

Advanced structural searching using ChemAxon tools

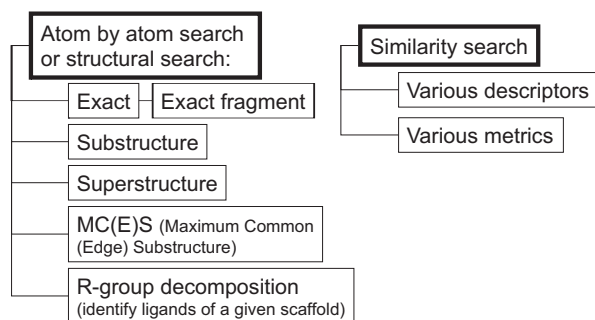
Szabolcs Csepregi*, Szilárd Dóránt, Nóra Máté, Miklós Vargyas, Péter Kovács, György Pirok, Ferenc Csizmadia
ChemAxon Ltd. 1037 Budapest, Máramaros köz 3/a, Hungary. *corresponding author, e-mail: scsepregi@chemaxon.com

Introduction

Molecular search techniques are invaluable tools in all cheminformatics systems including rational drug design, compound registration systems and laboratory information management systems. Often they provide a basis for more complex applications like functional group identification, bond cleavage, virtual reaction processing, standardization, toxic fragment identification, etc.

JChem, one of ChemAxon's major suites of programs, provides a very rich set of features related to structural search.

Search types in JChem

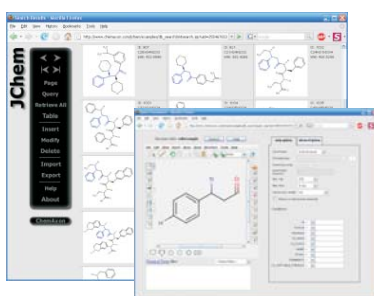


Structural search

Interfaces and Options

Interfaces

- JSP (Java Server Pages) web GUI for database
 - Similarity & structural search
 - Substructure highlighting
 - Additional constraints
 - Insert, modify, delete
- Command line utility: jcsearch: for files and DB
- Java API (programming interface)
 - isMatching() – Only to check matching
 - findFirst(), findNext() } Enumerate all possible matchings
 - findAll()
- Cartridge: access all functionality from SQL
- Chemical Terms



General options

- Order-sensitive hits
- Pre-assignment of query and target atoms
- Consider stereo or not, absolute stereo (ignore chiral flag)
- Timeout limit
- Exact charge/radical/isotope/query features/bond/stereo matching
- Double bond stereo: no check/checked/all double bonds
- Chemical Terms filter expression
- etc

Database search:

- Maximum search time/number of hits
- Additional SQL SELECT expression for prefiltering
- Output table
- Reverse hits mode (retrieve non-hits)

Database solutions

JChem Base provides a high-performance Java and JSP implementation, while JChem Cartridge offers a broadly accessible SQL interface through the Oracle® database.

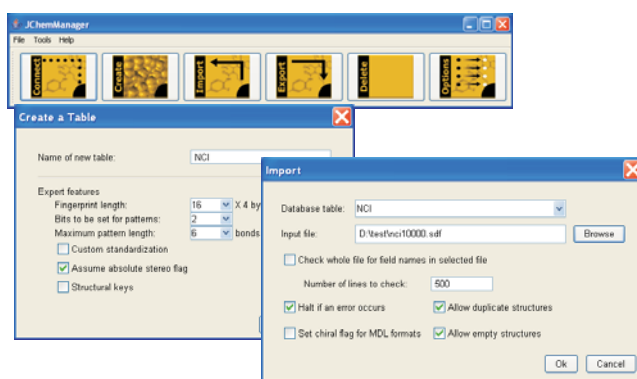
JChem has a two stage method for searching:

- Rapid pre-screening based on chemical hashed fingerprints (You can add your own definition of structural keys for even better performance.)
- Atom by atom search

Duplicate check at compound registration has another solution:

- Hash code: rapid primary filter
- Atom by atom search

Caching of structures and fingerprints allow top performance



Formats and platforms:

Supported formats:

- SMILES/SMARTS
- MDL molfile (v2000 and v3000)
- MDL SDF
- RXN
- RDF
- MRV
- CML
- PDB
- Sybyl molfile
- XYZ
- Image formats for export (JPG, PNG, SVG) etc.

Database engines:

- Oracle
- MySQL
- MS SQL Server
- PostgreSQL
- MS Access
- DB2
- etc.

OS:

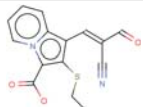
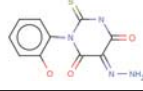
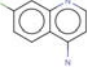

- any operating systems running Java:
 - Windows
 - Linux
 - Mac OS X
 - Solaris
 - etc.

