Application of Computer in Chemistry SSC 3533

DATABASES IN CHEMISTRY

Prof. Mohamed Noor Hasan
Dr. Hasmerya Maarof
Department of Chemistry

CC BY NC SA

Outline

- Types of Databases
- Examples of Databases in Chemistry
- Structural and sub-structural Searching
- Similarity Measures

Introduction

- Using computer to handle large amount of data associated with chemical compounds
- Information on a compound: literature, physicochemical properties, spectra, structures, reactions, etc.
- The system for storing and retrieving these data is called information system – consists of program and database
- Chemists need to search the database to find
  - Properties related to a structure
  - Compounds with similar structures
  - Compounds that contain certain functional groups or substructures

Types of Database

Literature Database

- Bibliographic
- Full Text
- Patent

Factual Database

- Numeric
- Metadata
- Directory
- Catalogs

Structure Database

- Structure
- Reaction

Literature Database

- **Chemical Abstract Service (CAS)**
  - Most comprehensive databases of disclosed research in chemistry and related sciences
  - Including a large collection of substance information, the CAS REGISTRY

- **ScienceDirect**
  - One of the largest online collections of published scientific research
  - over 8.5 million articles from over 2500 journals
  - over 6,000 e-books, reference works, book series and handbooks.

Literature Database

- **SCOPUS**
  - Covering abstracts and citations and web-based research tool
  - 15,800 peer-reviewed journals in the scientific, technical, medical and social sciences fields.
- **ISI**
  - Academic literature database
  - Provides access to many databases and other resources: covering about 8,700 leading journals

Factual Database

- **Beilstein Database**
  - Covers organic chemistry data from 1771 to date.
  - Based on Beilstein's Handbuch der organischen Chemie
  - Contains over 10.3 million structures, 10.6 million reactions, 2.1 million citations and 320 million property records.
  - Also contains over 900,000 original author abstracts from 1980-present,
  - EcoPharm database containing details of bioactive compounds

Factual Database

- **Gmelin Database**
  - Covering inorganic and organometallic compounds from 1772 to date.
  - Based on Gmelin Handbuch der anorganischen Chemie,
  - Comprises over 2.5 million compounds, including glasses, alloys, ceramics, minerals and coordination compounds, 1.9 million reactions and 1.3 million citations.
  - Also contains over 500,000 titles, abstracts and keywords.

CAS Registry

- Contains information on all the chemical compounds published in the literature since 1957
- More than 56 million organic/inorganic substances and more than 62 million sequences, updated daily
- Each substance is identified by CAS registry number
  - CAS Registry Number is a unique numeric identifier only for one substance
  - Has no chemical significance
  - For example, 58-08-2 is the CAS Registry Number for caffeine
- Available from: SciFinder, STN

Solutions for:

- Researchers
- IP Professionals
- Information Professionals
- Academics

Let us do your searching >>

Current Customers:

Find all of your product resources in our Support & Training section.

Latest News:

- STN Open Office Hours - 06/09

Comprehensive... Authoritative... Reliable

CAS, the global leader in chemical information, and a division of the American Chemical Society, provides the most comprehensive databases of disclosed research in chemistry and related sciences, including the **world's largest collection of substance information**, the CAS REGISTRY℠.

CAS makes this information available to researchers through SciFinder and STN, the best search and retrieval tools for scientists and information professionals.

Colors of Chemistry - Pistachio Green

Americans have loved ice cream ever since Dolley Madison served it at her husband's second inaugural in 1813. President Ronald Reagan even designated July as "National Ice Cream Month." From color and flavor to formulations, emulsifiers, and stabilizers, chemists must juggle many variables to deliver this deceptively simple treat.

CAS Launches Free Web-Based Resource "Common Chemistry" for General Public

"COMMON CHEMISTRY"™

Registry Number and Substance Counts

CAS is the leading provider of organic, inorganic, and sequence substance information.

The Latest CAS Registry Number® and Substance Count		
Date	**07/29/2009 10:14:32 EST**	
Count	**48,998,606**	organic and inorganic substances
	61,117,515	sequences
CAS RN	**1169929-67-2**	is the most recent CAS Registry Number

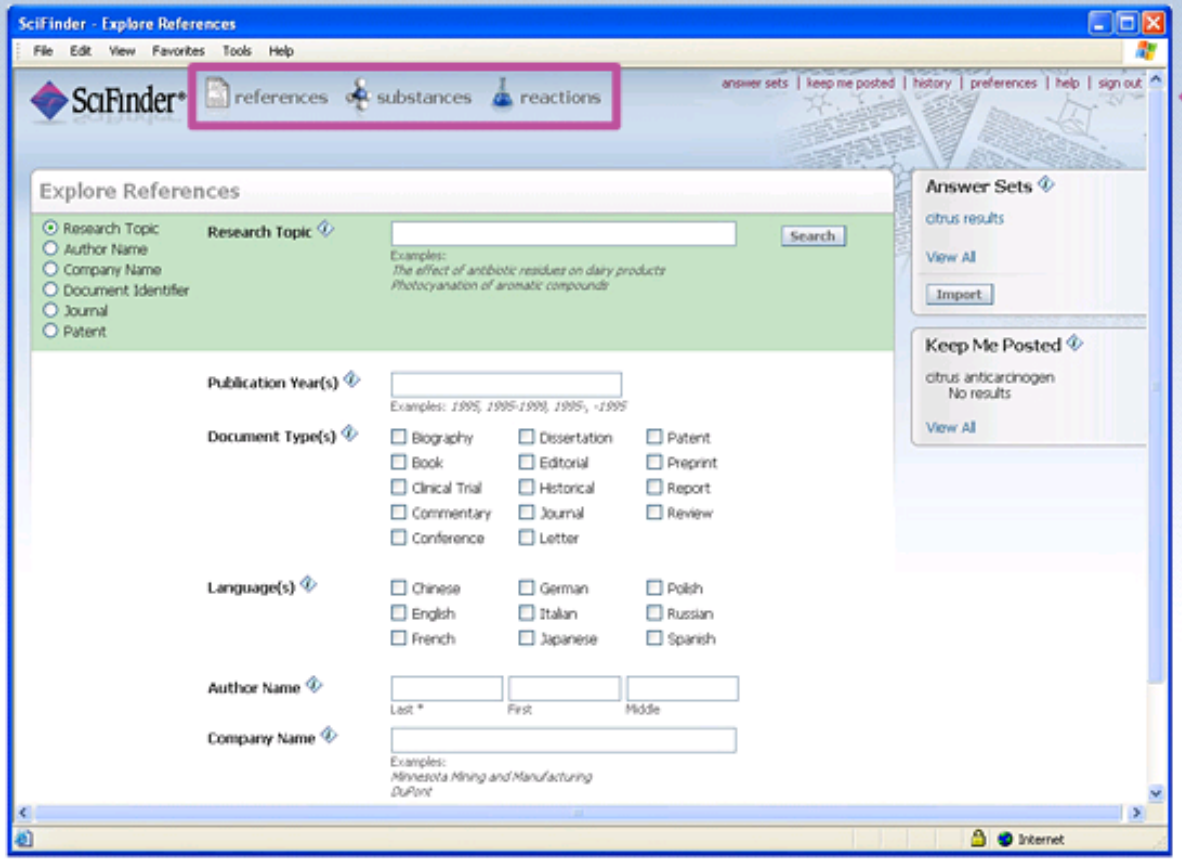
CAS also provides specialized databases of chemical reactions, regulated chemicals, commercially available chemicals and Markush substance information.

