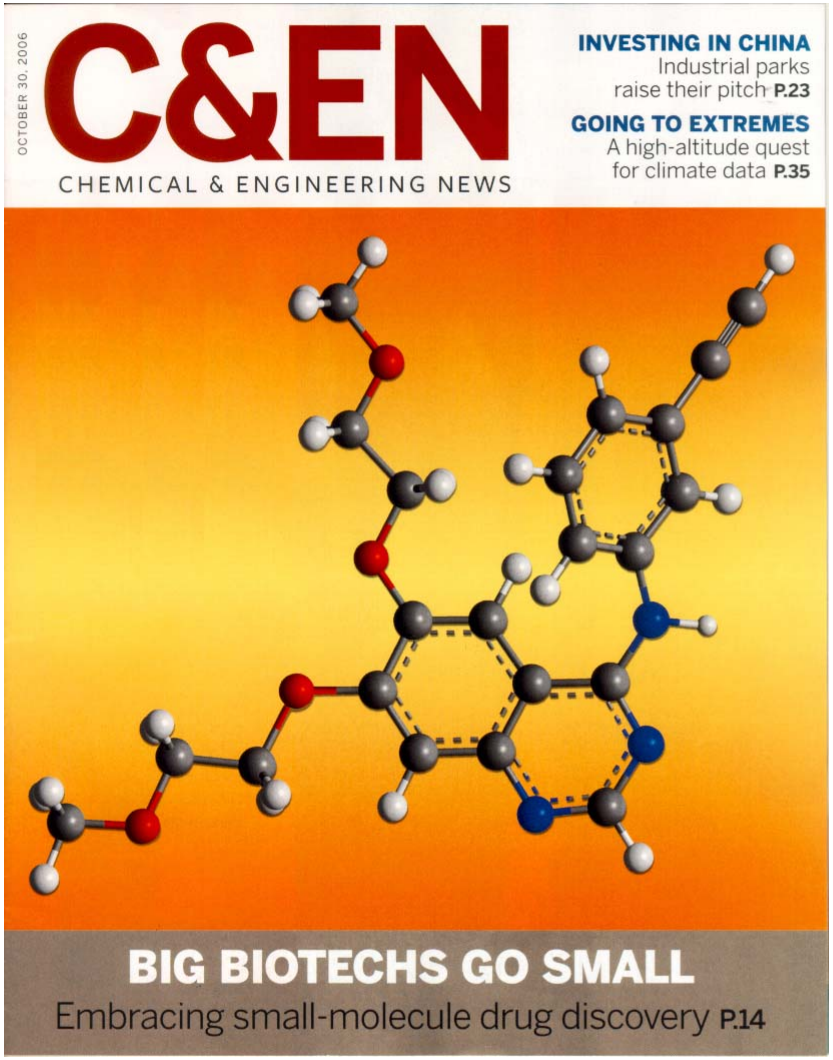
How to create a web-based molecular structure database with free software

Norbert Haider

Department of Drug Synthesis
Faculty of Life Sciences, University of Vienna

norbert.haider@univie.ac.at

small molecules are still going strong....



OCTOBER 30, 2006

C&EN

CHEMICAL & ENGINEERING NEWS

INVESTING IN CHINA
Industrial parks
raise their pitch **P.23**

GOING TO EXTREMES
A high-altitude quest
for climate data **P.35**

BIG BIOTECHS GO SMALL

Embracing small-molecule drug discovery **P.14**

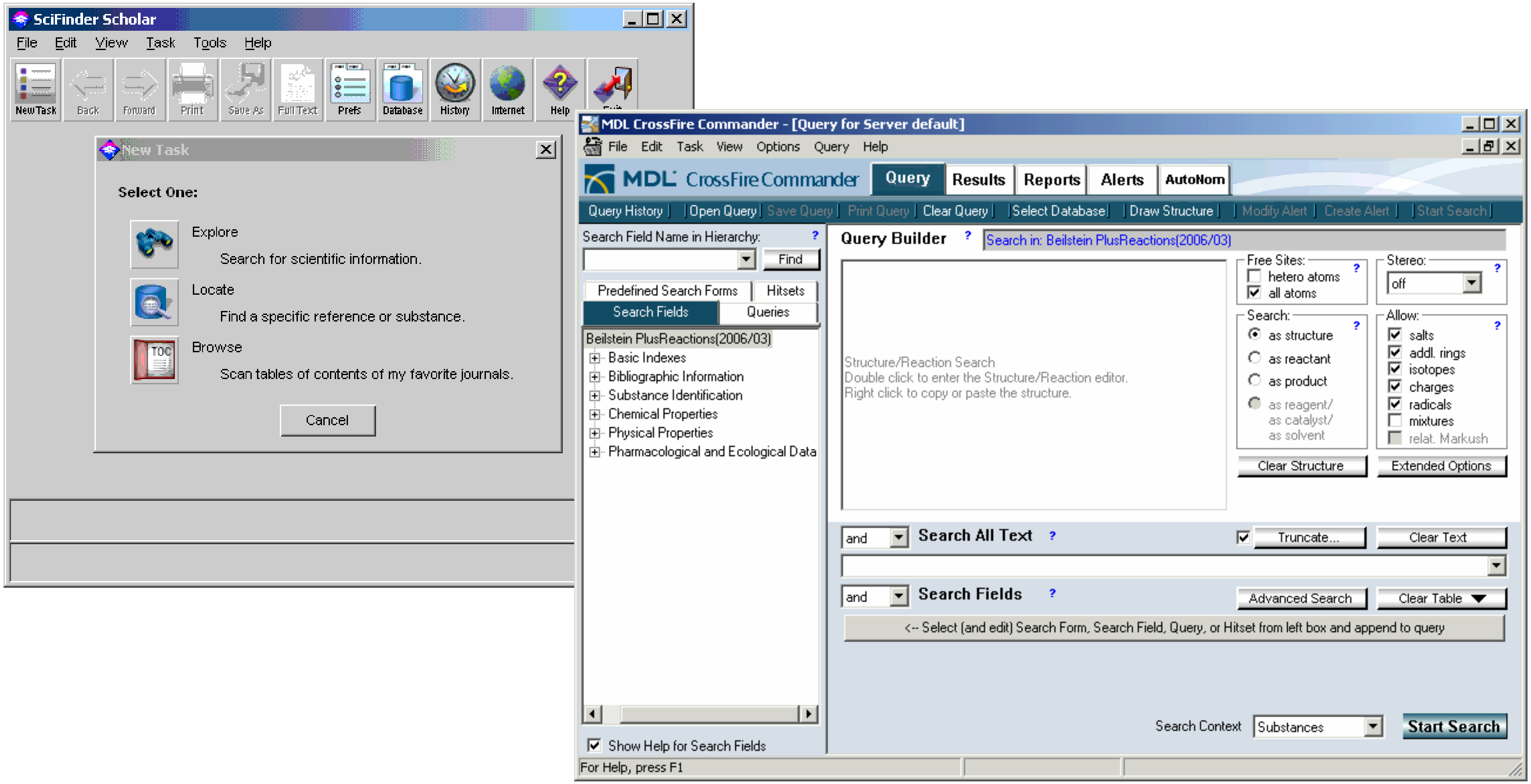
- revival of small molecules in drug discovery at "classical" pharmaceutical companies

