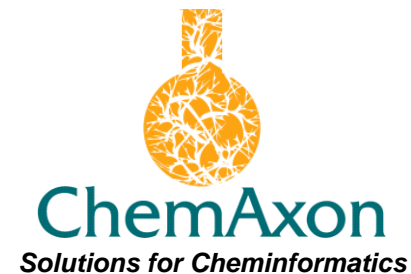




# JChem enters Formula 1 (and formula search enters JChem)

May 2011, European User Group Meeting

**Szabolcs Csepregi**



- ChemAxon chemical database tools
- Quick intro to JChem back end
  - JChem Base
  - Cartridge
  - Web Services
  - Markush search
- Recent developments, plans

# The JChem family – back & front end

## Visualization



### Marvin

Structure, query & reaction editor, viewer & visualization

## Property Prediction



### Calculator Plugins

Structure property prediction & calculation

#### Selected calculations listing

- pKa, Major microspecies
- logP, logD
- Charge
- Tautomerization
- Stereoisomer
- Conformation and 3D alignment
- Topology Analysis
- Molecular Surface Area
- Markush Enumeration
- Hydrogen bond donor/ acceptor
- Structural Frameworks
- Structure to Name
- etc...

## Chemical DB – toolkit



### JChem Cartridge

JChem/Oracle integration



### JChem Base

Structure searching & db access



### Standardizer

Chemical business rules processing



### Structure checker

Batch structure file validation and correction

## Chemical DB – desktop



### Instant JChem

Structure db management, search & prediction



### JChem for Excel

Enabling chemistry in Excel

## Nomenclature



### Name to Structure

Import & search chemical names

## Enumeration



### Reactor

Enumeration via reaction modelling

## Library analysis



### JKluster

Clustering & diversity analysis



### Fragmenter

Decomposition to fragments and R-groups



### Screen

HT pharmacophore screening

## Add-ons



### Markush Search

Store & search Markush structures



### JChem Webservices

Web services integration interface