Specialized Substance Collections Count		
CASREACT®	**18,188,051**	Single and multi-step reactions
CHEMLIST®	**248,365**	Inventoried/regulated substances
CHEMCATS®	**34,497,943**	Commercially available chemicals
MARPAT®	**814,774**	Searchable Markush structures

SciFinder®

SciFinder Provides
Easy Access via the Web

REFERENCES SUBSTANCES REACTIONS



You can explore in three ways:

- Explore references
- Explore substances
- Explore reactions

*Please select an **explore option** from the menu above*

CAS
A division of the American Chemical Society

NCI Database

- Contains 250,251 structures
- Every record contains at least the NSC number and the chemical structure in connection table format
- Enhance searching http://129.43.27.140/ncidb2/
- http://nci.cambridgesoft.com/
- DTP Human Tumor Cell Line Screen: compounds tested for evidence of the ability to inhibit the growth of human tumor cell lines
- DTP AIDS Antiviral Screen: compounds tested for evidence of anti-HIV activity.

Public Access Databases

1. ChemDB
2. ChemSpider
3. eMolecules
4. NIST Chemistry WebBook
5. Zinc
6. PubChem
7. RCSB Protein Data Bank
8. TOXNET

PubChem Text Search

PubChem Compound | GO

PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. If you would like to learn more about how to use the PubChem resources, please go to our help page.

New PubChem3D is released consisting of a single theoretical 3D conformer per compound record and includes a new similarity relationship (Similar Conformers).

New XLogP is now updated to version 3.0 of the algorithm. InChI and InChIKey are now updated to use the version 1.0.2 standard.

More PubChem announcements ...

PubChem Compound: Search unique chemical structures using names, synonyms or keywords. Links to available biological property information are provided for each compound.

PubChem Substance: Search deposited chemical substance records using names, synonyms or keywords. Links to biological property information and depositor web sites are provided.

PubChem BioAssay: Search bioassay records using terms from the bioassay description, for example "cancer cell line". Links to active compounds and bioassay results are provided.

Structure Search: Search PubChem's Compound database using a chemical structure as the query. Structures may be sketched or specified by SMILES, MOL

ChemDB: The UC Irvine ChemDB

UCI ChemDB Featured Sections

ChemicalSearch: Find Chemicals by Various Criteria

Find a chemical by basic criteria like molecular weight and predicted logP, or by the more abstract notion of structural similarity.

Virtual Chemical Space: Retro-Synthesis and Combinatorial Library Design

Interactively deconstruct target compounds into component precursors and reconstruct similar building-blocks into combinatorial libraries representing the "virtual chemical space" near the target compound.

Reaction Explorer: Synthesis Explorer and Mechanism Explorer

Interactive system for learning and practicing reactions, syntheses and mechanisms in organic chemistry, with advanced support for the automatic generation of random problems, curved-arrow mechanism diagrams, and inquiry-based learning.

Datasets: For Machine Learning and Searching Experiments

Various available chemical datasets annotated with interesting properties to train and test machine-learning prediction and searching methods.

Supplements: Articles and Support Material

Information & Announcements

ChemDB Update Publication

Updates to the ChemDB search tools, including full-text search and virtual chemical space, are described in an article published in *Bioinformatics*. Please cite this article if you use any data or tools from the system.

Many recent articles on related subjects are available under the Articles & Presentations section.

Toolkits

Smi2Depict
Generates 2D Images from SMILES

Babel
Molecule File Format Converter

MolInfo
Calculate / Predict Molecular Properties

Reaction Processor
Product Library Generation

Pattern Match Counter
Counts Functional Groups (sub-structures)

Pattern Count Screen
Screens Molecules by Functional Group Count

MSFragment

ocw.utm.my

ChemSpider™

Building Community for Chemists

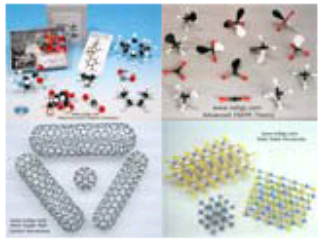
Search...