- growing interest in small molecules also at major Biotech players

➔ database technologies for structure handling are essential IT tools

molecular structure databases for “small molecules”

DB using proprietary server software + proprietary client software, e.g. CAS SciFinder, MDL Crossfire


SciFinder Scholar
File Edit View Task Tools Help
New Task
Back
Forward
Print
Save As
Full Text
Prefs
Database
History
Internet
Help
Exit
New Task

Select One:

Explore

Search for scientific information.

Locate

Find a specific reference or substance.

Browse

Scan tables of contents of my favorite journals.

TOC
Cancel
MDL CrossFire Commander - [Query for Server default]
File Edit Task View Options Query Help
MDL CrossFire Commander
Query
Results
Reports
Alerts
AutoNom
Query History | Open Query | Save Query | Print Query | Clear Query | Select Database | Draw Structure | Modify Alert | Create Alert | Start Search
Search Field Name in Hierarchy:
Find
Predefined Search Forms
Hitsets
Search Fields
Queries
Beilstein PlusReactions(2006/03)
Basic Indexes
Bibliographic Information
Substance Identification
Chemical Properties
Physical Properties
Pharmacological and Ecological Data
Query Builder

Search in: Beilstein PlusReactions(2006/03)

Structure/Reaction Search
Double click to enter the Structure/Reaction editor.
Right click to copy or paste the structure.

Free Sites: □ hetero atoms ☑ all atoms

Stereo: off

Search: ● as structure ○ as reactant ○ as product ○ as reagent/ as catalyst/ as solvent

Allow: ☑ salts ☑ addl. rings ☑ isotopes ☑ charges ☑ radicals □ mixtures □ relat. Markush

and | Search All Text | ☑ Truncate... | Clear Text

and | Search Fields | Advanced Search | Clear Table

<-- Select (and edit) Search Form, Search Field, Query, or Hitset from left box and append to query

☑ Show Help for Search Fields

Search Context: Substances

For Help, press F1

molecular structure databases for “small molecules”

- DB with access via WWW:
  using common web browser as client

  ➔ what about the server?

  ➔ what about the client’s capabilities to generate query structures?

  ➔ how to display the results??

server software: possible solutions

- €€€ ($$$) commercial products:
  - Oracle (SQL database) + add-on (“cartridge”), e.g. from MDL, CambridgeSoft
  - other commercial products like JChem by ChemAxon

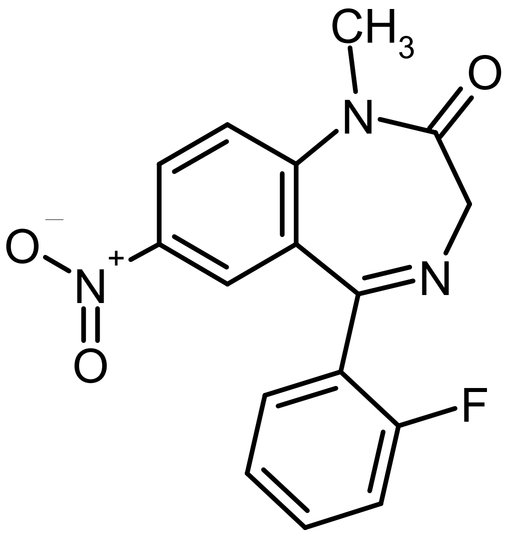
- free software ➔ availability? usability? frameworks, toolkits:
  - CDK: Chemistry Development Kit
  - OpenBabel (obgrep)
  - writing our own software
➔ **checkmol/matchmol**

checkmol: the very beginning

- PharmXplorer project (NML): eLearning portal developed by universities of Graz, Innsbruck, and Vienna ➔ http://www.pharmxplorer.at/
- funding for (wo)manpower, not IT infrastructure
- open-source solutions preferred
- main component: "information platform" including a database of all drug compounds on the Austrian market
- initially no structure/substructure search
- alternative: search by **functional groups**

checkmol: the very beginning

- need to assign functional groups to approx. 2500 chemical structures



→

- imine
- aryl fluoride
- tertiary carboxamide
- lactam
- nitro compound
- aromatic compound
- heterocycle

"manual" assignment
time? quality?

↔

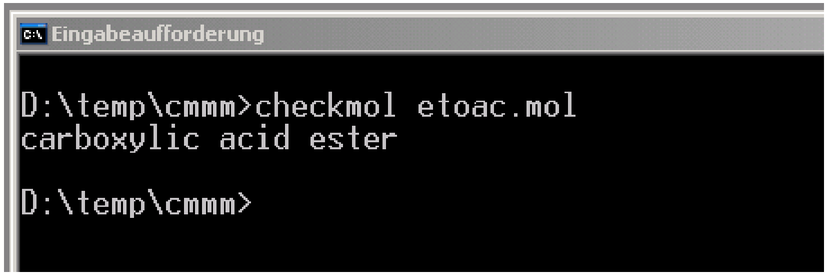
automatic assignment
checkmol

what checkmol does:

- read input structure (MDL molfile format):

```
CCOC(C)=O
JME 2003.05
Ethyl acetate
  6  5  0  0  0  0  0  0  0  0999 V2000
    4.8486    0.0742    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2124    2.1000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.6153    0.7368    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4248    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.2124    0.7000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  4  1  0  0  0  0
  2  6  1  0  0  0  0
  3  6  2  0  0  0  0
  4  5  1  0  0  0  0
  5  6  1  0  0  0  0
M  END
```

- analyze input structure
- write output:



