Data Mining in the Chemical Industry

Glenn J. Myatt, Ph.D.
Partner, Myatt & Johnson, Inc.
glenn.myatt@gmail.com

Overview of presentation

- Overview of the chemical industry
  - Example of the pharmaceutical industry
- Chemistry-based data mining
  - Parsing, representation, matching, generating descriptors, data mining approaches
- Examples of how data mining is used within the chemical industry
  - Identifying diverse compounds
  - Analysis of high throughput screening
  - Safety prediction

Overview of the chemical industry

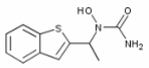
Chemical industry overview

- The **chemical industry** refers to an industry involved in the production of chemicals
- The industry includes:
  - petrochemicals
  - agrochemicals
  - **pharmaceuticals**
  - polymers
  - paints
  - oleochemicals

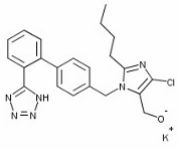
Pharmaceutical industry

- World's largest manufacturing industry
- Pharmaceutical sales: $534.8 billion (2005)
  - USA(~41%); Europe(~24%); Japan(~13%)
- R&D Time and Costs
  - 13.6 years to discover a new drug
  - Rising costs of drug discovery
    - 1976 - $54 million
    - 1987 - $231 million
    - 2002 - $802 million
- Only 1/3 of known diseases can be treated effectively

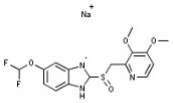
Examples of drugs



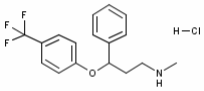
Zyflo by Abbott
Anti-asthmatic agent



Cozaar by Merck
Treatment for hypertension and congestive heart failure

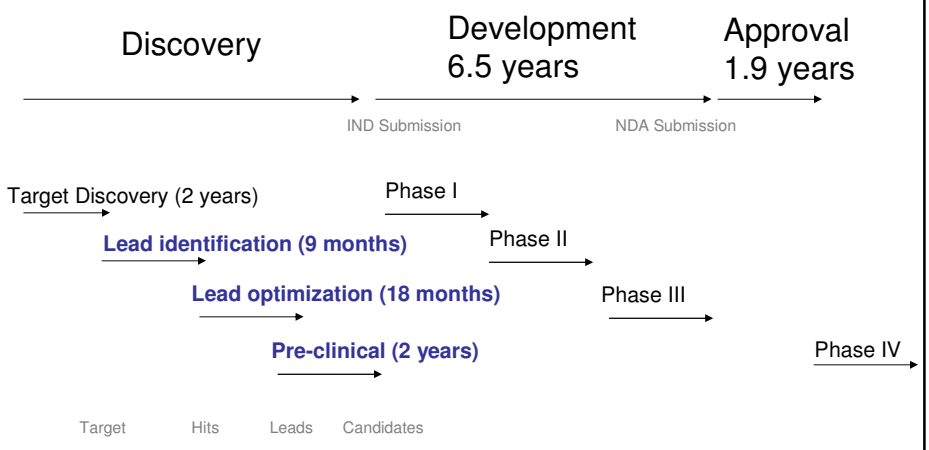


Protonix by Wyeth
Anti-ulcer agent



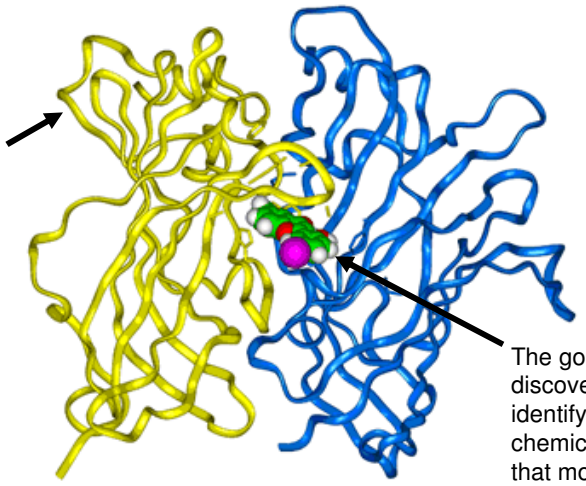
Prozac by Lilly
Antidepressive agent

Drug discovery process



Target discovery

A **target** is a protein that plays a role in the disease

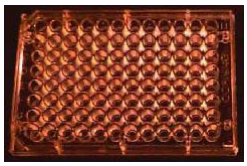


The goal of drug discovery is to identify a new chemical (**drug**) that modifies the behavior of the protein