Home | Search | Services | Resources | About | Login

- Home
- Search
- Predictions
- Deposit Data
- ChemSpider Blog
- ChemSpider News
- Spinneret Webzine
- Open Chemistry Web

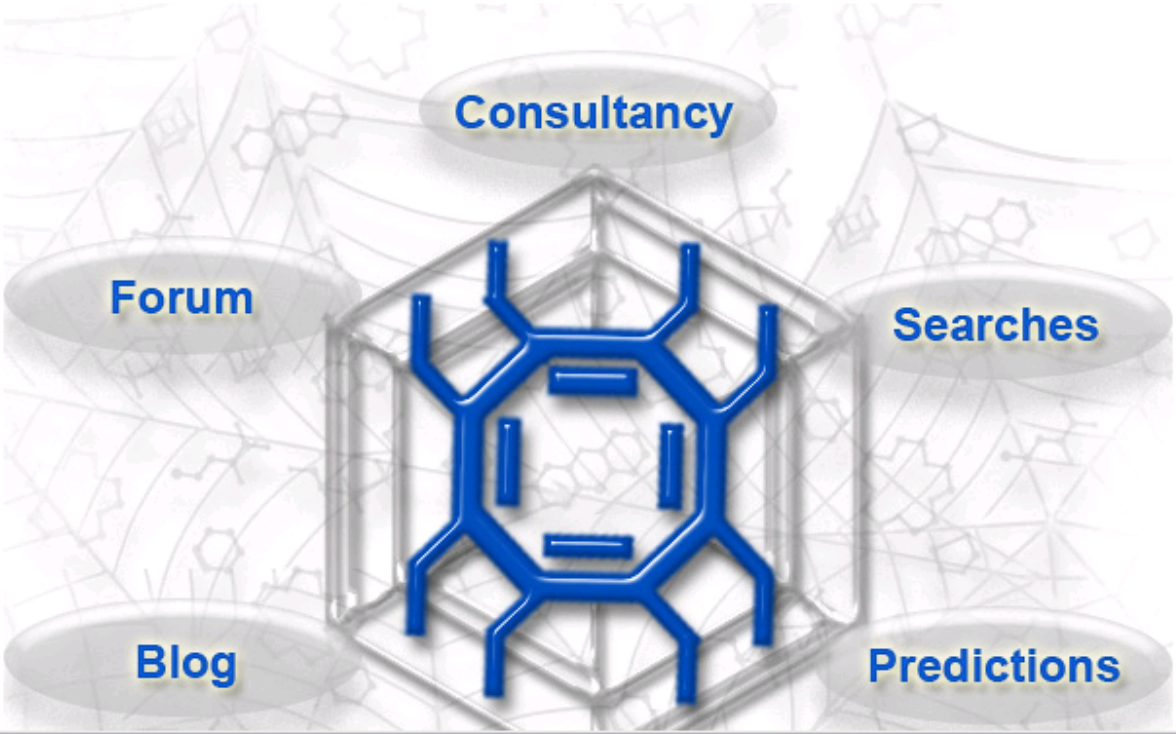
Ads on ChemSpider

ChemSpider Consultancy



Prestwick Chemical

ChemSpider is a free access service providing a structure centric community for chemists. Providing access to millions of chemical structures and integration to a multitude of other online services ChemSpider is the richest single source of structure-based chemistry information.



Advertise Sponsor

Gold Sponsors

Agilent Technologies

Thermo Scientific

Silver Sponsors

Waters
THE SCIENCE OF WHAT'S POSSIBLE.™

Chemical Patent Search

Bruker

Bronze Sponsors

ocw.utm.my

eMolecules

Find **Suppliers and Information** for over 10 Million Unique Chemicals!

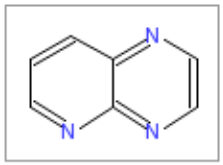
Want to see Prices?

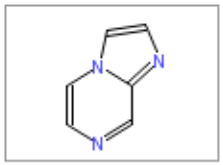
Can't find your favorite vendor? Request it!

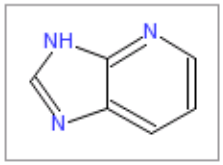
Shop ChemDiv Online!

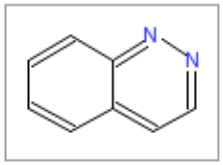
- Daily availability updates
- **Next-day shipping**
- 25% off web orders!

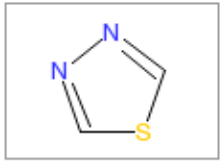
Try These Example Searches:

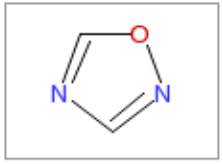

Pyrido-pyrazines

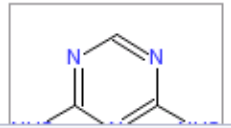

Imidazo-pyrazines

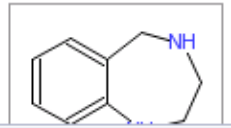

Imidazo-pyridines


Cinnolines

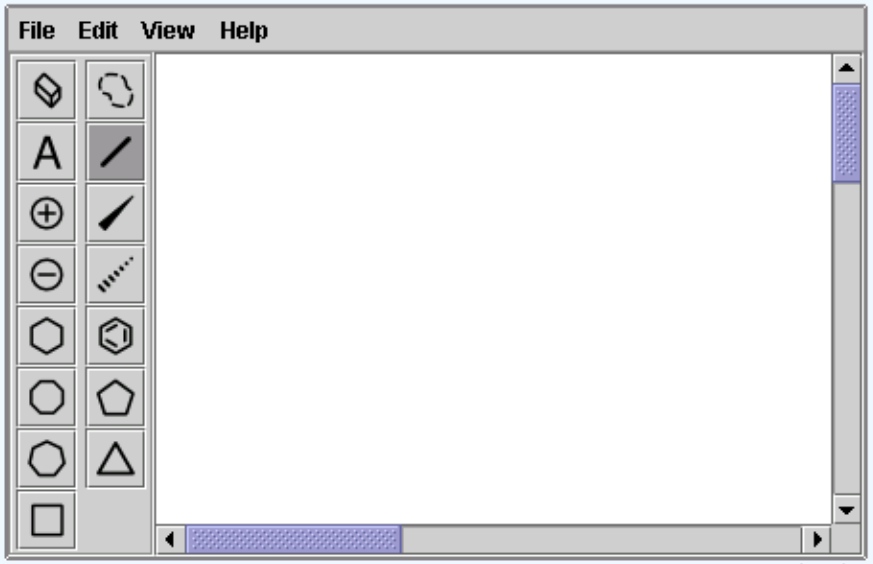

Thiadiazoles


Oxadiazoles





Structure Search



pick editor

Choose a Category:

- [x] Building Blocks
- [x] Screening Compounds
- [ ] All Compounds

(what are categories?)