JChem Base performance

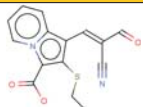
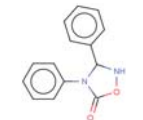

Compound registration

Number of compounds	Elapsed time	
	Duplicates not checked	Duplicates checked
10,000	22 s	35 s
100,000	2 min 33 s	4 min 16 s
200,000	4 min 53 s	8 min 19 s

Substructure search in a table of 3 million compounds:

Query	Number of hits	Search time (s)
	12	0.2
	936	0.4
	4,608	0.7
	65,208	5.6

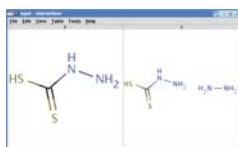
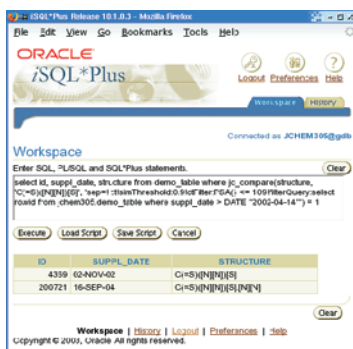
Similarity search: Tanimoto >0.8

Query	Number of hits	Search time (s)
	24	1.6
	156	1.6
	336	1.6

Server parameters: Windows XP; Athlon X2 2.6GHz desktop PC, 4GB RAM; Oracle 9i.

JChem Cartridge for Oracle

Oracle is extended to support chemical database operations using JChem Cartridge for Oracle



SQL Examples:

Substructure search displaying ID, SMILES codes, and molweight:

```
SELECT cd_id, cd_smiles, cd_molweight
FROM my_structures
WHERE jc_contains(cd_smiles,
  'CC(=O)Oc1ccccc1C(O)=O') = 1;
```

Similarity search filtered with predicted pK_a values, which displays predicted $\log P$ and $\log D$ values:

```
SELECT cd_id, jc_evaluate( cd_smiles, 'logP' ) ,
  jc_evaluate( cd_smiles, 'logD(7.4)' ),
  jc_evaluate( cd_smiles, 'pKa' )
FROM my_structures
WHERE
  jc_tanimoto(cd_smiles, 'CC(=O)Oc1ccccc1C(O)=O')
  >= 0.8;
```

The Chemical Terms language enhances Cartridge usage:

- New interface to ChemAxon API features from SQL - accessible from non-Java programs.
- Enhanced performance of certain SQL queries.

SQL Examples using Chemical Terms:

Number of compounds in table "nci_10m" containing 3-bromoindole and conforming the Lipinski rule of 5:

```
SELECT count(*)
FROM nci_10m
WHERE jc_compare(structure, 'Br1c1nc2ccccc12',
  `sep=! t:s!ctFilter:(mass() <= 500)
  && (logP() <= 5) && (donorCount() <= 5)
  && (acceptorCount() <= 10)') = 1;
```

Compounds in table "nci_10m" containing guanine and restricting TPSA, molecular weight, rotatable bond and aromatic ring counts:

```
SELECT cd_structure
FROM nci_10m
WHERE jc_compare(structure,
  '[#7]C1=NC2=C(N=CN2)C(=O)N1',
  `sep=! t:s!ctFilter:(PSA() <= 200)
  && (rotatableBondCount() <= 10)
  && (mass() <= 500)
  && (aromaticRingCount() <= 4) ') = 1;
```

Query features

Structural search in JChem handles a wide range of query features for atoms, bonds and molecules. JChem provides full SMARTS® substructure search support. The supported range of query features are detailed below.

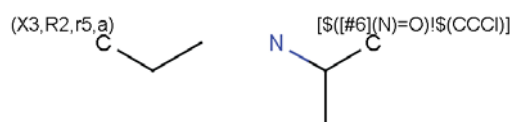
Atom query features

- Query atom types: any, hetero, list, not list
- Pseudo atoms e.g. "Resin"
- Explicit lone pairs (also matches to implied lone pairs)
- Query features:

Symbol	Description
H<n>	Total hydrogen count
a	Aromatic
A	Aliphatic
R<n>	Ring count in SSSR
r<n>	Ring size in SSSR
v<n>	Valence
X<n>	Connectivity
D<n>	Degree
h<n>	Implicit H count
rb<n>	rb* Ring bond count *: as drawn
s<n>	s* Substitution count *: as drawn
u	Unsaturated atom
The following only in SMARTS atoms:	
& ; , !	Logical operators
\$(<smarts>)	Recursive SMARTS
+0, -0	Zero charge

- Charge, isotope, radical
- SMARTS® atoms: all query features supported
- Link nodes

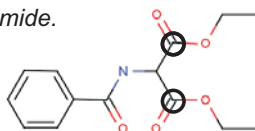
Examples of query and SMARTS® atoms:



Example: Carbonyl C, but not amide.