Lead identification

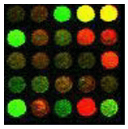
- Pharmaceutical companies have historical collections of chemicals (1 million approx.)
- These chemicals will be screened against the target assay (test that indicates whether a chemical modifies the behavior of the target)
- A chemical showing a positive response is considered a hit at this stage
- Lead series (sets of similar chemicals) will be uncovered through an analysis of the data (both chemical and biological screening data)

High throughput screening



The test is performed on small plates

The process of screening the chemicals is automated



The results of the testing are automatically read

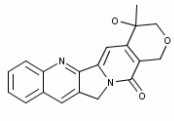
Screening results

By analyzing the screening data, lead series can be identified

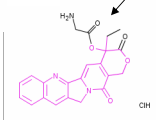
Class

Biological screening data

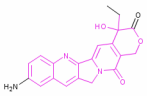
Chemical



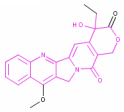
8.0



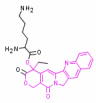
8.0

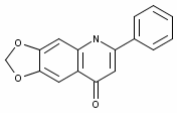


7.92

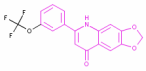


7.66

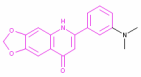




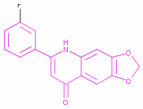
6.92



6.49



6.08



Lead optimization

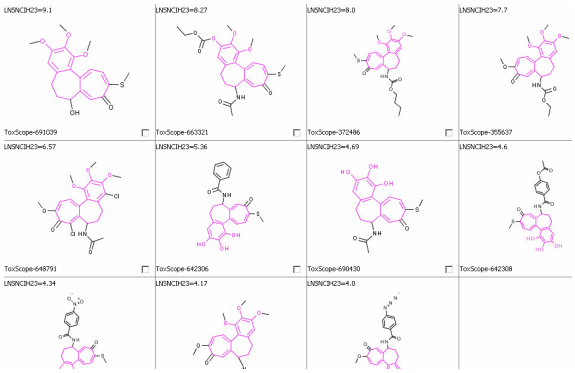
Lead series

Design new compounds

Synthesize or acquire compounds

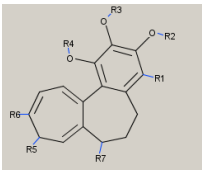
Screen

Lead optimization



Synthesize close analogs to determine how substituents affect the biological response data

Lead optimization



Optimize over multiple rows of data:
(1) primary screening data
(2) selected screening data
(3) ADME screening data

R1	R2	R3	R4	R5	R6	R7			
H	Me	Me	Me	=O	SMe	OH	9.10	9.15	9.23
H	Me	H	H	=O	OMe	NHAc	8.00	4.40	4.73
H	Me	Me	Me	=O	OMe	NHAc	7.20	7.20	6.78
CHO	Me	Me	Me	=O	OMe	NHAc	6.92	2.91	4.00
H	Me	Ac	Ac	=O	SMe	NHAc	5.65	5.91	4.00
H	Me	Me	Me	=O	OMe	C10…	5.11	4.43	4.38
H	Me	Me	Me	=O	H	NHAc	4.82	4.00	7.41
H	H	H	Me	=O	OMe	NHAc	4.43	4.61	9.40

