

2010 – a year of JChem

Sept 2010, US User Group Meeting

Szabolcs Csepregi



ChemAxon
Solutions for Cheminformatics

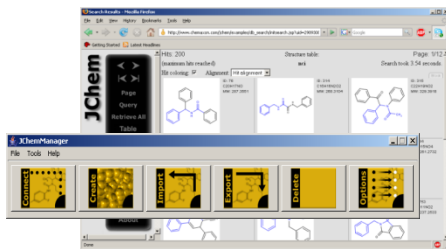
Contents

- ChemAxon chemical database tools
- Main features of JChem Base, Cartridge
- Example interfaces: JSP, ASP, AJAX examples
- Integration with other CXN products
- Recent developments, plans

The JChem DB family

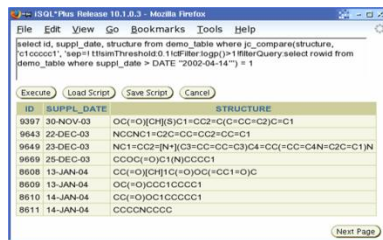
JChem Base

The chemical engine



JChem Cartridge

Oracle SQL integration

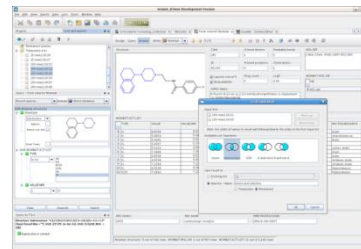


```
select id, suppl_date, structure from demo_table where jc_compare(structure, 'C1CCCC1', 'sup=1 (fromThreshold=0.1) $Filer (sup)=1') filter Query select rowid from demo_table where suppl_date > DATE '2002-04-14' = 1
```

ID	SUPPL_DATE	STRUCTURE
9397	30-NOV-03	CC(O)(CH)S(C)1=CC2=C(C(=CC=C2)C=C1
9643	22-DEC-03	NC(=O)1C2=CC=CC=C2C(=O)C1
9649	23-DEC-03	NC1=CC2=NC(=C)C=CC(=C3)C4=CC(=CC=C4N=C3)N1
9689	25-DEC-03	CCOC(=O)C1(N)CCCC1
8608	13-JAN-04	CC(O)(CH)1C(O)OC(=CC1=O)C
8609	13-JAN-04	CC(O)COC1CCCC1
8610	14-JAN-04	CC(O)COC1CCCC1
8611	14-JAN-04	CCCCCCCC

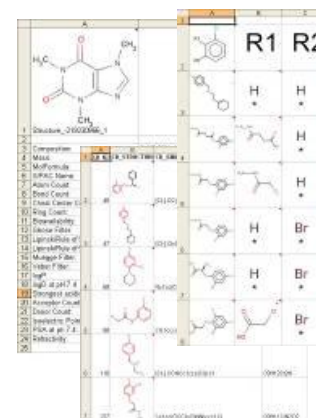
Instant JChem

Desktop application for scientist



JChem 4 Excel

MS Excel integration



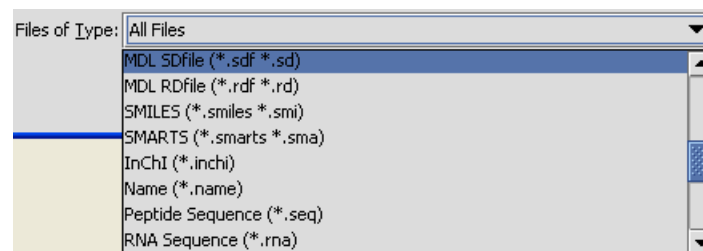
Compatibility and interfaces

Chemical file formats:

- MDL mol/rxn/sdf/rdf (v2.0 and v3.0), SKC, CDX, CDXML
- Smiles
- CML, MRV (Marvin)
- IUPAC and traditional names
- InChI, mol2, PDB, etc.

New New New

Markush DARC New



Database engines:

- **Oracle**, **MySQL**, MS SQL Server, MS Access, PostgreSQL, IBM DB2, **Derby**, **Composite**, etc.