SMARTS® query:

[\$ (C=O) ! \$ (C(N)=O)]



Bond and fragment query features

- Query bond types:

- Any,
- single or double,
- single or aromatic,
- double or aromatic

- Bond topology



- SMARTS® bonds:

Symbol	Description
- = #	Single, double, triple
:	aromatic
& , ; !	Logical operators
@	Ring bond
/ \ / ? \ ?	Directional bond (cis/trans)

- Component level grouping:

Symbol	Description
(C.C)	Same component
(C).(C)	Different component
C.C	No component restrictions

E/Z Double bond stereo searching

- Levels of check:

- All
- Only marked double bonds (MDL terminology: stereo care flag)
- No double bond stereo check



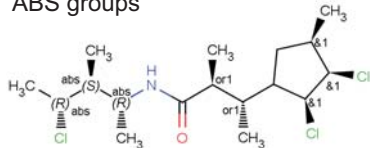
Depiction	Meaning
	Cis
	Trans
	Cis or trans (unknown)
	Not trans
	Not cis

Tetrahedral Chirality

- Stereo bond types:

- Relative stereo configuration

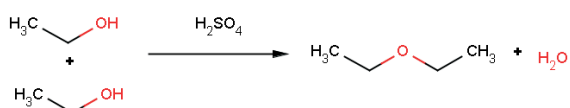
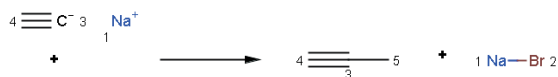
- Chiral flag model
- Enhanced stereo representation: AND<n>, OR<n>, ABS groups



Up	Down	Up or down

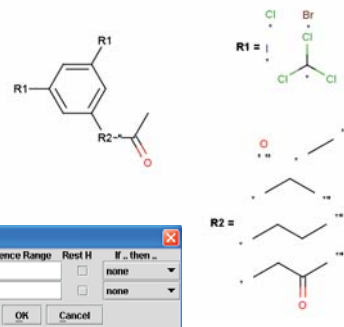
Reaction search

- Reactants, agents, products
- Transformation recognition (mapping)
- Stereospecific reactions (inversion, retention)
- Reactant grouping



R-group search

- Scaffold, R-group definitions
- Monovalent, divalent R-groups
- R-logic
 - Occurrence
 - If-then
 - RestH



Explicit and implicit hydrogens

- H representations:

- Explicit
- Implicit
- Query H count (total or implicit)

	Considered in structural search		
	Explicit H	Implicit H	Query H count
Query	✓	✗	✓
Target	✓	✓	✗

Search example:

Target	Query	H ₃ C-CH ₂ -OH	H ₃ C-CH ₂ -O-H	H ₂ O	H ₃ C-C(=O)-CH ₃
O-H	✓	✓	✓	✓	✗
O	✓	✓	✓	✓	✓
O (H1)	✓	✓	✗	✗	✗

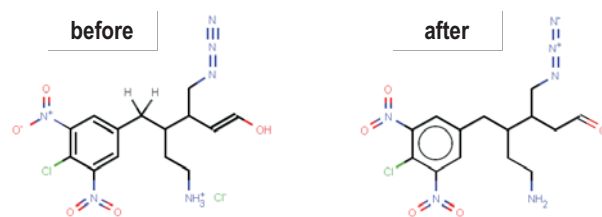
Canonicalization

JChem Standardizer can automatically perform structure canonization during import defined by a rule-based configuration file.

Supported features include:

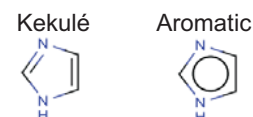
- Explicit hydrogen removal
- Aromatization
- Mesomers (e.g. representations of nitro, azide)
- Tautomers (e.g. oxo/enol forms)
- Counterion removal
- Stereo representation

Canonicalization example:

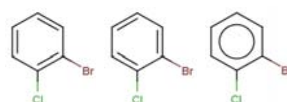


Aromatization

There are two commonly used representations for aromaticity:



Problem: The two Kekulé representations below don't match, so all molecules are transformed to an aromatic representation (right). We provide two types of aromatization:



- ChemAxon-type
- Daylight-type

The Chemical Terms Language

The Chemical Terms language was developed to add a new dimension to cheminformatics tasks like searching, virtual reaction processing, structure filtering, ordering, etc.

Elements of the language:

- Structure matching functions (describing functional groups, reaction sites, similarity, match count...)
- Property calculations (partial charge distribution, pK_a , $\log P$, $\log D$, major microspecies, electrophilicity, TPSA, polarizability, number of rotatable bonds/HB acceptors/donors/rings, exact mass, etc)
- Arithmetic and logic operators
- Extensible: your own Java plugins can be easily added.