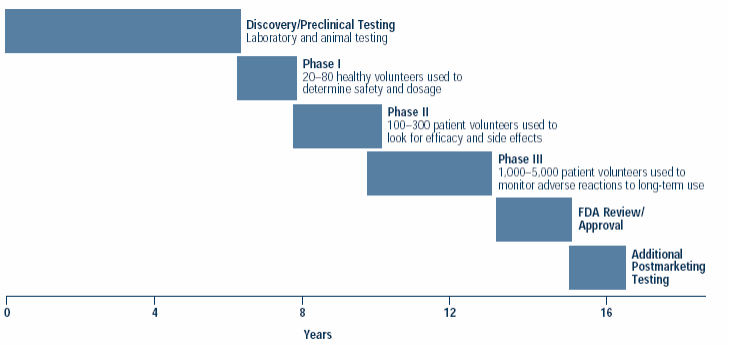
Pre-clinical

- Use of model system
- Absorption
- Distribution
- Metabolism
- Elimination
- Toxicity (safety)

IND submission and patent application

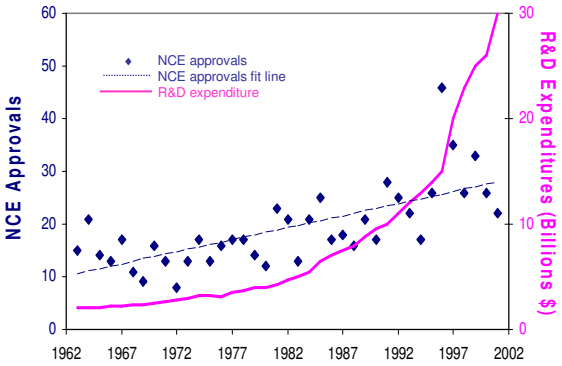
- Prior to clinical trials –
  - submit IND to Regulatory Agencies (FDA)
    - Allow company to conduct a clinical trial
  - Patent
    - Exclusivity period from approval date

Drug development



Source: Pharmaceutical Research and Manufacturers of America, based on data from Center for the Study of Drug Development, Tufts University, 1995.

Issues



R&D expenditures adjusted for inflation
Source: Tufts CSDD Approved NCE Database, PhRMA

Compounds withdrawals

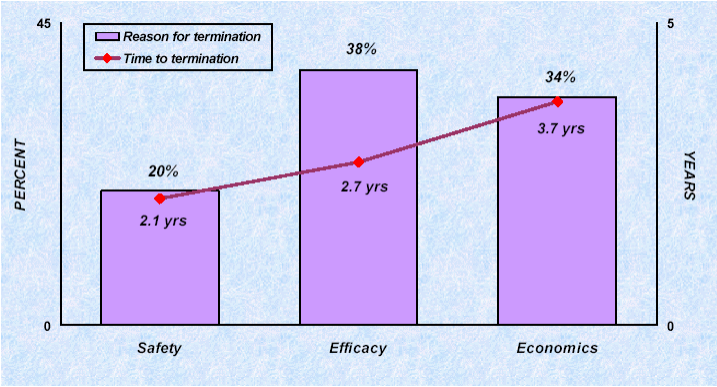
Safety-Based Drug Withdrawals Over Last 5 Years

1997	1998	1999	2000	2001
	Seldane		Rezulin	
Redux	Duract	Hismanal	Propulsid	Raplon
Pondimin	Posicor	Raxar	Lotronex*	Baycol

Combined peak sales potential for these drugs alone was in excess of $11 Billion

Source: IBM

Why drugs fail?



Source: DiMasi, Clin Pharm Ther, 2001;69:297-307

Issues

- Making sense of large volumes of data
  - Genomics, HTS, Lead optimization, ADME(T), Clinical trials
- Integration of data across silos
- Accelerating the pace of drug discovery
- Reducing compound attrition

Chemistry-based data mining

Data mining in chemistry - Chemoinformatics

Journals

Journal of Chemical Information and Computer Sciences
Journal of Computer-Aided Molecular Design
Journal of Molecular Graphics and Modelling

Books

An Introduction to Chemoinformatics (Hardcover) by Andrew R. Leach, Valerie J. Gillet

Chemoinformatics: A Textbook (Hardcover) by Johann Gasteiger (Editor), Thomas Engel (Editor)

Meetings

International Conference on Chemoinformatics
Discovery Knowledge & Informatics 2007
The Fourth Joint Sheffield Conference on Chemoinformatics