New

All operating systems through:

- Java API (JChem Base)
- .NET API (JChem Base)
- SQL (Cartridge)
- JChem Web Services



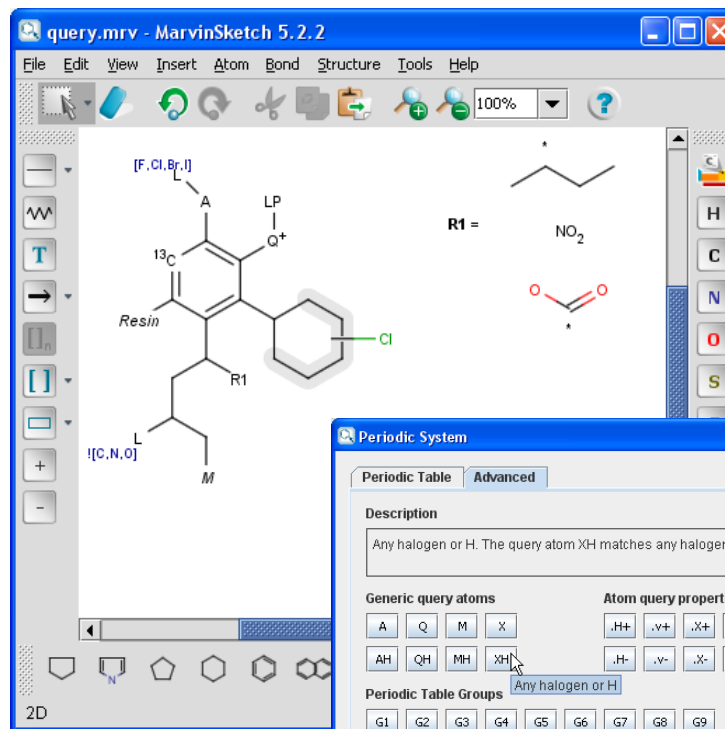
ORACLE®



Structure searching: features

- Substructure, Similarity, Full, etc. search types
- Wide range of query atoms
- Reaction search features
- Polymers
- Position variation
- Hit coloring
- R-group decomposition
- Homology groups **New**
- Full structure search with R atoms **New**

...



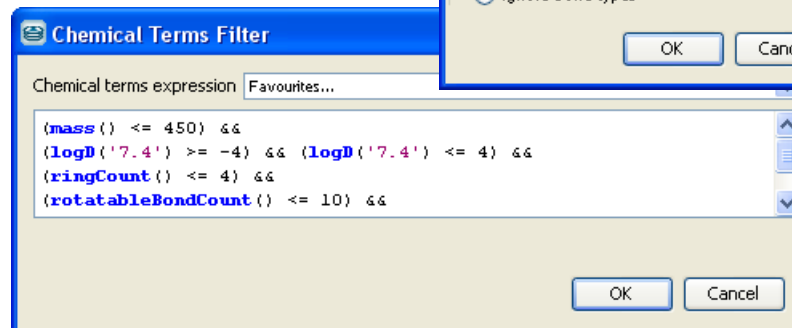
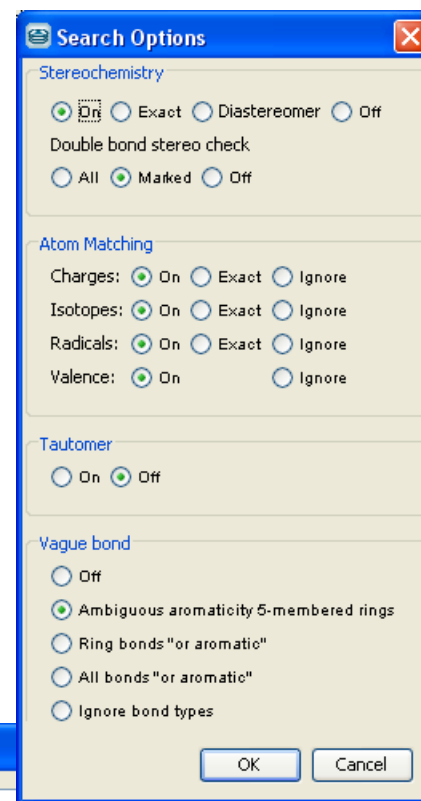
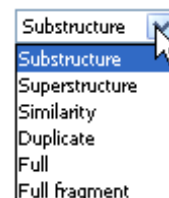
The 'Periodic System' dialog box is open, showing the 'Advanced' tab. The 'Description' field contains the text: 'Any halogen or H. The query atom XH matches any halogen atom or hydrogen.' The 'Generic query atoms' section includes buttons for A, Q, M, X, AH, QH, MH, and XH. The 'Atom query properties' section includes buttons for .H+, .v+, .X+, .R+, .r+, .rb+, .s+, .h+, .D+, .u, .H-, .v-, .X-, .R-, .r-, .rb-, .s-, .h-, .D-, .a/A. The 'Periodic Table Groups' section includes buttons for G1 through G18. The 'Special nodes' section includes buttons for LP, Pol, and *. The 'R-groups' section includes buttons for R1 through R32. The 'Custom Property' section includes buttons for Type: R-group, Alias, Pseudo, SMARTS, and Value. The 'Value' field is empty. A 'Close' button is located at the bottom right of the dialog box.

www.chemaxon.com/conf/Structural_Search.ppt

Structure searching: options

Some selected structure search options:

- Stereo on/off
- Ignore charge/isotope/radical/valence/polymers, etc.
- Vague bond matching options
- Chemical Terms filter
- Tautomer search
- Inverse hit list
- Maximum search time / number of hits
- Combine with non-structure conditions
- Ordering of results
- etc.