Chemical Terms examples

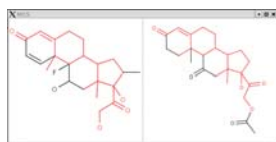
searching	<pre>match("olefine.mol") && !match("c1ccncc1") && (atomCount(16) == 0) (mass() < 300);</pre>
goal functions	<pre>inhibitor = inhibitor.mol; (similarity(inhibitor, pharmacophore_tanimoto) > 0.8) && (similarity(inhibitor, chemical_tanimoto) < 0.5);</pre>
filtering	<pre>(mass() <= 500) && (logP() <= 5) && (donorCount() <= 5) && (acceptorCount() <= 10);</pre>

Similarity search

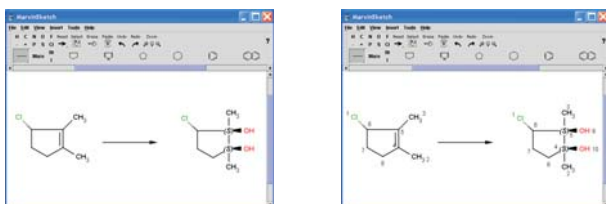
- Descriptors:
 - Chemical hashed fingerprint
 - 2D (topological) pharmacophore fingerprint
 - BCUT
 - Structural keys
 - Hypothesis fingerprints: minimum, average
- Dissimilarity Metrics:
 - Tanimoto: standard, scaled, asymmetric
 - Euclidean: standard, normalized, weighted, asymmetric
 - Optimized for a set of actives

Maximum Common (edge) Subgraph

JChem can compute the maximum connected common (edge) subgraph (MCS).



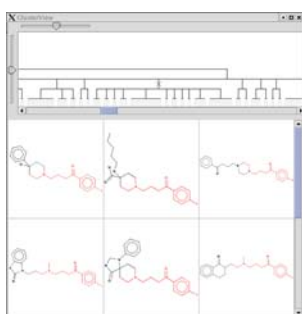
Application: reaction automapping in Marvin



The Library MCS (LibMCS) program rapidly creates a hierarchy of MCS-es on a library.

Applications:

- Identification of the most frequently occurring MCS.
- Focused set analysis
- Clustering based on common substructures

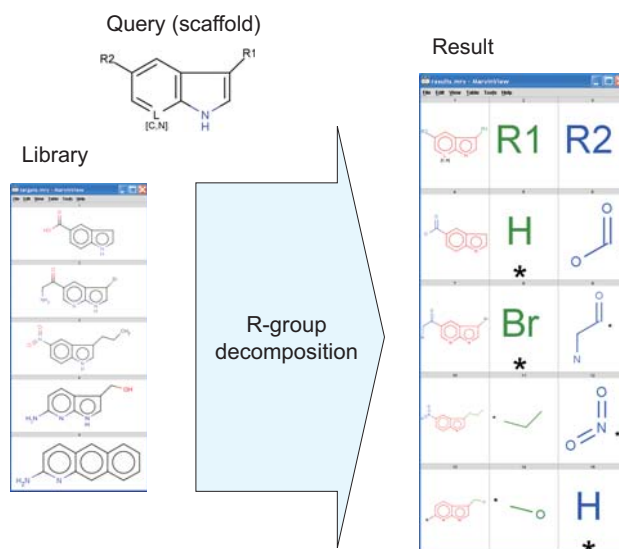


Hierarchical MCS clustering performance

Library	Library size	Time(s)	Clusters	Top level clusters	No. of levels
NCI (small molecules, random, diverse sets)	500	0.7	301	51	5
	1000	2.4	521	93	5
	5000	102.2	2,240	301	5
d2 inhibitors (medium sized molecules, low diversity)	500	0.4	81	7	5
	1000	0.6	83	8	5
Thrombin inhibitors (medium sized molecules, moderate diversity)	1000	2.1	127	17	5
	3000	3.9	145	18	5

R-group decomposition

JChem is able to identify the ligands of a given scaffold at specified substitution positions:



Conclusions

Structural search techniques provide useful tools for chemists and cheminformaticians and are often the basis of more complex tasks.

ChemAxon's JChem suite contains a broad range of chemical search facilities with a rich set of features. This adds convenience and flexibility to users formulating queries. The presented benchmark results illustrate the high performance of JChem search.

Chemical Terms as a general chemical expression format enables chemists to define complex conditional expressions in a standard format for applications like structure queries and advanced pharmacophore point definitions.

Implementation

All software applications mentioned above are parts of ChemAxon's JChem 3.1.2 software suite (100% Java).

The Chemical Term evaluation engine used in presented applications supports the public plugin API of ChemAxon and uses their prediction plugins.

Hardware and software requirements: any system running Java Runtime Environment 1.4 or above.

To download or evaluate online implementations please visit <http://www.chemaxon.com/jchem>