Substructure Search

Exact Structure Search

Choose Specific Supplier

Search Named Chemicals

Search

Name: Advil, Ibuprofen CAS Num: 15687-27-1 SMILES: S=C=NC

Search from a List

Upload a File... Enter a List...

Did you know...

For **common elements**, just move the cursor over the atom and type its lowercase symbol: c, h, b, n, o, p, s, f, I (for Cl), r (for Br), i, and t (for Sn).

Mouse over the atom and type "**N**":

next hint...

eMolecules for Suppliers

We can boost your online chemical sales! Add your catalog, advertise, create your own e-commerce solution!

Learn more...

Featured Suppliers

TimTec – Diverse compounds

Analyticon – Natural derivatives

Focus Synthesis – Precursors

Pharmacore – Small molecules

Otava Chemicals – Fine organics

Life Chemicals – Targeted

ocw.utm.my

UTM

RCSB PDB
PROTEIN DATA BANK

A MEMBER OF THE PDB

MyPDB: Login | Register

An Information Portal to Biological Macromolecular Structures

As of Tuesday Feb 10, 2009 there are 55795 Structures | PDB Statistics

Quick Tips :

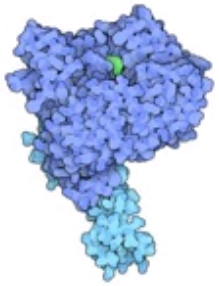
Try the *Web Services API* for software developers using C/C++, Java, Python and Perl. Click here.

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the **wwPDB**, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

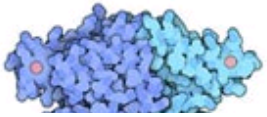
Molecule of the Month: Auxin and TIR1 Ubiquitin Ligase



Plants, like animals, have hormones that deliver chemical messages between distant cells. Charles Darwin and his son discovered this over a century ago--they noticed that if they shined a light on the tips of grass shoots, the stems bend to bring the entire shoot towards the light. Somehow, a message was being sent from the tip down to the stem. You might also have observed the action of hormonal signals in plants: when you prune a tree to make it more bushy, you are modifying the traffic of plant hormones. Both of these effects are caused by the phytohormone auxin.

Read more ... | Previous Features

PSI Featured Molecule: CBS domain protein TA0289



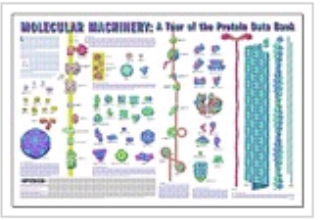
Researchers at the PSI MCSG have recently determined the structure of a protein with a new combination of two familiar protein modules, and made the first steps towards uncovering the function of this unusual new family of

News

- Complete News
- Newsletter
- Discussion Forum
- Job Listings

10-February-2009

Tools for Education



The General Education section of the RCSB PDB offers a variety of resources for teachers and students interested in learning about protein and nucleic acid structures.

More >>

02-December-2008

PDB Archive Version 3.15 to be Released

More >>

NMRShiftDB Links

Developers' page
Sponsoring
Media coverage
Static name list
Links
FAQ
Contact

Hall of Fame

	Name	Contributions
1	E. Willighagen	921
2	S. Dathe	505
3	P. Braeutigam	439
4	S. Kuhn	389
5	N. Prakash	350
6	B. Patel	305
7	M. Gericke	181
8	N. Kuznik	120
9	K. Bohn	111
10	R. Ellinger	76
11	A. Dransfeld	56
12	K. Bartusseck	26
13	M. Mitchell	20
14	J. Bitzer	19
15	L. Ernst	17

About NMRShiftDB

NMRShiftDB is a NMR database (web database) for organic structures and their nuclear magnetic resonance (nmr) spectra. It allows for spectrum prediction (13C, 1H and other nuclei) as well as for searching spectra, structures and other properties. Last not least, it features peer-reviewed submission of datasets by its users. The NMRShiftDB software is open source, the data is published under an open content license. Please consult the documentation for more detailed information.

News about NMRShiftDB

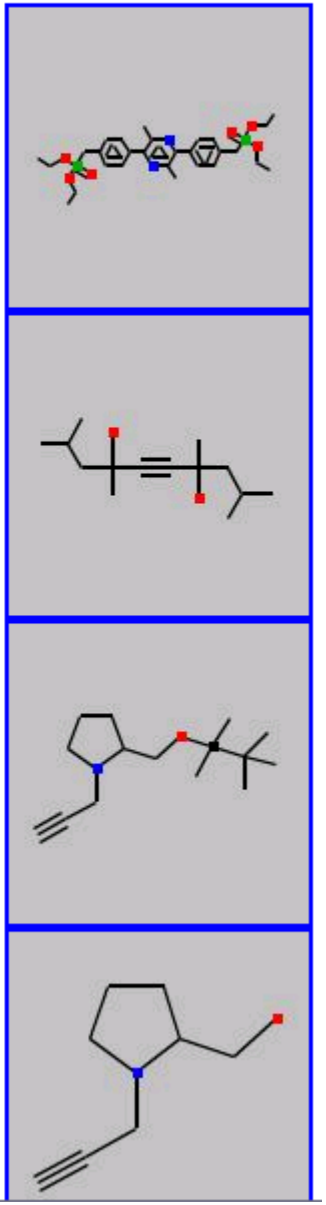
NMR prediction paper published 2009-01-28 11:32 - NMRShiftDB
A paper on prediction of 1H-NMR spectra using the data from NMRShiftDB has been published in BMC Bioinformatics.It can be read electronically on here.
Read More »