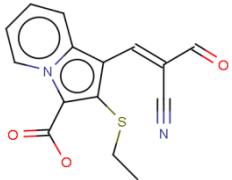
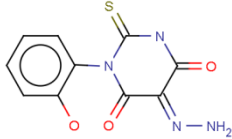
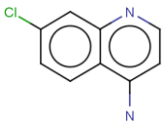
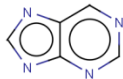


Structure search: performance

Compound registration:

Number of compounds	Elapsed time	
	Duplicates not checked	Duplicates checked
10,000	21 s	25 s
100,000	2 min 4 s	2 min 34 s
200,000	4 min 24 s	5 min 13 s

Substructure search in PubChem (19.5 million compounds):

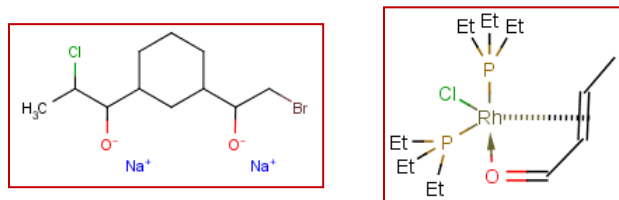
Query	Number of hits	Search time
	2	0.91 s
	93	0.98 s
	6,001	1.30 s
	146,256	5.66 s

JChem Base 5.2.2,
Intel Quad Q6600 2.4GHz,
8GB RAM; Oracle 10.2.0.3

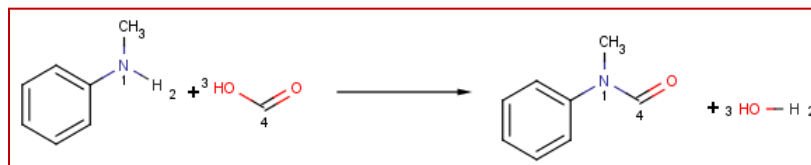
Table types

Control allowed chemical structures and available operations

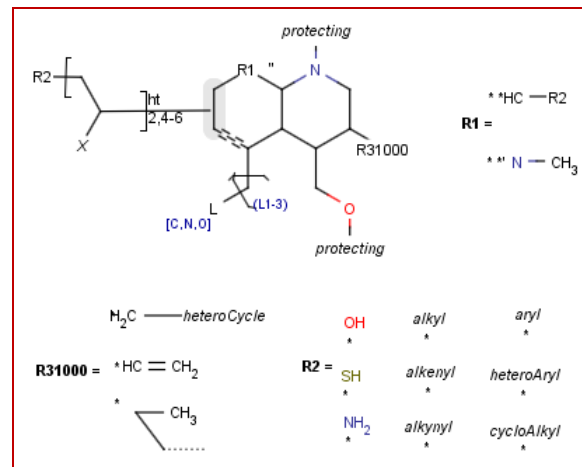
- Molecule



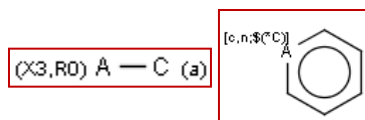
- Reaction



- Markush



- Query



- Any structure

Example web applications

Java Server Pages (JSP) example

- Marvin applets are used for query drawing and structure visualization

The screenshot displays the ChemAxon web services interface. On the left, the 'Query Parameters' window shows search settings: Search type: Substructure, Similarity type: Chemical Hashed Fingerprint, Dissimilarity threshold: 0.1, Max. hits: 200, Max. time: 3 min, and Return prev. results: No. The main window shows search results for 'nci' with a grid of chemical structures and their IDs and molecular weights. A table below the grid lists the results:

No.	Structure	ID	Formula	Molweight
1		1	C ₇ H ₆ O ₂	122.1213
2		2	C ₁₄ H ₈ N ₂ S ₄	332.487
3		3	C ₆ H ₃ ClN ₂ O ₅	218.551
4		4	C ₇ H ₃ N ₃ O ₂ S	145.14

AJAX example

- Back-end is JChem Web Services
- No Java is needed for browsing

Integration – other ChemAxon tools

- Standardizer — customizable chemical representation
- Calculator plugins — properties by Chemical Terms
Calculated columns
- Screen — alternative similarity types and metrics
- Tautomer support:
 - Tautomer search – duplicate or SSS
 - Tautomer duplicate filter option
- Marvin — Query drawing and structure visualization
Provides the most consistent interface and back-end.