Virtual Discovery. Computer-Aided Drug Design and Screening

Web sites

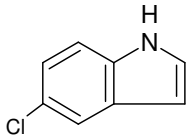
http://www.chemoinf.com/

http://www.iainm.demon.co.uk/indexnew.htm

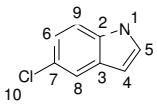
Issues to consider when data mining chemical data

- Parsing
  - Chemical data
  - Related information
- Matching
  - Exact
  - Substructure
  - Similarity
- Descriptors
- Data mining methods

How to convert a chemical into a form to be read by a computer



Use of a connection table



Number of atoms: 10
Number of bonds: 11

Atoms:

Atoms ID	Type
1	N
2	C
3	C
4	C
5	C
6	C
7	C
8	C
9	C
10	Cl

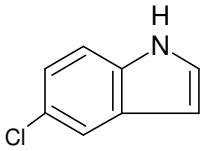
Bonds:

From atom	To atom	Type
5	1	Single (1)
2	3	Single (1)
6	7	Double (2)
1	2	Single (1)
7	8	Single (1)
8	3	Double (2)
3	4	Single (1)
2	9	Double (2)
9	6	Single (1)
4	5	Double (2)
7	10	Single (1)

Computer readable representations of chemicals

- Describe the connection table in a computer readable form:
  - **MOLFILE and SD File**
    - Contain information on the atoms (including coordinates), bonds, connections and associated information
  - SMILES, WLN, CML, ….

MolFile example

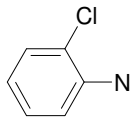


Chemical Structure-1

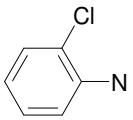
```
Chemcial Structure-1

10 11  0  0  0  0  0  0  0  0999 V2000
    3.8226   -0.4964    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.0380   -0.7541    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0418   -1.5806    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8297   -1.8311    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.3101   -1.1585    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6194   -0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6141   -1.5732    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3232   -1.9875    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3297   -0.3428    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.8958   -1.9792    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  5  1  1  0  0  0  0
  2  3  1  0  0  0  0
  6  7  2  0  0  0  0
  1  2  1  0  0  0  0
  7  8  1  0  0  0  0
  8  3  2  0  0  0  0
  3  4  1  0  0  0  0
  2  9  2  0  0  0  0
  9  6  1  0  0  0  0
  4  5  2  0  0  0  0
  7 10  1  0  0  0  0
M  END
```

Matching chemical structures

Aromaticity

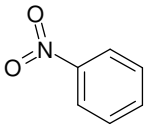


A

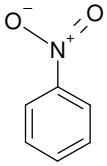


B

Representation

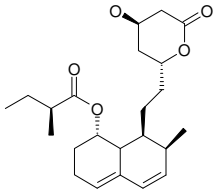


A

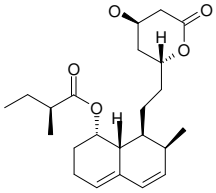


B

Stereochemistry

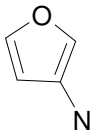


A

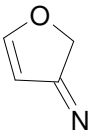


B

Tautomerism

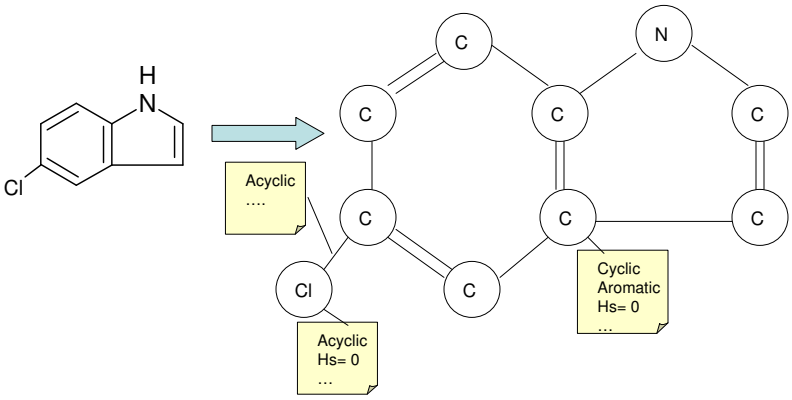


A



B

Need to annotate atoms and bonds (and the whole compound) to effectively search



All atoms and bonds can be annotated with additional information

Graphs

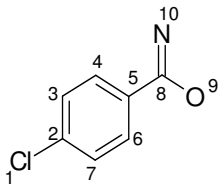
- Nodes (atoms) and edges (bonds) have properties
  - Non-calculated
    - E.g.; Charge, bond type, atoms type, ….
  - Calculated
    - Cyclic, number of hydrogens, ….