Bioclipse-based client available 2008-06-18 15:45 - NMRShiftDB
Speclipse, the new standalone client for NMRShiftDB, based on Bioclipse and therefore Eclipse, is available for Windows and Linux systems at here - it provides convenient access to NMRShiftDB functions including offline editing of entries.
Read More »

CUBIC NMRShiftDB server moved 2008-04-15 15:39 - NMRShiftDB
The server running at CUBIC was moved to the NMR labs of the University of Mainz, thanks to the help of the NMR department of Organic Chemistry running these facilities. It can now be reached at here
Once we've completed the OS upgrade of our main servers at the Max-Planck-Institute for Chemical Ecology in Jena, the Mainz server will be integrated into the load scheduling system and you may get automatically redirected to it again.
Read More »

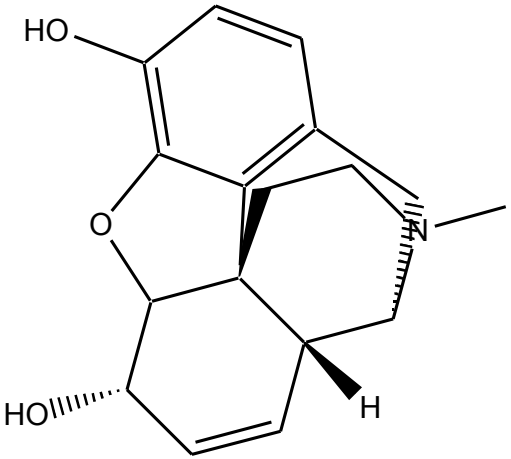
Server Upgrade 2008-04-15 12:22 - NMRShiftDB

Latest Addition

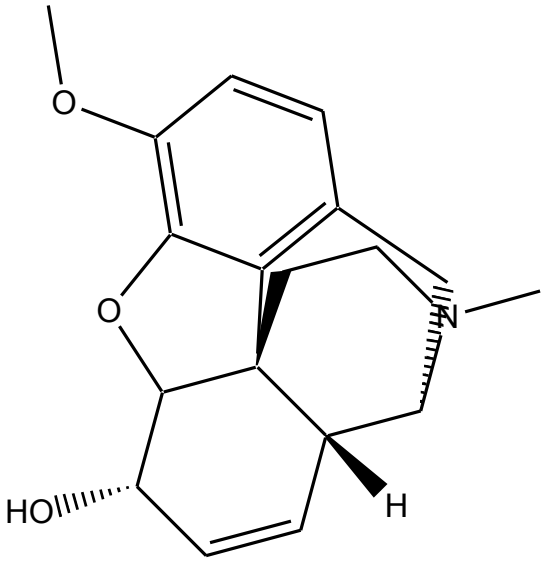


Similarity

- Similarity property principle
- Structurally similar molecules are expected to exhibit similar physical properties or, similar biological activities