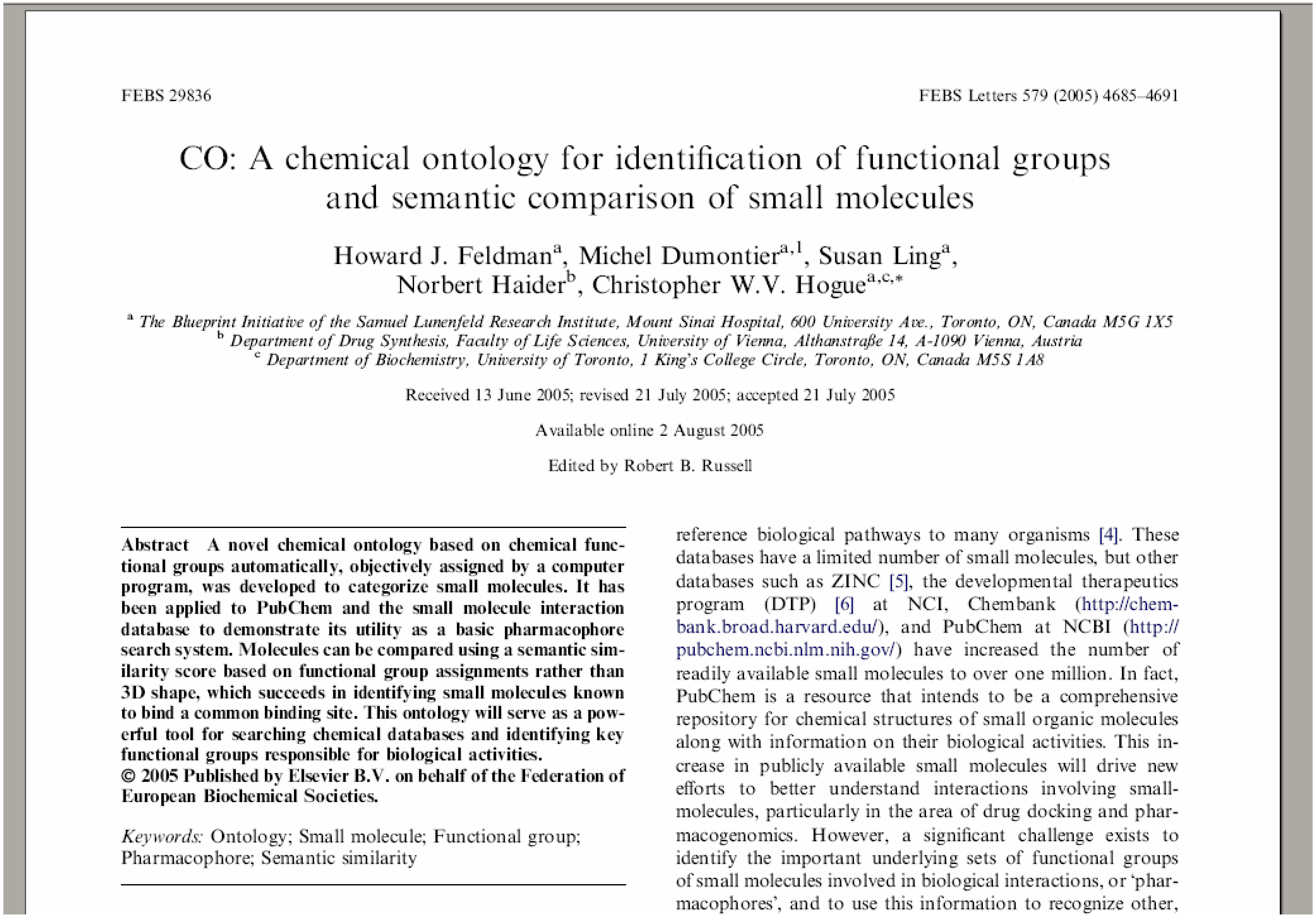
```
D:\temp\cmmm>checkmol etoac.mol
carboxylic acid ester

D:\temp\cmmm>
```

checkmol features

- supported input formats:
  - MDL mol
  - Alchemy mol
  - SYBYL mol2
- supported output formats:
  - text (English or German)
  - 8-digit code, e.g C3NOC000
  - binary code (bitstring)
- currently approx. 200 functional groups
  http://merian.pch.univie.ac.at/~nhaider/cheminf/fgtable.pdf

checkmol applications: a chemical ontology



FEBS 29836 FEBS Letters 579 (2005) 4685–4691

CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules

Howard J. Feldman[a], Michel Dumontier[a,1], Susan Ling[a], Norbert Haider[b], Christopher W.V. Hogue[a,c,*]

[a] *The Blueprint Initiative of the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave., Toronto, ON, Canada M5G 1X5*
[b] *Department of Drug Synthesis, Faculty of Life Sciences, University of Vienna, Althanstraße 14, A-1090 Vienna, Austria*
[c] *Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada M5S 1A8*

Received 13 June 2005; revised 21 July 2005; accepted 21 July 2005

Available online 2 August 2005

Edited by Robert B. Russell

Abstract A novel chemical ontology based on chemical functional groups automatically, objectively assigned by a computer program, was developed to categorize small molecules. It has been applied to PubChem and the small molecule interaction database to demonstrate its utility as a basic pharmacophore search system. Molecules can be compared using a semantic similarity score based on functional group assignments rather than 3D shape, which succeeds in identifying small molecules known to bind a common binding site. This ontology will serve as a powerful tool for searching chemical databases and identifying key functional groups responsible for biological activities.

© 2005 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

Keywords: Ontology; Small molecule; Functional group; Pharmacophore; Semantic similarity

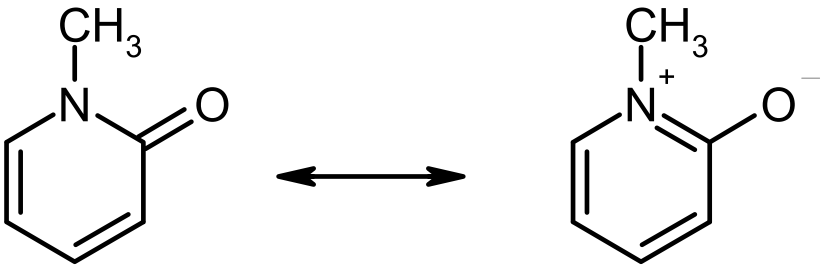
reference biological pathways to many organisms [4]. These databases have a limited number of small molecules, but other databases such as ZINC [5], the developmental therapeutics program (DTP) [6] at NCI, Chembank (http://chembank.broad.harvard.edu/), and PubChem at NCBI (http://pubchem.ncbi.nlm.nih.gov/) have increased the number of readily available small molecules to over one million. In fact, PubChem is a resource that intends to be a comprehensive repository for chemical structures of small organic molecules along with information on their biological activities. This increase in publicly available small molecules will drive new efforts to better understand interactions involving small-molecules, particularly in the area of drug docking and pharmacogenomics. However, a significant challenge exists to identify the important underlying sets of functional groups of small molecules involved in biological interactions, or 'pharmacophores', and to use this information to recognize other,

checkmol internals

- ring search algorithm:
  SAR = set of all rings (max. 1024 rings)