Integration – Cartridge extras

- JChem index (arbitrary table structure)
- Communication with Oracle optimizer
- Reaction based enumeration (Reactor)
- Format conversions – image generation also
- Markush enumeration (Calculator plugins)
- Property predictions through Chemical Terms (Calculator plugins)
- Indexing of user fingerprints **New**

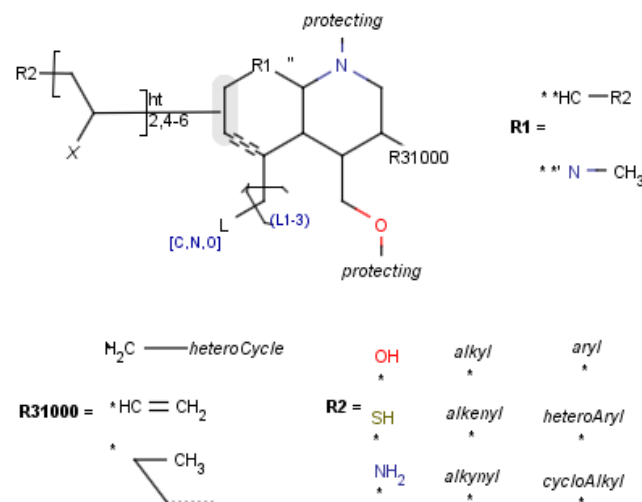
Markush structures

Markush structure registration and search

- Markush features

- R-groups
- Atom lists, bond lists
- Position variation bond
- Link nodes and repeating units
- Homology groups

- Compatible enumeration plugin



What's new

What's new: JChem Base

5.3.X

- Homology variation in queries (alkyl, aryl, etc.)
- R-group decomposition API integrated with JChemSearch
- Markush output of R-group decomposition
- Improved undefined R-atom matching feature
- New vague bond level options (improvements in handling fused aromatic ring systems)
- Reduced memory need for ABAS

- Pre-regeneration for less downtime at version updates
- JChem Manager GUI improvements (Standardizer GUI, MD)
- New method of handling log tables and register caches, to improve cache update mechanism on batch loading.

What's new: Cartridge-specific

5.3.X

- Markush tables and indexes
- User-defined fingerprints for similarity search
- Screen molecular descriptors for similarity search
- Increased security: encrypted passwords and usernames
- Improved exception handling (e.g. file format errors)

Under development

Plans: JChem Base & Cartridge

JChem Base

- Multi-threading of fingerprint operations (screening, similarity)
- Similarity search results visualization: MCS highlighting
- Symmetry handling in R-group decomposition
- Homology properties (alkyl C1-6, heteroaromatic N1-2, etc.)
- Maximum Common Substructure search type
- R-group decomposition on GUI-s
- Arbitrary table structure (JChem index table)
- JChem Server, JChem grid

5.4

Cartridge

- R-group decomposition
- Further search speed improvements (SSS, similarity)

Example web applications

Query drawing

Hit alignment,
coloring

Search types,
options

New features

The image displays two screenshots of the JChem web application interface. The top screenshot, titled "Search Results - Mozilla Firefox", shows a search results page for the query "nci". It displays 200 hits (maximum reached) and a search time of 3.54 seconds. The results are presented in a table with columns for ID, Formula, and Molecular Weight (MW). Several chemical structures are shown as hits, including benzamide derivatives. The bottom screenshot, titled "Query Parameters - Mozilla Firefox", shows the query configuration page for the same query. It includes a "main options" section with fields for Search type (Substructure), Similarity type (Chemical Hashed Fingerprint), Screening config (Default), Dissimilarity threshold (0.1), Max. hits (200), Max. time (3 min.), and Search prev. results (No). The "advanced options" section contains a "Conditions" table with fields for Id, Formula, Molweight, CD_HASH, CD_FLAGS, and CD_SORTABLE_FORMULA. A "Chemical Terms filter" section at the bottom contains a text area with the following filter rules:

```
(mass() <= 500) &&  
(logP() <= 5) &&  
(donorCount() <= 5) &&  
(acceptorCount() <= 10)
```

Summary

- JChem Base and Cartridge are comprehensive and efficient
- Good team players – open to integration and extensions
- Continuous development, improvements in the pipeline