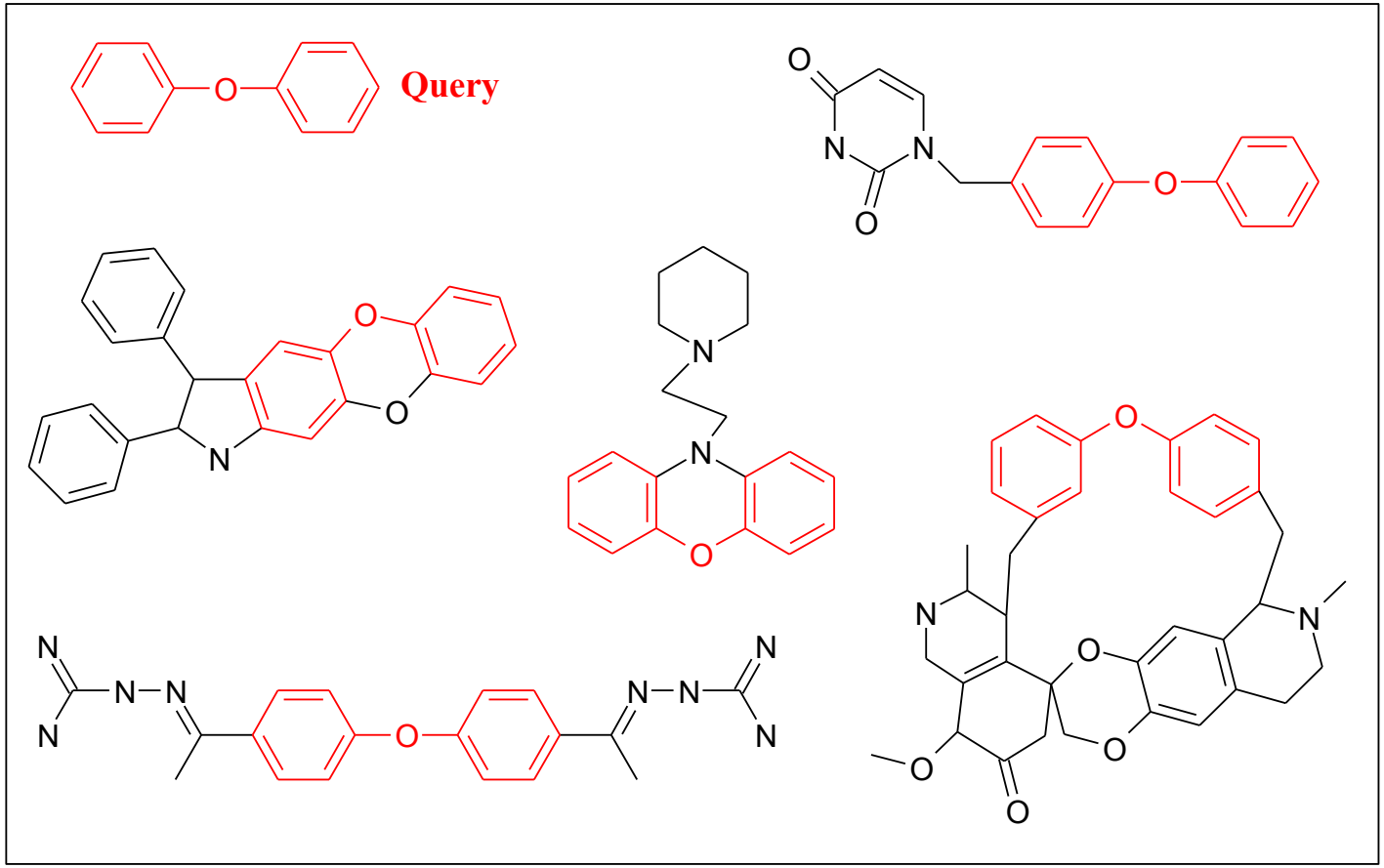


Morphine



Codeine

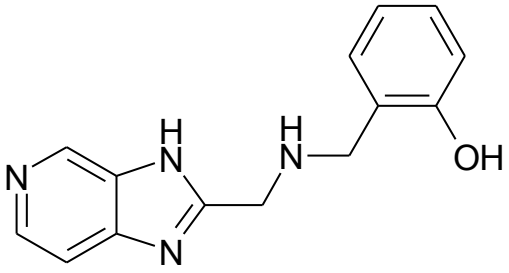
Substructure Searching

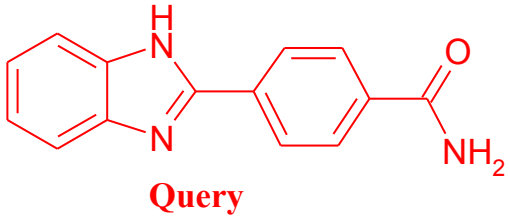


Substructure Searching

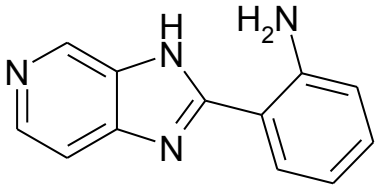
- Identifies all the molecules in the database that contain a specified substructure
- Example: identify all structures that contain a particular functional group such as carboxylic acid, benzene ring or C_5 alkyl chain
- The first step of a substructure search is a screening process to eliminate molecules that cannot possibly match the substructure query. One method of screening is by the **bitstrings** representation of query structure with molecules in the database