### JChem for SharePoint

Live chemistry, search and structure based properties

New

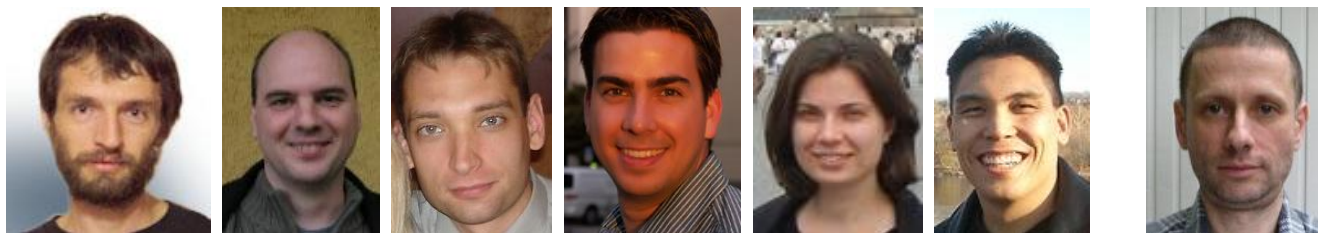
# Development team – back end

---

## JChem Search Engine and Markush



## JChem Base, Web Services and Cartridge

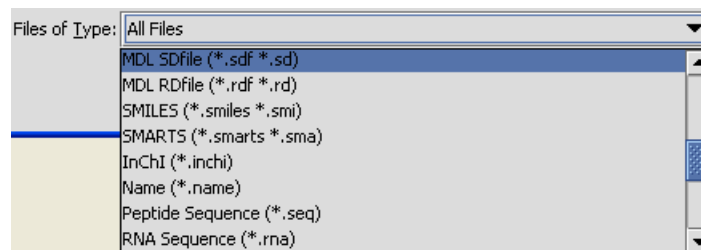


# Compatibility and interfaces

## 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.

Markush DARC



## Database engines:

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

## All operating systems:

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

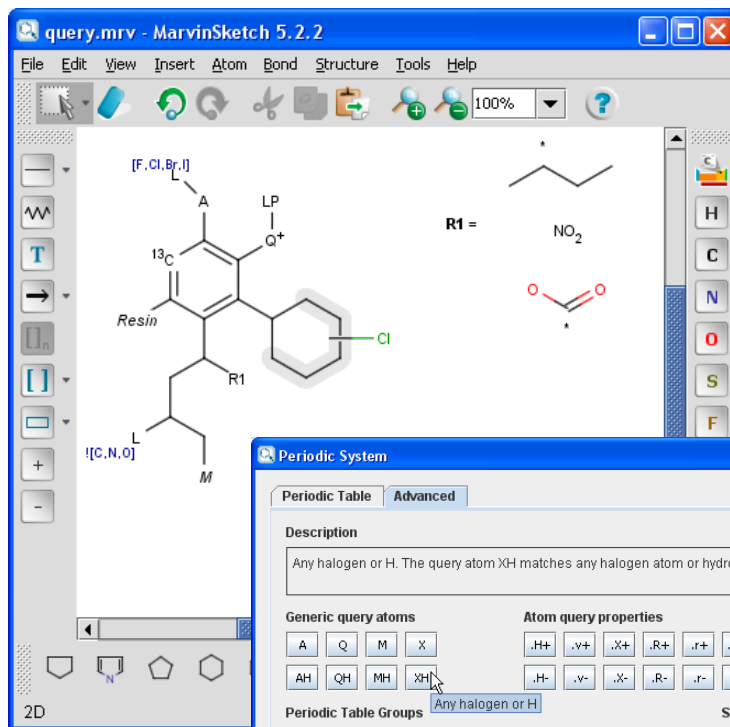


ORACLE®



# Structure search: features

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



The screenshot shows the 'Periodic System' dialog box, which is used for defining query atoms and R-groups. It has two tabs: 'Periodic Table' and 'Advanced'. The 'Advanced' tab is selected, showing the following sections:

- Description:** A text area containing the description: "Any halogen or H. The query atom XH matches any halogen atom or hydrogen."
- Generic query atoms:** A grid of buttons for A, Q, M, X, AH, QH, MH, and XH. The XH button is highlighted with a tooltip that says "Any halogen or H".
- Atom query properties:** A grid of buttons for .H+, .v+, .X+, .R+, .r+, .rb+, .s+, .h+, .D+, .u, .H-, .v-, .X-, .R-, .r-, .rb-, .s-, .h-, .D-, and .ajA.
- Periodic Table Groups:** A grid of buttons for G1 through G18.
- Special nodes:** Buttons for LP, Pol, and \*.
- R-groups:** A grid of buttons for R1 through R32.
- Custom Property:** A section with a 'Type' dropdown (set to 'R-group') and buttons for Alias, Pseudo, SMARTS, and Value. Below it is a 'Value' text input field.