- fallback to SSR = set of small rings
  (size < 13 atoms, no “envelope rings”)

- aromaticity detection based on Hückel rule
  (4n + 2 π electrons) + mesomeric structures, e.g.



building a web database: how to start

- basic software package: LAMP
  - Linux: operating system
  - Apache: web server
  - MySQL: relational database management system
  - PHP: scripting language
- user input: e.g., „show me all compounds with an ester function“ (selection from listbox etc.) ➔ transformed into SQL query, e.g. SELECT mol_id FROM mol_fg WHERE fgcode LIKE ‘C3O2C000’;
- display list of hits, dynamic generation of HTML output with PHP

functional group search is fine, but....

the next step:
structure/substructure search

- extension of **checkmol**‵s capabilities
- comparison of two structures:
  matchmol

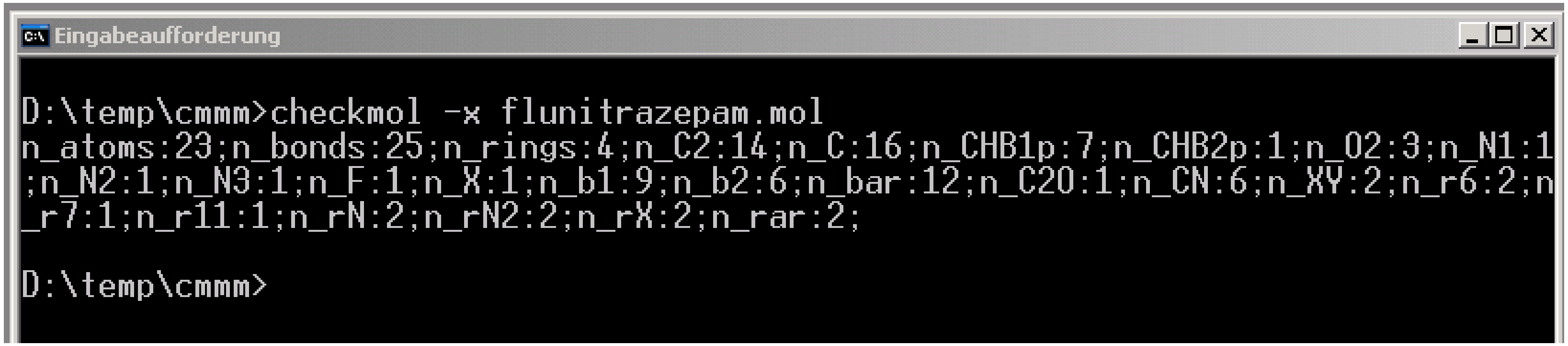
structure/substructure search: workflow

a two-stage process saves CPU time:

- **preselection**: removes as many candidate structures as possible, based on structural features
- **atom-by-atom** matching of the remaining candidate structures with the query structure

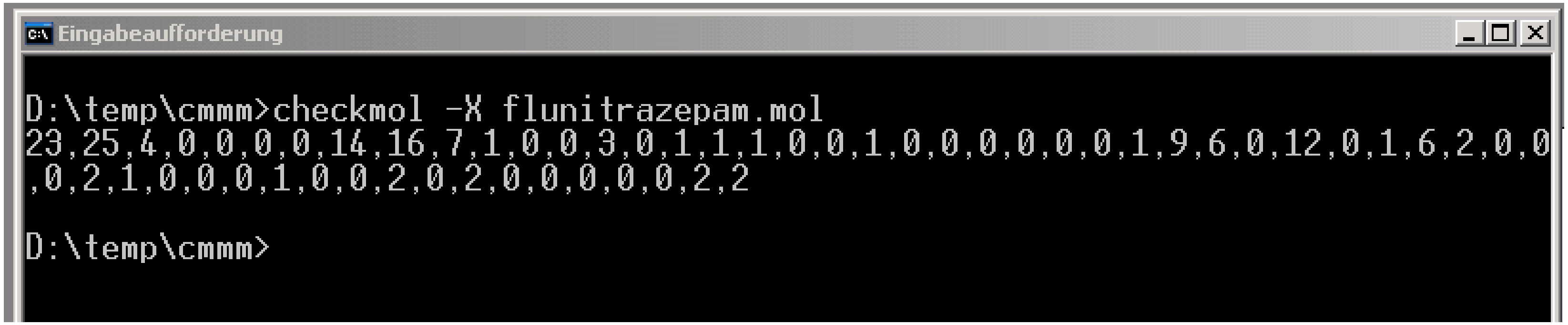
checkmol features: molecular statistics

structural descriptors for rapid preselection, can be stored in a MySQL table



```
D:\temp\cmmm>checkmol -x flunitrazepam.mol
n_atoms:23;n_bonds:25;n_rings:4;n_C2:14;n_C:16;n_CHB1p:7;n_CHB2p:1;n_O2:3;n_N1:1
;n_N2:1;n_N3:1;n_F:1;n_X:1;n_b1:9;n_b2:6;n_bar:12;n_C2O:1;n_CN:6;n_XY:2;n_r6:2;n
_r7:1;n_r11:1;n_rN:2;n_rN2:2;n_rX:2;n_rar:2;

D:\temp\cmmm>
```



```
D:\temp\cmmm>checkmol -X flunitrazepam.mol
23,25,4,0,0,0,0,14,16,7,1,0,0,3,0,1,1,1,0,0,1,0,0,0,0,0,0,1,9,6,0,12,0,1,6,2,0,0
,0,2,1,0,0,0,1,0,0,2,0,2,0,0,0,0,0,2,2

D:\temp\cmmm>
```

list of all descriptors: **`checkmol -l`**

checkmol features: molecular statistics

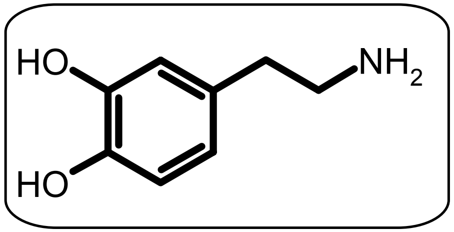
whenever a new molecule is stored in the database
➔ molstat "fingerprints" are calculated and stored in "molstat" table

whenever a query structure is submitted for exact search, its molstat values are translated into an SQL query, e.g.:

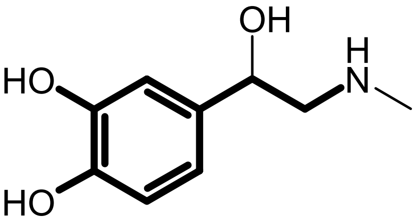
SELECT mol_id FROM molstat WHERE (n_atoms = 23) AND (n_bonds = 25) AND (n_rings = 4) AND ...

substructure search: n_atoms >= 23   etc.