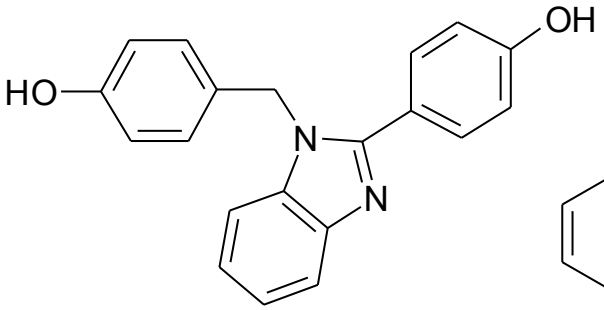
Structure Searching

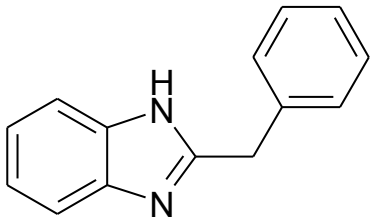


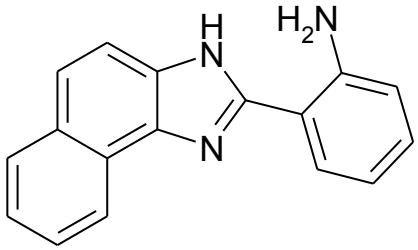


Query









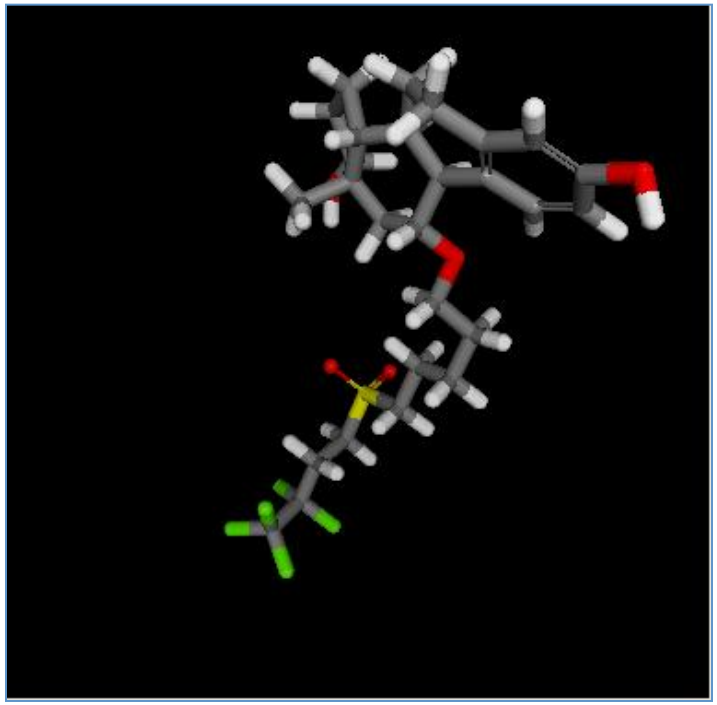
Structure Searching

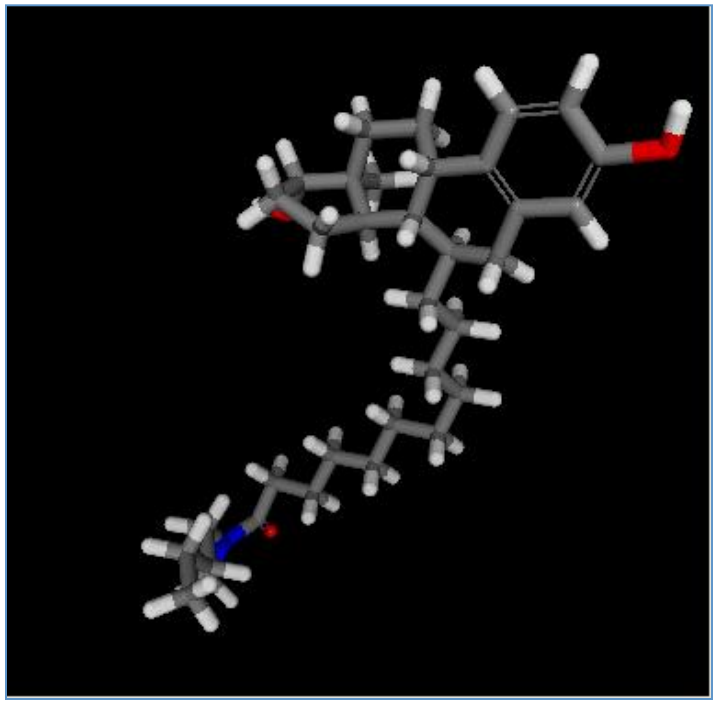
- Molecules that are similar in structure probably have similar properties.
- We can use similarity coefficient as a measure of similarity between two structures
- Example of similarity coefficients: Euclidean distance, Tanimoto coefficient, Manhattan distance

Reference:

- Willett, P., Barnard, J.M., Downs, G.M. "Chemical similarity searching" J. Chem. Inf. Comput. Sci. 1998, **38**, 983-996.

The Similarity Problem

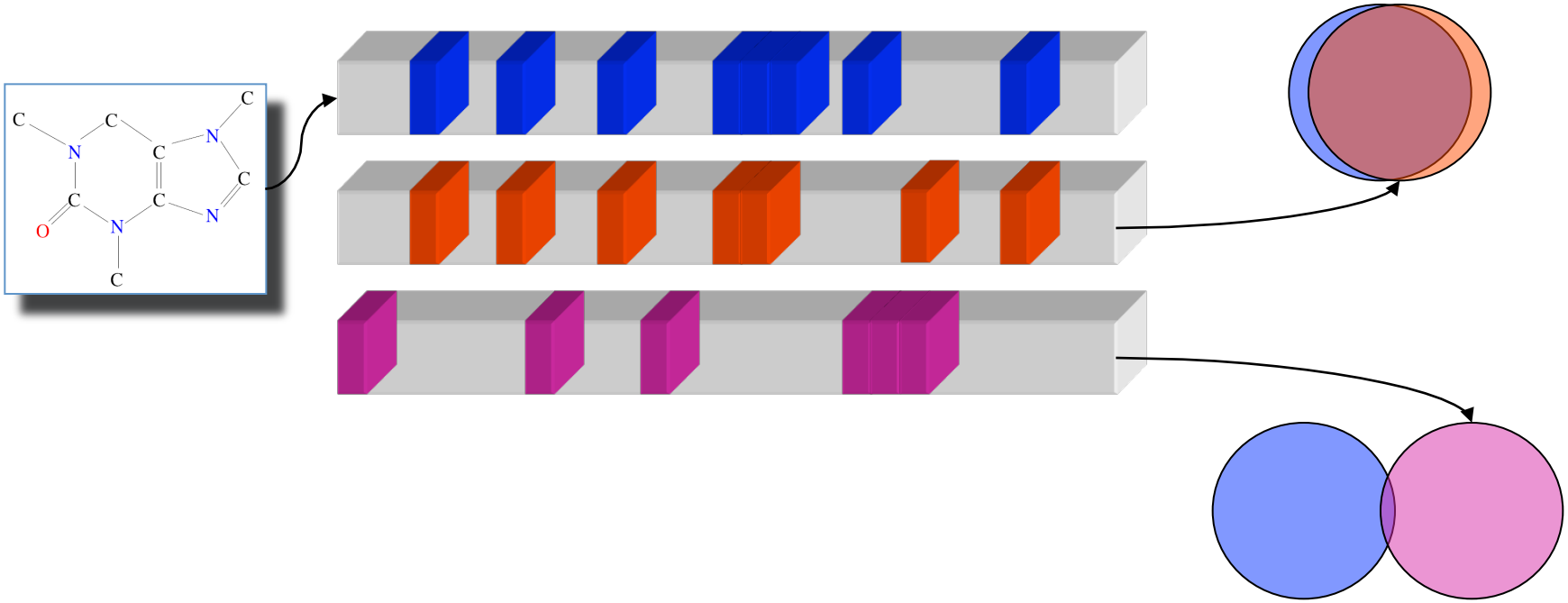




How similar?

The Fingerprint Approximation

- Fingerprint bit similarity approximates chemical feature similarity.