Perception of chemical features

- Rings and chains
- Aromaticity
- Stereocenters
- Olefinic Double Bonds
- **Hydrogens**
- Hybridisation levels
- Canonicalization
- Symmetry
- ….

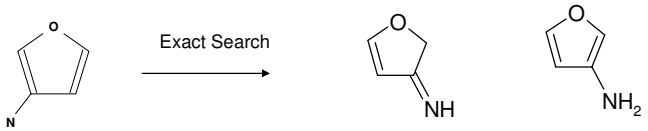
Hydrogen perception example

Now compound is in a graph we can easily determine the number hydrogens using the atom type, attached bonds and charge



Atom#	Type	Hydrogens
1	Cl	0
2	C	0
3	C	1
4	C	1
5	C	0
6	C	1
7	C	1
8	C	0
9	O	1
10	N	1

Exact matching

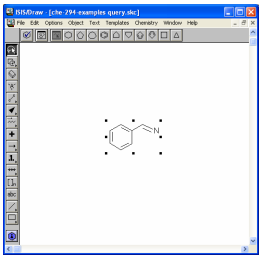


Query

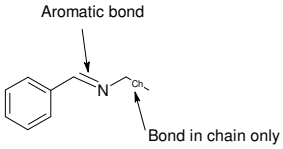
Chemicals selected from a database

Substructure queries

- Structure drawing packages
- Add additional restriction on the atom and bonds:
  - Cyclic/acyclic
  - Number of hydrogens
  - Closed to substitution
  - …..
- MOLFILE



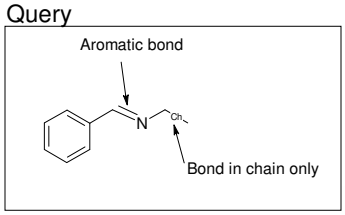
MOLFILE representation of a query

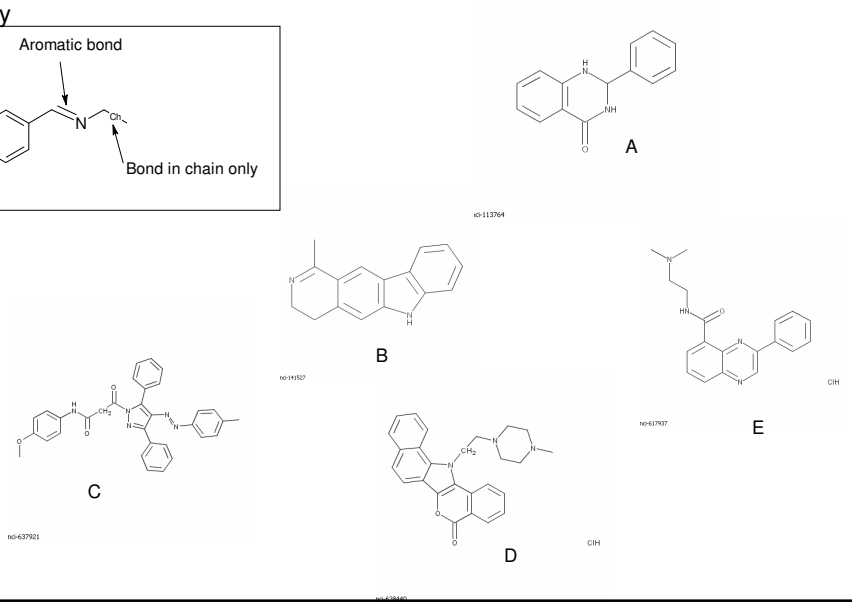



```
 -ISIS-  10290307032D

 10 10  0  0  0  0  0  0  0  0999 V2000
    4.8000   -2.8083    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    1.7444   -2.8125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7391   -3.6357    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4482   -4.0500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.1631   -3.6424    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.1645   -2.8160    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4547   -2.4053    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8750   -2.3958    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.5125   -2.3917    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.2250   -2.8042    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  5  6  1  0  0  0  0
  3  4  1  0  0  0  0
  6  7  2  0  0  0  0
  7  2  1  0  0  0  0
  2  3  2  0  0  0  0
  6  8  1  0  0  0  0
  8  1  4  0  0  0  0
  4  5  2  0  0  0  0
  1  9  1  0  0  0  0
  9 10  1  0  0  2  0
M  END
```