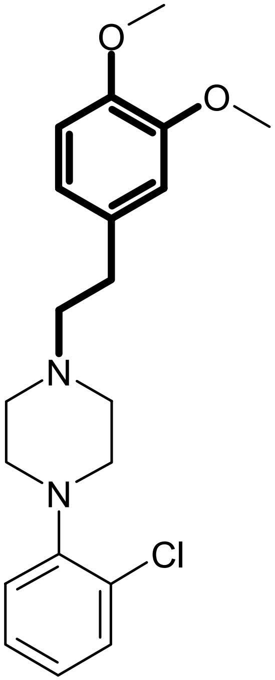
subgraph isomorphism: atom-by-atom matching



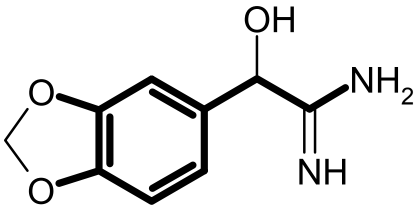
query structure



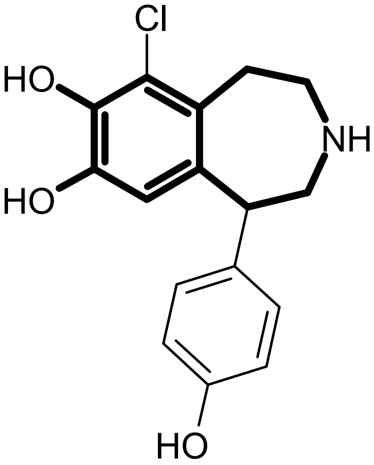
adrenaline



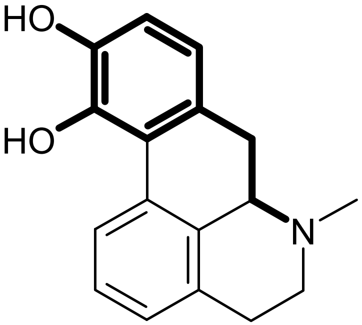
mefeclorazine



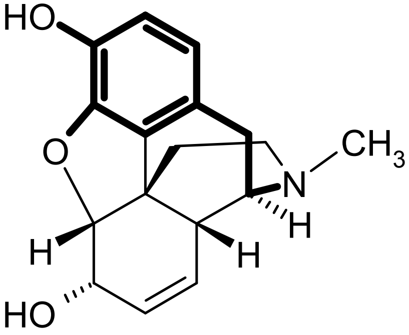
olmidine



fenoldopam



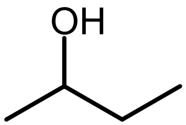
apomorphine



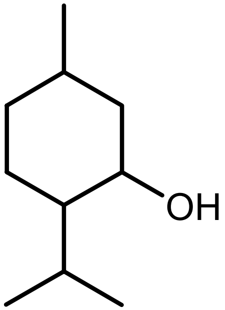
morphine

strategies for atom-by-atom matching

example: is 2-butanol a substructure of menthol?



query structure (Q)



candidate structure (C)

a) "brute force" approach:
take any atom of Q and match it against every atom of C (compare all bonds, all neighbors, all neighbors of neighbors, etc.)

b) pick the most "unique" atom of Q and match it against its possible counterparts in C:

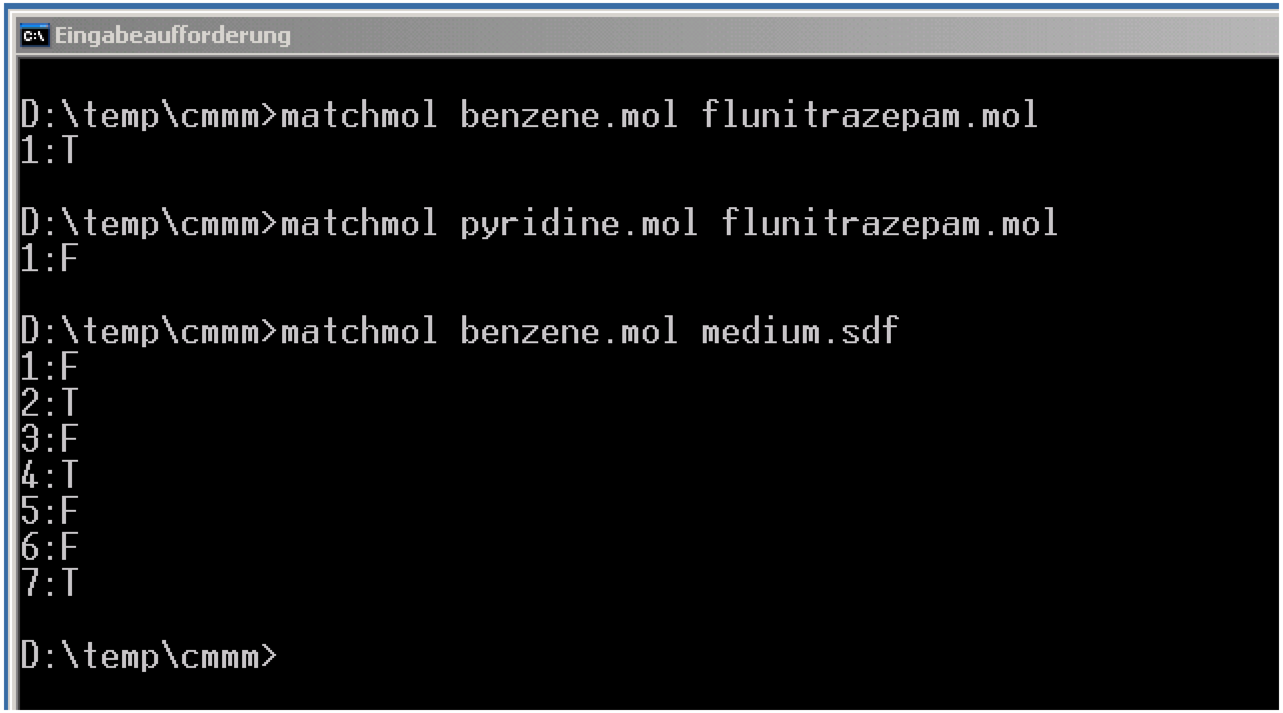
- highest degree of branching *or*
- highest number of heteroatom substituents *or*
- heteroatoms themselves *or*
- highest rank by Morgan's algorithm etc.