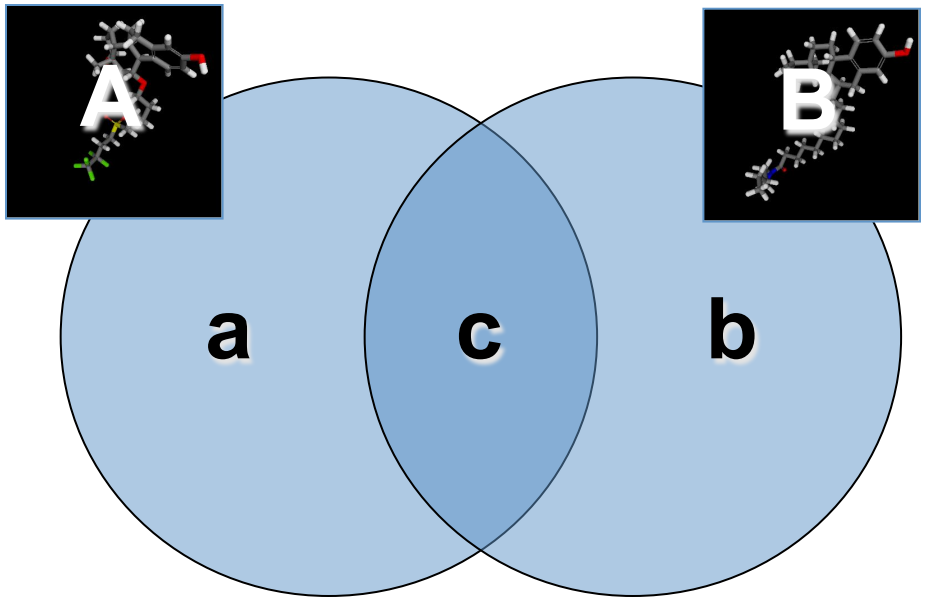


Similarity Measures

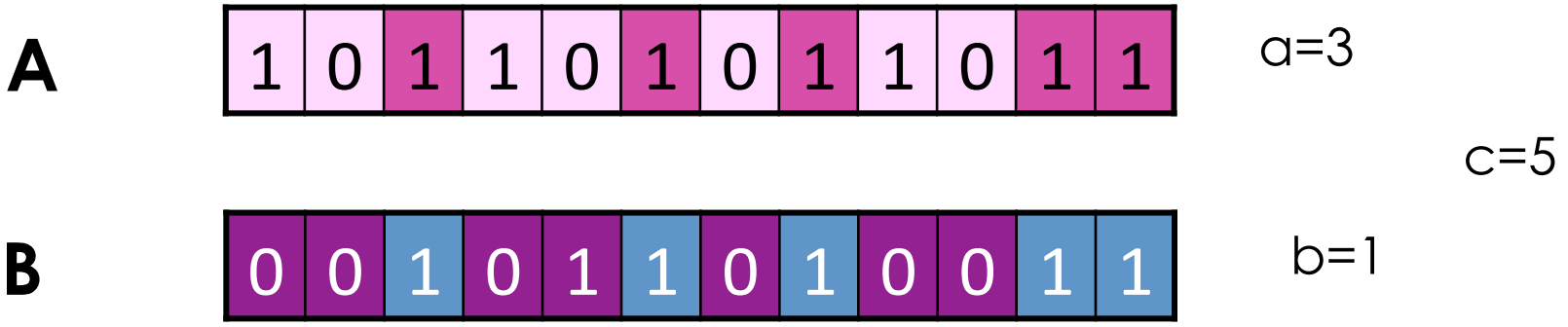
- There are many ways of measuring similarity (or distance) between bit/count vectors:
  - Euclidean
  - Cosine
  - Exponentials
  - Tanimoto/Jaccard
  - Tversky
  - MinMax
  - Hamming/Manhattan

Similarity Measures: Tanimoto

- Tally features:
  - Unique (a,b)
  - Both on (c)
  - Both off (d)
- Similarity Formula
  - Tanimoto=c/(a+b+c)



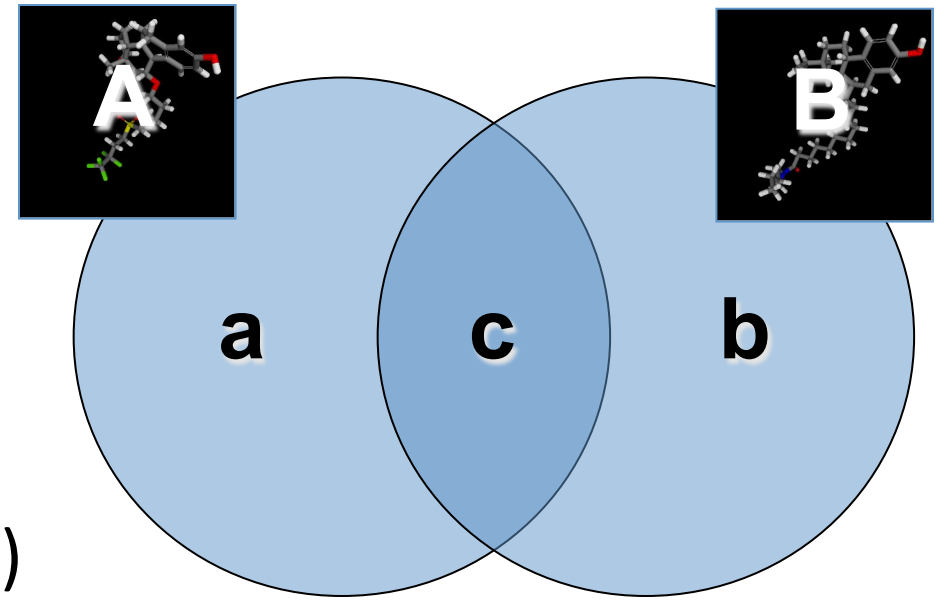
Tanimoto example



$$S_{AB} = 5/(3 + 1 + 5) = 0.56$$

Similarity Measures: Tversky

- Tally features:
  - Unique (a,b)
  - Both on (c)
  - Both off (d)
- Similarity Formula
  - Tanimoto=c/(a+b+c)
  - Tversky(α,β)=c/(αa+βb+c)



Similarity Measures

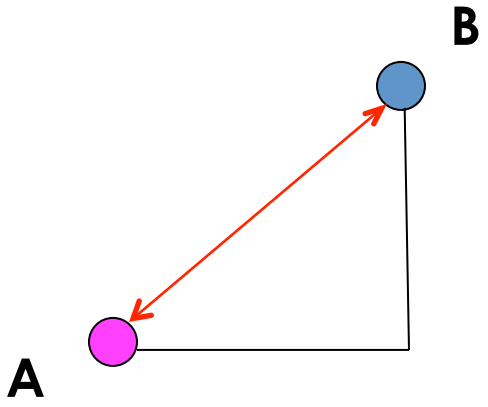
Measure	Range	Formula
Cosine	0.0,1.0	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
Dice	0.0,1.0	$\frac{2.0*c}{(a+c)+(b+c)}$
Euclidean	0.0,1.0	$\sqrt{\frac{c+d}{a+b+c+d}}$
Hamming/Manhattan	1.0,0.0	$\frac{(a+b)}{(a+b+c+d)}$
Tanimoto/Jaccard	0.0,1.0	$\frac{c}{a+b+c}$

Distance Measures

- Euclidean distance

$$D_{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2 \right]^{1/2}$$

- Range: 0 to $\propto$



- Hamming (Manhattan) Distance

$$D_{AB} = \sum_{i=1}^{N} |x_{iA} - x_{iB}|$$

- Range: 0 to $\propto$