Substructure search example

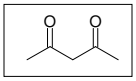
Query

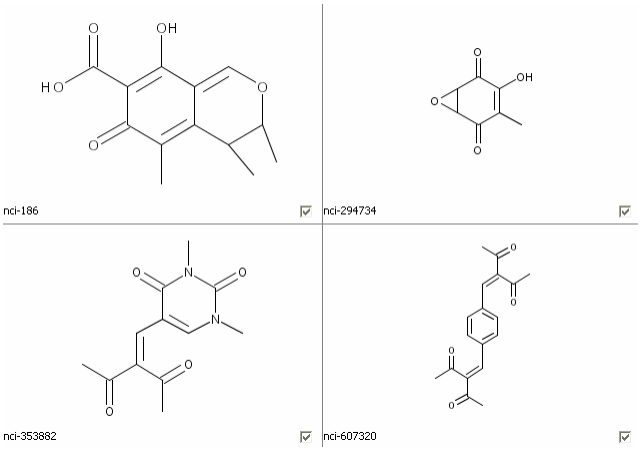




Substructure search example

Query



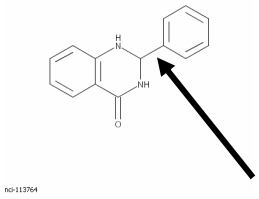


Molecular Descriptors

- Number of hydrogen bond acceptors
- Number of hydrogen bond donors
- Number of rings
- **Number of rotatable bonds**
- Molecular weight
- Hydrophobicity
- Molar refractivity
- Topological indices
- Kappa shape indices
- Electrotopological state indices
- Polar surface area
- **2D fingerprints**
- Atom-pairs and topological torsions
- Pharmacophore keys

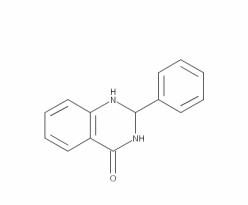
Rotatable Bonds

- Single
- Acyclic
- Non-terminal

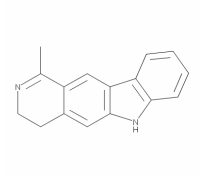


Rotatable bond

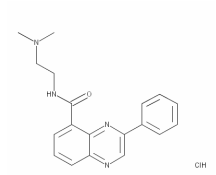
Calculating rotatable bonds



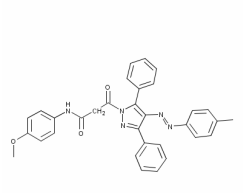
Rotatable bonds = 1



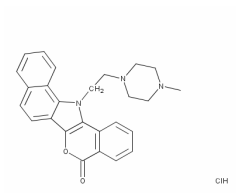
Rotatable bonds = 0



Rotatable bonds = 6



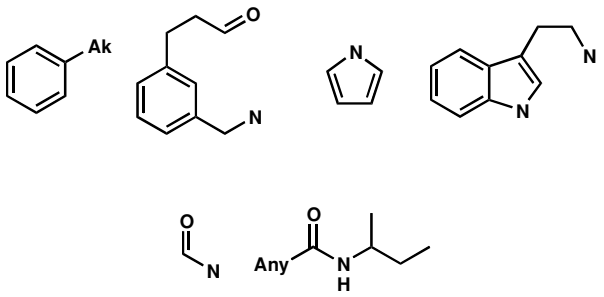
Rotatable bonds = 10



Rotatable bonds = 3

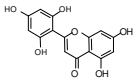
2D Fingerprints

Makes use of a dictionary of fragments (pre-defined substructures)



Describing chemicals using fragment descriptors

Fragment dictionary

Chemicals											
				— OH	HO	O	O	O O N	O	O O	.....
	0	1	1	1	1	1	0	0	0	0	
	1	0	0	1	0	0	0	1	0	0	
....											

