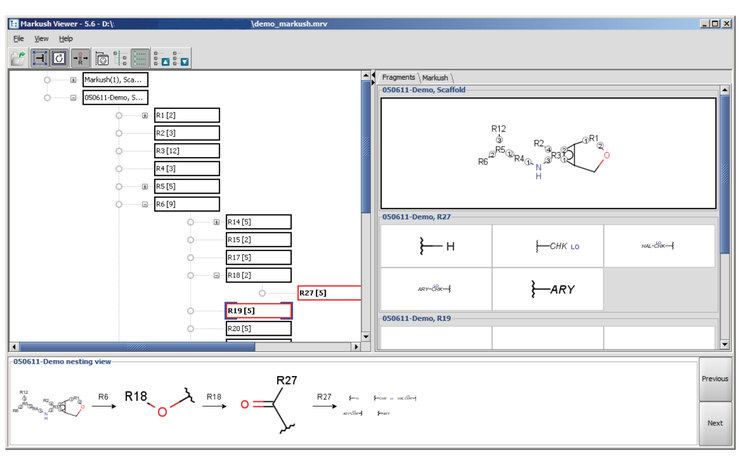
Off - query atom can match homology atom only if it is the same homology atom.

On - query atom matches homology group representing a larger set of structures. Homology atoms can also match homology atoms covering a larger set of structures.



Markush viewer

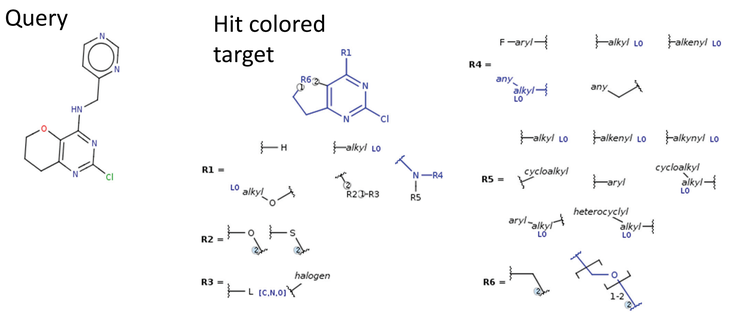
Markush Viewer is a desktop application of ChemAxon to view the various R-group definitions of Markush structures in an organized way. Markush Viewer represents the hierarchical nature of a Markush structure in graphical form. The structures of an opened Markush file will be identified, classified and organized as scaffolds and attached R-groups to help you follow each nested structure.



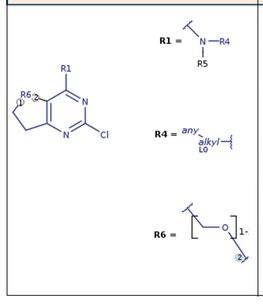
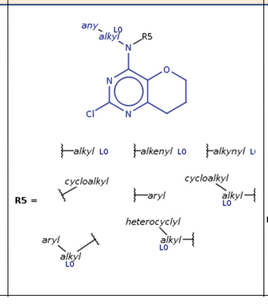
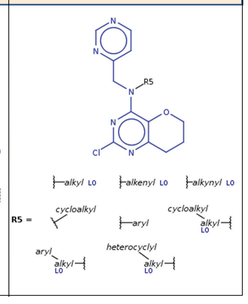
Hit Visualization Example

Query

Hit colored target



Markush display options:

original Markush: (with unused Rgroup Def. removal option)	Markush reduction to hit	Markush reduction to hit plus homology expansion
		

Markush as query

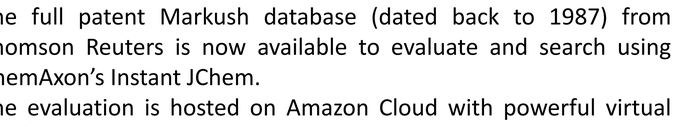
The query can also be a simple Markush structure. Different Markush variations: atom lists, bond lists, repeating units, position variation bond, homology group, R-groups, etc. are supported on the query side. Query atom properties are also supported: substitution count, ring bond count, hydrogen count, aromatic/aliphatic, …

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.:

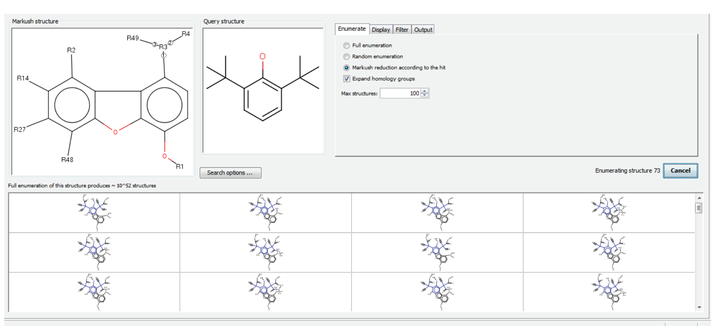
Query	Target	Hit (blue-specific, green homology)



Markush enumeration

Sampling the Markush space

- Types: full, random, sequential,
- Homology groups can be enumerated using a sample set of substructures.
- Enumerated structures can be aligned to the Markush core or to the query structure (Markush reduction). They can also be colored according to their R-groups or the query structures.
- A Chemical Terms filter can be applied to the enumerated structures to pick out specific structures (e.g. drug-like molecules).
- Enumerated structures can be displayed on the screen or exported to a file.



Full Patent Database Search

- The full patent Markush database (dated back to 1987) from Thomson Reuters is now available to evaluate and search using ChemAxon's Instant JChem.
- The evaluation is hosted on Amazon Cloud with powerful virtual machine, and secure connections.
- Useful features are implemented into the search interface:
  - Search both Markush and Exemplified structure databases
  - Export exemplified structures
  - Retrieve patent documents
  - Add notes to patents for easy review

Future work

- Improve search speed performance
- Further query features, e.g. homology broad translation in selected atoms; narrow translation; etc.
- 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