ChemIDplus: A Free, Web-Based Portal to a Variety of Compound-Based Information

Mitch Miller, American Chemical Society, 10 September 2003
George F. Hazard Jr., Vera W. Hudson, Christopher Hilt, Jenny Fang, David Mayer, and Larry Callahan

LION

Definition of ‘Integration’

Uniting data from multiple sources about a set of items of interest in a way that turns the data into information and allows a savvy person to turn the information into knowledge

Outline

- The application in context
- The current system
- Futures
- Wrap-up

The Application in Context

The Problem

- Lots of information about lots of compounds
  - How can it all be searched coherently?
    - Multiplicity of names, identifiers
    - Structures
  - Once you find a compound of interest, how can you get to all information?
- Updating, extending information
  - New compounds
  - New information about existing compounds
- What types of compounds qualify for ChemIDplus?
  - Drugs
  - Environmental contaminants
  - Industrial compounds

The Solution

- ChemIDplus: Web-based search, browse system for chemical databases
- Free to all
- Database architecture
  - Data in Oracle
    - Text data in Oracle (>360,000 records)
    - Structures in MDL ISIS/Host (RCG) database (> 2D 173,000 structures and growing)
    - 3D structures calculated using CORINA from Molecular Networks
      - Up to 5 conformations per molecule
      - Updated once per month
      - Application only provides access to first conformation

Current Application

Environment

- Interface: Web
  - Browser-neutral, client-neutral
- Server: iPlanet Web Server 6
- MDL's Chemscape
- Data access:
  - Searching via Hview
  - Retrieval through Hview and LiveWire
  - Chemical structures rendered
    - Chime
    - ChemSymphony-based Java applets

Searching in ChemIDplus

- Structures (3%)
  - Substructure (2%)
  - Similarity (<1%)
    - Facility for one-button ‘follow-up’ similarity search
  - Exact (<1%)
- Text:
  - Names (51%)
  - Registry numbers (45%)
  - Locator codes (<1%)
  - Classification codes (<1%)
- Statistics for the past 3 months

Database Operations

- Searching using the publicly available Web application
- Text data addition, revision using a separate Web-based Database Maintenance Application
- Structures registered using ISIS/Base
- 3D structures registered using HostCLI

Name Searches

- Very easy to misspell chemikal names!
- Trap cases where exact name searches retrieve no hits
- Perform a ‘fuzzy’ name search
  - Implemented algorithm developed by Antonio Zamora
    - Computes hash code for input name
    - Compares hash to stored index for registered names
    - Produces ranked list of names
- Gives user a list of near matches
- Improved user satisfaction

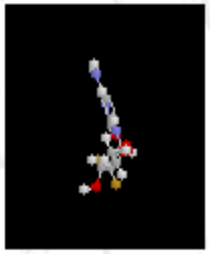
ChemIDplus Usage

- Search for data and/or structure
- Browse hit list
- Drill down to interesting molecules
  - Browse data from NLM database
  - Use locator codes to view data on other systems

Current Search Page

Division of Specialized Information Services, NLM
ChemID*plus* Chemical Search Input Page
A TOXNET Resource

Welcome to a new and improved version of ChemID*plus*!
New Last Updated: February 2, 2003



ChemID*plus*
367368 Records
163083 Structures

Powered by Chemscape TM
http://www.mdli.com/chemscape

Structure Input Box

Structure Search Type: Substructure Search

Display 10 Results

Display chemical structures using
Chime
Change Now
Java

Data Search Type:
Name/Synonym | Equals

valeum

And optional **Locator Code** limit is
Please choose one

Search | Clear | Help

LION

Your search for chemical name "valeum" did not retrieve any matches.

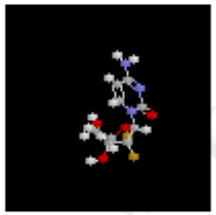
Here are some similar names:

Click on the one you're interested in

(your desired name may not be first on the list)

Valium
Valeo
Valu-7
Valu-8
Valux
Vallene
Value 100
Valinum [Latin]
Vaulen
Vaderm
Velium
Velmol
Vulm 993
VULM 993
Vulm-411
Vulm 411
Vulm 111
Valerin
Valetan

LION

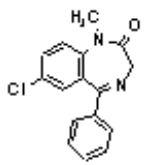


Diazepam [USAN:BAN:INN:JAN] RN: 439-14-5

Molecular Structure | Names and Synonyms | Classification Codes | Formulas | Notes

Registry Numbers | Locator Codes | Entire Record | Main Query Page

ChemIDplus



Enlarge Structure

File Locator

- AIDSLINE
- CANCERLIT
- CCRIS
- ClinicalTrials.gov
- DART/ETIC
- DSL
- EINECS
- EMIC
- GENETOX
- HSDB
- MEDLINE
- MEDLINEplus
- MESH

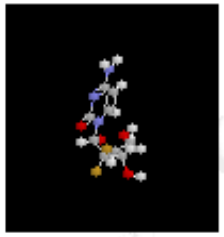
Welcome to the ChemIDplus Locator Page

The window to the left displays one or more ChemIDplus Locator Codes which link to other sites that carry information about the chemical you have retrieved. You may click on:

- A hyperlinked Locator Code such as TOXLINE to retrieve data from that resource. Results will be displayed in this window.
- The information icon (i) to get a description of the source of the Locator Code and the scope of the search.