matchmol: checlmol's companion

development of **matchmol**

- a simple command-line utility
- input: two (or more) structures
- output: "yes" or "no" ("T" or "F")
- exact or substructure matching
- input can be taken from
  - 2 files (MDL mol or SDF)
  - standard input (SDF format, first structure is always the query structure)
- output is written to standard output

matchmol: usage

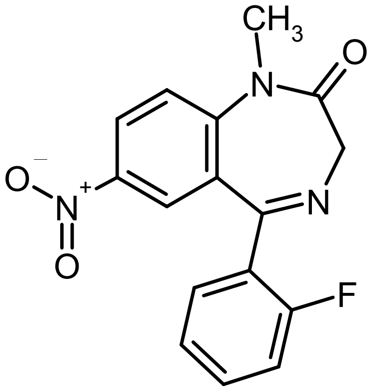


```
D:\temp\cmmm>matchmol benzene.mol flunitrazepam.mol
1:T

D:\temp\cmmm>matchmol pyridine.mol flunitrazepam.mol
1:F

D:\temp\cmmm>matchmol benzene.mol medium.sdf
1:F
2:T
3:F
4:T
5:F
6:F
7:T

D:\temp\cmmm>
```



selected command-line options:

-x exact match

-s strict comparison of atom types and bond types

-m output in MDL mol/SDF format, for example

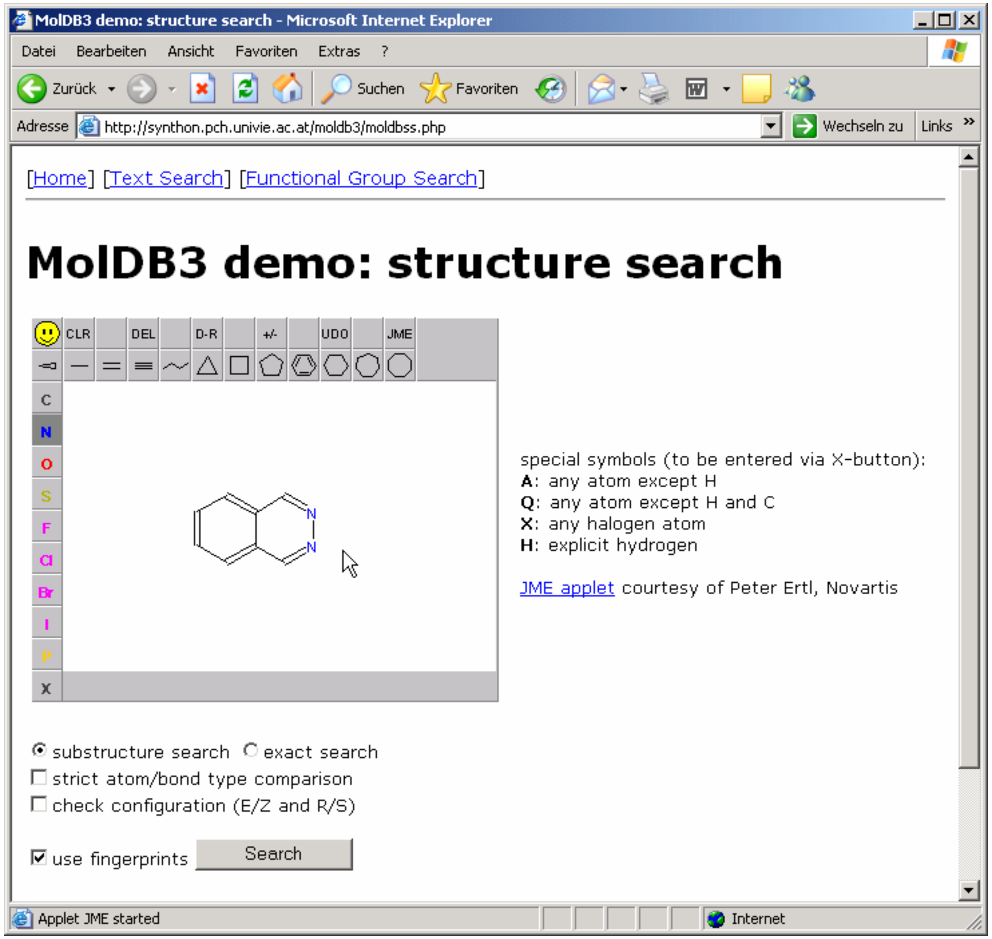
```
matchmol -m uracil.mol maybridge.sdf > maybridge-uracils.sdf
```

building a web database: integrating the parts

user interface:

- PHP scripts for dynamic HTML generation and MySQL database connectivity

- Java applet for structure input (molfile format) ➔ JME (P. Ertl)


MolDB3 demo: structure search - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück
Suchen
Favoriten
Adresse http://synthon.pch.univie.ac.at/moldb3/moldbss.php
Wechseln zu
Links
CLR
DEL
D-R
+/-
UDO
JME
C
N
O
S
F
Cl
Br
I
P
X
N
N
Search
Applet JME started
Internet

[Home] [Text Search] [Functional Group Search]

MolDB3 demo: structure search

special symbols (to be entered via X-button):
A: any atom except H
Q: any atom except H and C
X: any halogen atom
H: explicit hydrogen

JME applet courtesy of Peter Ertl, Novartis

● substructure search ○ exact search
□ strict atom/bond type comparison
□ check configuration (E/Z and R/S)