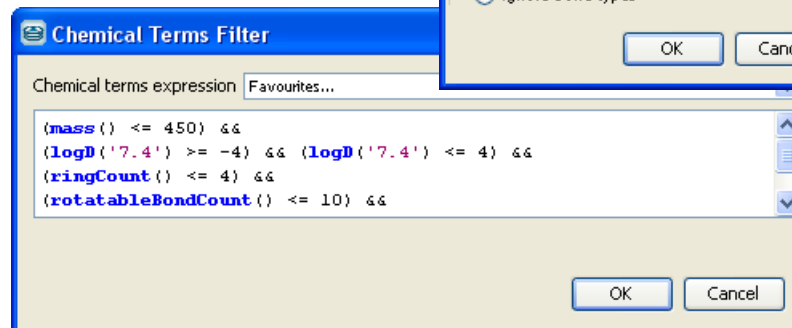
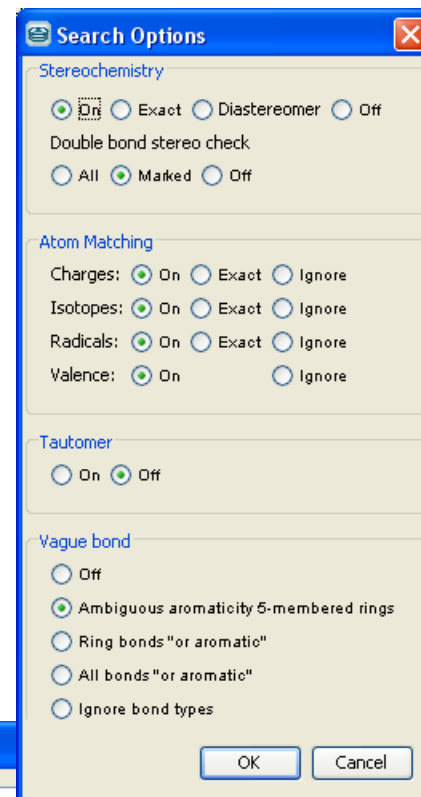
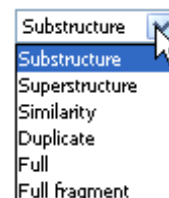
A 'Close' button is located at the bottom right of the dialog box.

[www.chemaxon.com/conf/Structural\\_Search.ppt](http://www.chemaxon.com/conf/Structural_Search.ppt)

# Structure search: 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.

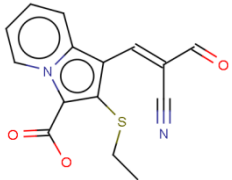
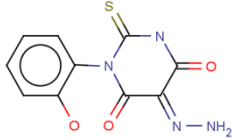
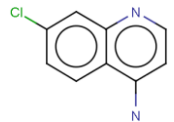
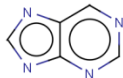


# Structure search: performance

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):

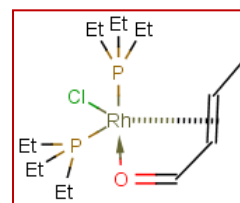
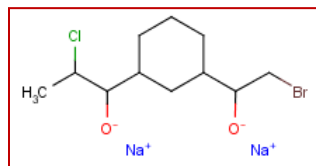
Query	Number of hits	Search time
	2	0.91 s
	93	0.98 s
	6,001	1.30 s
	146,256	5.66 s

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

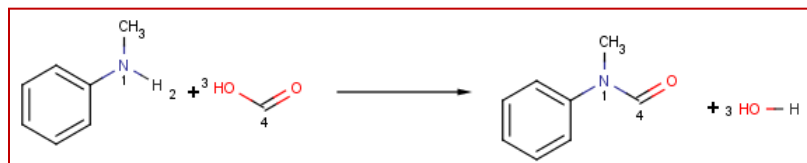
# Table types

## Control allowed chemical structures and available operations

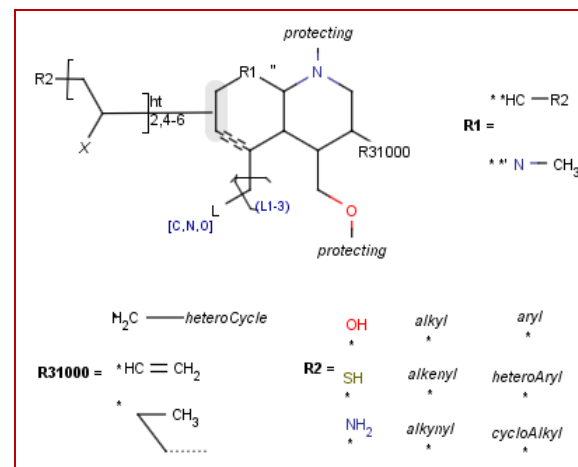
- Molecule



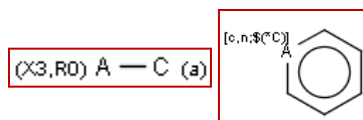
- Reaction



- 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, Screening config: Default, Dissimilarity threshold: 0.1, Max hits: 200, Max time: 3 min, Search prev. results: No, Return non-hit: 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 <sub>7</sub> H <sub>6</sub> O <sub>2</sub>	122.1213
2		2	C <sub>14</sub> H <sub>8</sub> N <sub>2</sub> S <sub>4</sub>	332.487
3		3	C <sub>6</sub> H <sub>3</sub> ClN <sub>2</sub> O <sub>5</sub>	218.551
4		4	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub> O <sub>2</sub> S	145.14