The window on the top of the screen displays buttons that you may click on to change the data displayed for this chemical or to change your search mode. Hold your mouse pointer over a button to see a more detailed description.

LION



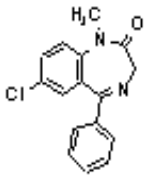
Diazepam [USAN:BAN:INN:JAN] RN: 439-14-5

Molecular Structure | Names and Synonyms | Classification Codes | Formulas

Notes | Registry Numbers | Locator Codes | Entire Record

Main Query Page

ChemIDplus



Enlarge Structure

File Locator

- AIDSLINE
- CANCERLIT
- CCRIS
- ClinicalTrials.gov
- DART/ETIC
- DSL
- EINECS
- EMIC
- GENETOX
- HSDB
- MEDLINE

HSDB | Download | Details | Other Files | TOXNET Home

Contents — Contract all categories | Expand all categories | Select | Clear

- FULL RECORD
- Human Health Effects
- Emergency Medical T
- Animal Toxicity Studi
- Metabolism/Pharmaco
- Pharmacology

DIAZEPAM

CASRN: 439-14-5

For other data, click on the Table of Contents

Human Health Effects:

Evidence for Carcinogenicity:

Evaluation: There is evidence suggesting a lack of carcinogenicity of diazepam to the breast and inadequate evidence for carcinogenicity at other sites in humans. There is inadequate evidence in experimental animals for the carcinogenicity of diazepam. Overall Evaluation: Diazepam is

Co

Future Developments

LION

Near Future Enhancements

- New architecture
- Retain database
  - Oracle, ISIS/Host database
- MDL Relational Chemistry Server
- Use Java servlets/JSP for server-side processing
  - Tomcat
- Chemical structure rendering:
  - GIFs
  - Chime
  - Applets based on Marvin beans

Near Term: ‘Lite’ Version

- Fast-track for searching text data (name, registry number, etc.)
  - No structure searching
- UI conforms to NLM TOXNET standard
  - http://toxnet.nlm.nih.gov/
- In testing now
- Due out soon
  - http://chem2.sis.nlm.nih.gov/chemidplus/
  - Expect some changes, down time before release
  - Please report any problems

‘Heavy’ version

- Based on new architecture
- Includes structure searching
- In development
- Additional data, searching

New data

- Quantitative toxicity data
- Numeric chemical property data
  - collected, calculated, contributed by Syracuse Research
  - melting point, boiling point, solubility, logP, Henry's Law
- Provide browsing of 3D conformations
- Searchable using the new 'heavy' version

New Documentation

- Provide more comprehensive ‘how-to’ guide
- Hypothesis: more users would search by structure if they knew how
- Offer pre-built searches

Longer Term Future (speculative)

- New data
  - Images
  - ADME predictions
  - ??? (an appeal)
- Enhanced searching
  - more complex queries
  - logic
- Cleaning up structures
  - experimentation using Cheshire

Challenges

- Keeping the server running
  - Hackers 2 years ago
  - ‘Scripters’ (a.k.a., non-traditional data downloaders)
    - Cause a huge load on the system
    - Block users at firewall level
    - Exclude CGI scripts at application level
    - Eventually, provide data download
  - Server overloads occasionally require restarts
  - Rapid bug fixes
- User support (remote problem diagnosis)
- Team communications
  - 4 or more organizations work together remarkably well
  - Weekly phone conferences
  - Lots of e-mail messages

Wrap-up

URLs

- Current system: http://chem.sis.nlm.nih.gov/chemidplus
  - Free, accessible to all
- Near-release ‘lite’ http://chem2.sis.nlm.nih.gov/chemidplus
  - Not guaranteed to be stable
- E-mail:
  - Project director: mike_hazard@nlm.nih.gov
  - Speaker: mitchell.miller@lionbioscience.com

The Team

- National Library of Medicine
  - George (Mike) Hazard, Jenny Fang, Vera Hudson, Philip Thomas
- Altum
  - Chris Hilt, David Mayer
- Cygnus
  - Larry Callahan, Jeff Zhang, Dave Rezvani
- MSD inc.
  - Sanjay Jain
- LION bioscience
  - Mitch Miller

Thank you all!

- Questions?