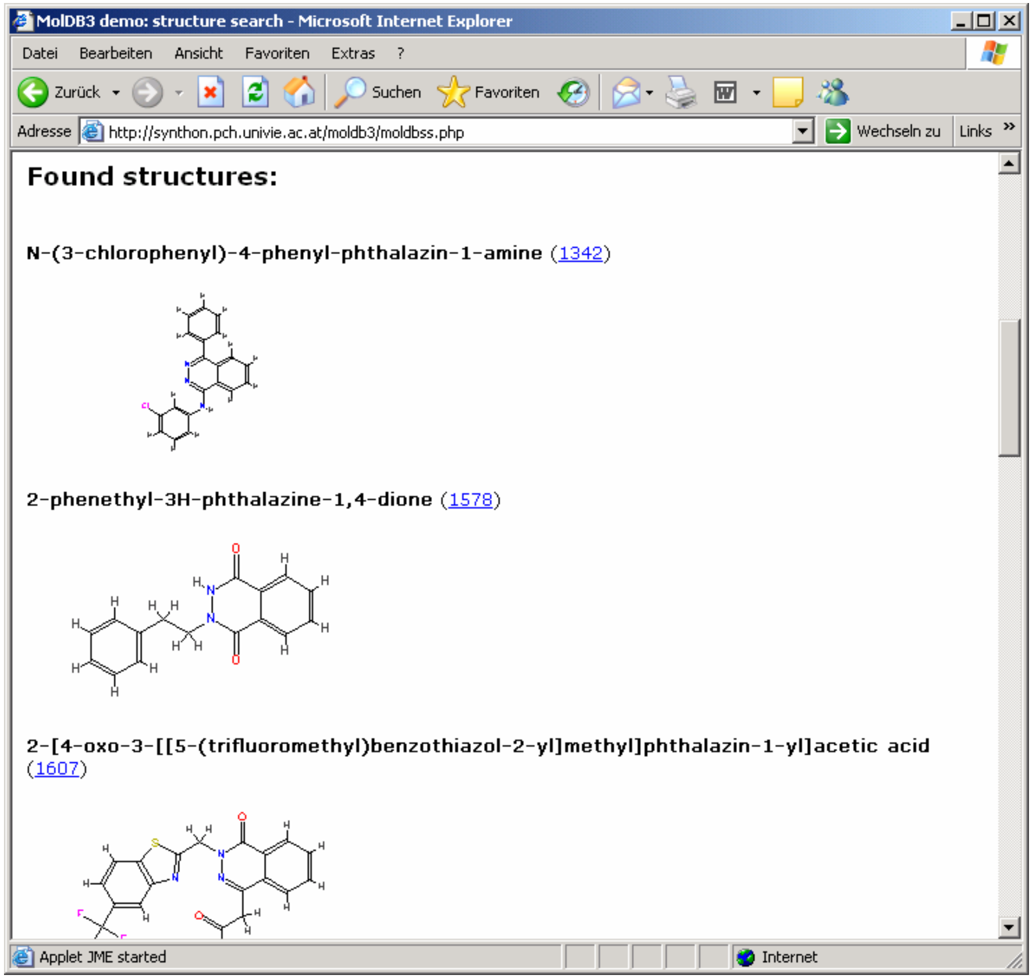
☑ use fingerprints

structure/substructure search: basic workflow

- user draws query structure in JME
- query structure (MDL molfile) is passed to **checkmol** ➔ molstat fingerprints
- preselection: SQL query, using molstat ➔ list of candidate molecules
- **matchmol** is invoked for each candidate structure in combination with query structure (atom-by-atom matching)
- if matchmol returns “T” ➔ hit!
- display hits in appropriate format

displaying the hits

- option A: using JME in “depict” mode


MolDB3 demo: structure search - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück
Suchen
Favoriten
Adresse
http://synthon.pch.univie.ac.at/moldb3/moldbss.php
Wechseln zu
Links
Internet

Found structures:

N-(3-chlorophenyl)-4-phenyl-phthalazin-1-amine (1342)

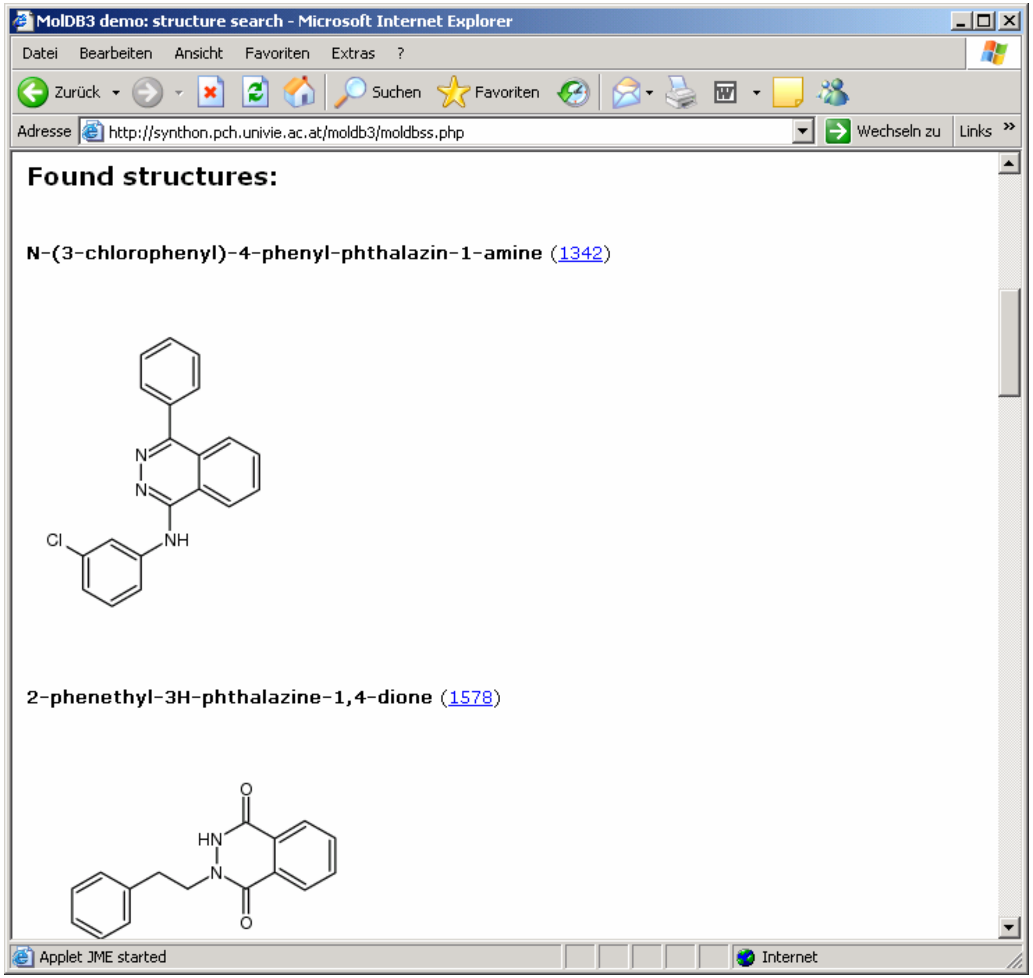
2-phenethyl-3H-phthalazine-1,4-dione (1578)

2-[4-oxo-3-[[5-(trifluoromethyl)benzothiazol-2-yl]methyl]phthalazin-1-yl]acetic acid (1607)

Applet JME started

displaying the hits

- option B: using static bitmap images


MolDB3 demo: structure search - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück
Suchen
Favoriten
Adresse http://synthon.pch.univie.ac.at/moldb3/moldbss.php
Wechseln zu
Links
N
N
NH
Cl
O
HN
N
O
Applet JME started
Internet

Found structures:

N-(3-chlorophenyl)-4-phenyl-phthalazin-1-amine (1342)

2-phenethyl-3H-phthalazine-1,4-dione (1578)

software used:

mol2ps: generates PostScript graphics

GhostScript: renders PS files into bitmap graphics, e.g. PNG, GIF

performance considerations (1)