## AJAX example

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

The screenshot shows the 'Ajax interface to ChemAxon web services' window. It features a navigation menu on the left with options: Query, Export, Print, and About. The main area displays search results for 'nci' with a table of results:

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

# 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 to back end





# Integration – Cartridge & WS extras

---

## JChem Cartridge:

- JChem index (arbitrary table structure)
- Indexing of user fingerprints
- Communication with Oracle optimizer

## JChem Cartridge & Web Services:

-  Reaction based enumeration (Reactor)
-  Format conversions – image generation also
-  Markush enumeration (Calculator plugins)
-  Property predictions through Chemical Terms (Calculator plugins)

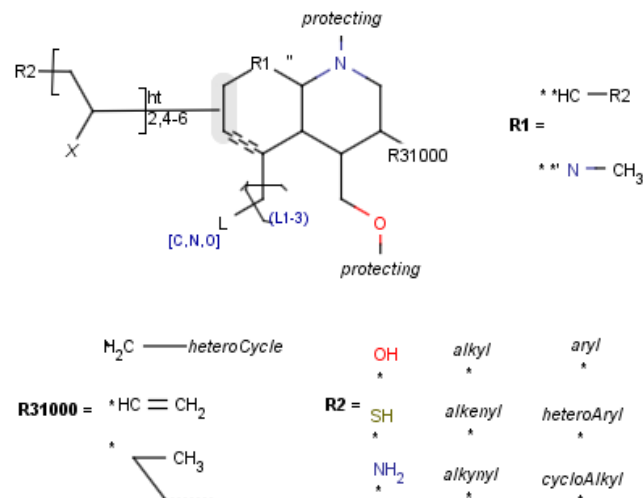
# Markush structures

## 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

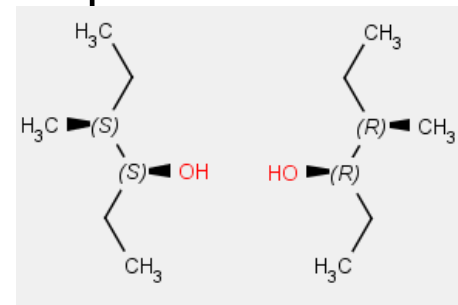
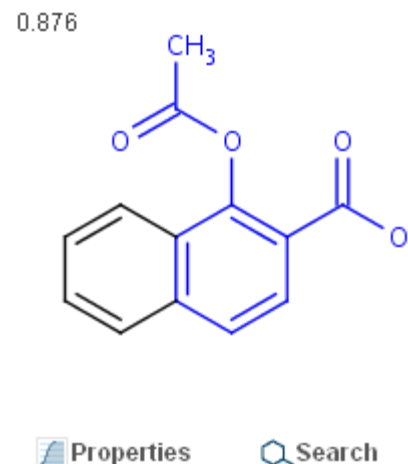
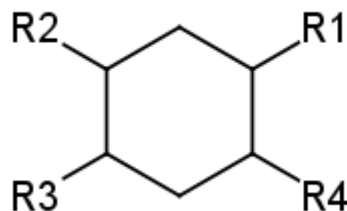


**What's new**

# What's new: Search engine

## 5.4

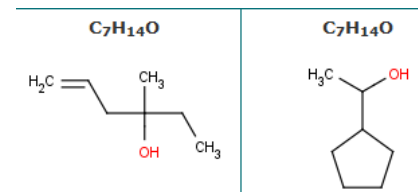
- Hit visualization of similarity search results using MCS
- More consistent R-tables for symmetrical scaffolds.
- Multi-threading enhancements:  
faster first results and similarity search
- Non-tautomer duplicate search on tautomer duplicate tables.
- Enantiomer stereo search option
- ECFP & FCFP in similarity search (with JChem Screen)



# What's new: Search engine

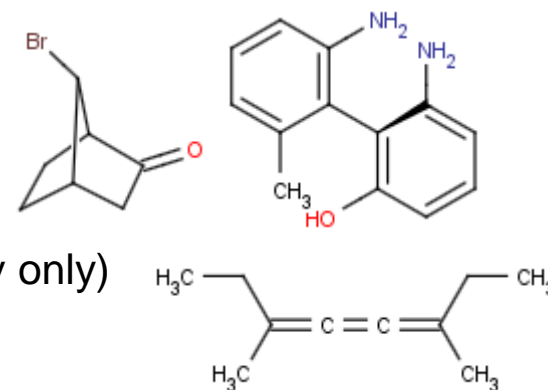
## 5.5 Just released!

- Sophisticated formula search (intervals, excluded atom type, components, polymers, isotopes, coefficients, combinatorial groups, etc.)

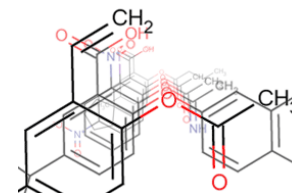


- New stereo types support:

- Syn, anti
- Cis/trans of cumulenes (in memory only)
- Axial: atrop, allenes and cumulenes (in memory only)



- Faster R-group search



# What's new: JChem Base, WS

## 5.4

- Duplicate filtering table option

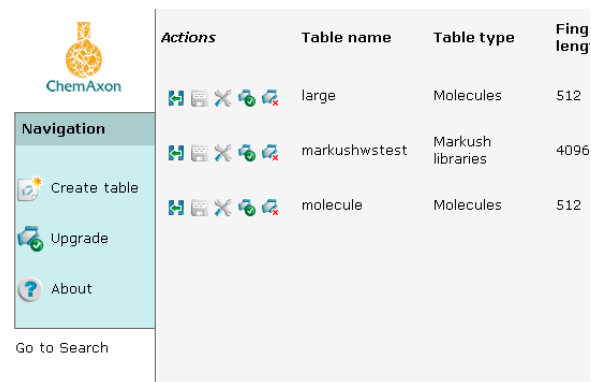


Filter out duplicate structures

- JChem Manager in AJAX & JChem Web Services

- New Web Services:

- Molecule search in lists
- Retrieve or export related table data
- Markush search and enumeration
- Batch insert & delete in JChem table
- Batch Chemical Terms evaluations



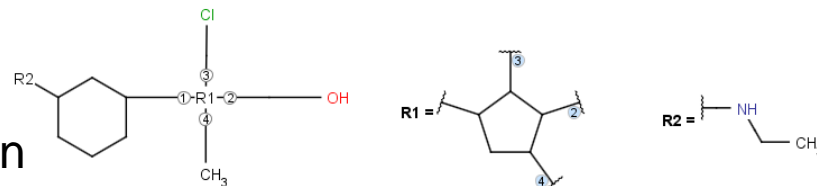
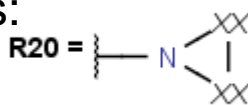
The screenshot shows the JChem Manager interface. On the left is a navigation sidebar with the ChemAxon logo and buttons for 'Create table', 'Upgrade', and 'About'. Below the sidebar is a 'Go to Search' button. The main area displays a table with the following data:

Actions	Table name	Table type	Fing leng
	large	Molecules	512
	markushwstest	Markush libraries	4096
	molecule	Molecules	512

# What's new: Markush

## 5.4

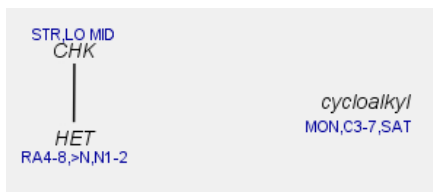
- New attachment point representation
- New homology groups:
  - Cyclyl,
  - RingSegment (XX)
- Properties for homology groups