- save time by reading aromaticity information from “tweaked” MDL molfiles

```
Salicyclic acid
  CheckMol                        TMF02:r0:m0

 10 10  0  1                  999 V2000
   -1.2139   -0.1916    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
   -1.2151   -0.9023    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
   -0.6003   -1.2569    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
    0.0120   -0.9019    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
    0.0091   -0.1880    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
   -0.6021    0.1628    0.0000 C   0 00  0  0  0  0  0  0  0  0  0  0
   -0.6035    0.8711    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0092    1.2265    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2176    1.2241    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.6214    0.1682    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  3 01  0  1  0  0
  5  6 02  0  1  0  0
  6  1 01  0  1  0  0
  1  2 02  0  1  0  0
  6  7  1  0  0  0  0
  3  4 02  0  1  0  0
  7  8  1  0  0  0  0
  7  9  2  0  0  0  0
  4  5 01  0  1  0  0
  5 10  1  0  0  0  0
M  END
```

leading “0” in charge column: atom belongs to aromatic ring

leading “0” in bond type column: bond belongs to aromatic ring

performance considerations (2)

- save time by reducing the number of matchmol calls (PHP shell calls via **`popen()`** function)
  - initial version: called matchmol with query structure + each single candidate structure
    number of shell calls = number of candidates
  - advanced version: "burst mode" ➔ uses assemblies of query structure + approx. 10 candidates (SDF format)
    ➔ number of shell calls reduced by ~90%
- use a faster shell: **/bin/ash** instead of /bin/bash
- store structures in DB records instead of files

performance considerations (3)

combination of

- molstat fingerprints (higher selectivity with larger query structures)

  and

- binary fingerprints (higher selectivity with smaller query structures): fragment dictionary, e.g. all common ring systems

binary fingerprints: basic principle

- create an SDF file with up to 62 entries, representing the fragment dictionary

- run matchmol in “fingerprint” mode, e.g.

  ```
  matchmol -F theophylline.mol fp01.sdf
  4644345705660416
  ```

  ➔ 00000000000**1**0000**1**0000000000000**1**000000000000000000000000000000000

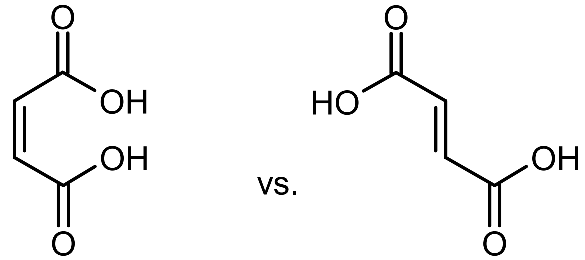
- output: decimal number representing a bitstring of 64 bits, each bit signals the absence (0) or presence (1) of a particular fragment in the input structure

- can be searched very efficiently in any SQL database by “bitwise AND” operation

extended features

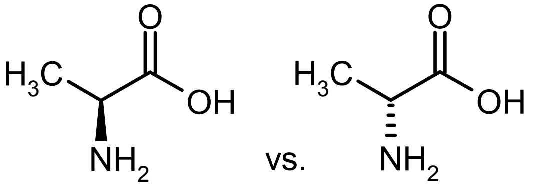
- E/Z geometry check

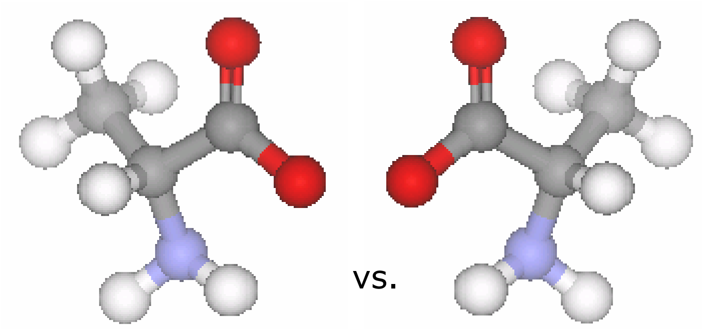
```
matchmol -g needle.mol haystack.mol
```



- R/S geometry (chirality) check

```
matchmol -G needle.mol haystack.mol
```





from tools to practical solutions

- release of **checkmol/matchmol** under the terms of the GPL (GNU General Public License): http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html
- release of **mol2ps** under the GPL:

  http://merian.pch.univie.ac.at/~nhaider/cheminf/mol2ps.html
- JME applet is available upon request from peter.ertl@pharma.novartis.com
- release of a fully functional package of PHP scripts, setup scripts (Perl) and installation instructions: **MolDB3**

  http://merian.pch.univie.ac.at/pch/download/chemistry/moldb/

  moldb3.tar.gz

MolDB3 package: getting started

- have your LAMP system up and running
- download & install all required software
- edit a simple configuration file
- import your structures + data from any SDF file:
  - automatic analysis + manual adjustments
  - automatic import, including
    - tweaking of molfiles
    - generation of functional group descriptors
    - generation of molstat and binary fingerprints
    - generation of 2D bitmap pictures (if desired)
- have fun!

example installations

- MolDB3 demo page (~10.000 structures): http://synthon.pch.univie.ac.at/moldb3/
- PubChem demo page (~100.000 structures): http://synthon.pch.univie.ac.at/pubchem/
- CSEARCH web frontend (~140.000 structures): http://nmrpredict.orc.univie.ac.at/csearchlite/
- MolBank (online journal), Austrian mirror site: http://at.mdpi.net/molbank/molbanksss.php

- related sites, using checkmol/matchmol:
  - SMID (Small Molecule Interaction Database) http://smid.blueprint.org/index2.php
  - Aurora Fine Chemicals online catalog http://www.aurorafinechemicals.com/chemicals-catalog.html
  - the pgchem::tigress project (PostgreSQL add-on) http://pgfoundry.org/projects/pgchem/

summary

- MolDB3 is a fully functional package for a web-based, searchable molecular structure database
- moderate requirements: standard PC, LAMP
- convenient data import from SDF files
- “chemical intelligence” is located in a compact command-line program: checkmol/matchmol
- reasonable performance for up to ~100.000 structures
- open source
- easily extendable

acknowledgements

- bm:bwk
  funding of PharmXplorer
- Rami Jbara, University of Vienna
  PharmXplorer PHP programming, 8-digit codes
- Alessandro Barozza, Procos S.p.A. (Italy)
  first Windows DLL version, bug reports
- Howard Feldman, The Blueprint Initiative (Canada)
  feature requests, bug reports, SMID integration
- Ernst-Georg Schmid, Bayer Business Services (D)
  feature requests, bug reports, C port, DLL, pgchem::tigress project