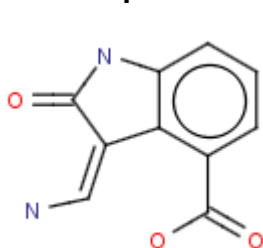
Edit properties of atom (cycloalkyl)

Property key	Property value
Deuterium or Tritium Count	
Ring Type	Monocyclic
Saturation	Not Specified
Additional Text Notes	
<Type new property name here>	

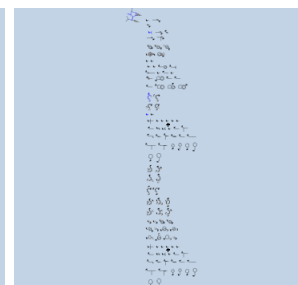
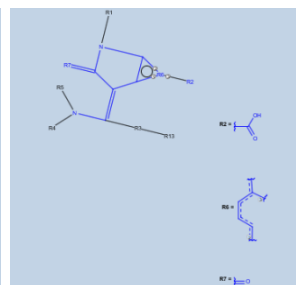
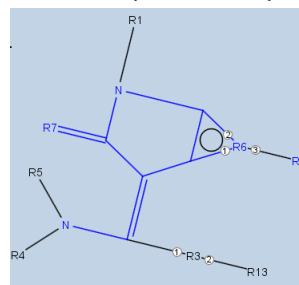
Ok Cancel



- IJC: GUI improvements (enumeration window, filtering, homology, etc.)
- Improved R-group hit visualization (IJC: 5.5)



No R-groups  
Relevant R-groups  
R-groups





**Under development**

# Plans: Markush search

---

- Further speed improvements – scale up to full MMS searching efficiently
- New query features:
  - Atomic query properties:
    - Atomic counts: substitution, ring and ring bond, H count,
    - unsaturation,
    - free sites, etc.
  - Full support of explicit H in queries
  - Option to switch on / off translation (of homology groups), equal translation
- Relevancy ranking of search hits

# Plans: JChem Base, Cartridge & WS

---

- Computational cluster / grid solution
- Maximum Common Substructure search type
- R-group decomposition on GUI-s & Cartridge
- R-group decomposition: pivot layout
- Arbitrary table structure (JChem index table) API support
- „One-click” installer for Cartridge

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 shows the search results page for a query. The URL is [http://www.chemaxon.com/jchem/examples/db\\_search/intsearch.jsp?uid=298930](http://www.chemaxon.com/jchem/examples/db_search/intsearch.jsp?uid=298930). The page shows 200 hits (maximum hits reached) for the query 'nci'. The search took 3.54 seconds. The results are displayed in a table with columns for ID, Formula, and MW. The first three hits are:

ID	Formula	MW
76	C20H17NO	287.2551
314	C16H16N2O2	268.3104
315	C22H19NO2	329.3918

The bottom screenshot shows the query parameters page for the same query. The URL is [http://www.chemaxon.com/jchem/examples/db\\_search/query.jsp?uid=29893](http://www.chemaxon.com/jchem/examples/db_search/query.jsp?uid=29893). The page shows the query structure table 'nci' and the query parameters. The query parameters are:

- Search type: Substructure
- Similarity type: Chemical Hashed Fingerprint
- Screening config: Default
- Disimilarity threshold: 0.1
- Max. hits: 200
- Max. time: 3 min.
- Search prev. results: No
- Return non-hits (inverse result set):

The query parameters page also includes a 'main options' section and a 'Conditions' section. The 'Conditions' section includes a table with columns for Id, Formula, Molweight, CD\_HASH, CD\_FLAGS, and CD\_SORTABLE\_FORMULA. The 'Chemical Terms' filter is set to 'Select filter'.

# Summary

---

- JChem back end is comprehensive and efficient
- Good team player – open to integration and extensions
- Continuous development, improvements in the pipeline