Data Mining Methods

- Clustering
- Decision trees
- Principal component analysis
- Support vector machines
- kNN
- Decision forests
- Bayesian networks
- Genetic algorithms
- ……

Identifying diverse chemicals

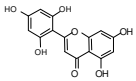
Selecting diverse chemicals from commercially available chemical collections is often used to supplement in-house screening sets

Approach to identifying diverse chemicals

- Select and calculate chemical descriptors
- Determine the similarity between chemicals
- Cluster chemicals based on these descriptors
- Select representative chemicals from each group generated

Describing chemicals using fragment descriptors

Fragment dictionary

Chemicals											
		O	O	— OH	HO	O	O	O O N	O	O O	.....
	0	1	1	1	1	1	0	0	0	0	
	1	0	0	1	0	0	0	1	0	0	
....											

Similarity

- Tanimoto is one example
  - $S_{AB} = c / (a + b - c)$
  - a is the number of bits set to one in A
  - b is the the number of bits set to one in B
  - c is the number of bits that are 1 in both A and B

Calculating similarity

Fragment dictionary

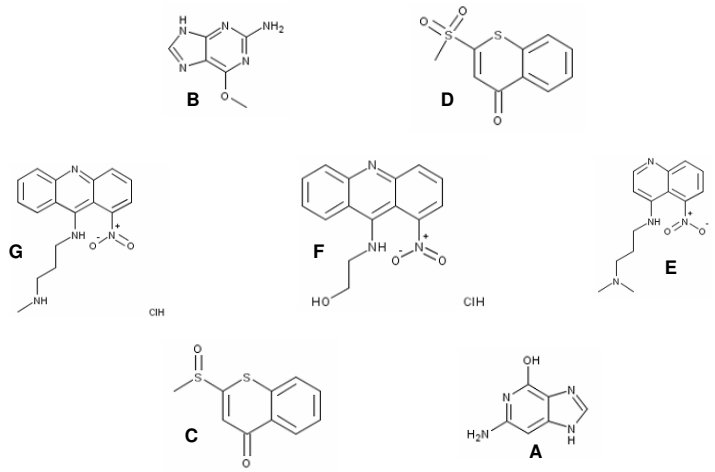
Chemicals											
				— OH							.....
	0	1	1	1	1	1	0	0	0	0	
	1	0	0	1	0	0	0	1	0	0	
....											

$S = c / (a + b - c)$

$S = 1 / (5 + 3 - 1)$

$S = 0.14$

Using 7 chemicals to illustrate



This type of analysis is usually performed on tens of thousands of chemicals

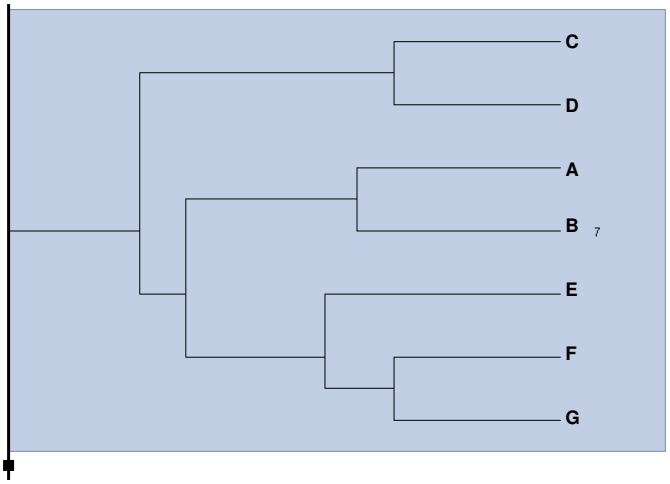
Fingerprint table

Chemicals

Fragment dictionary ids

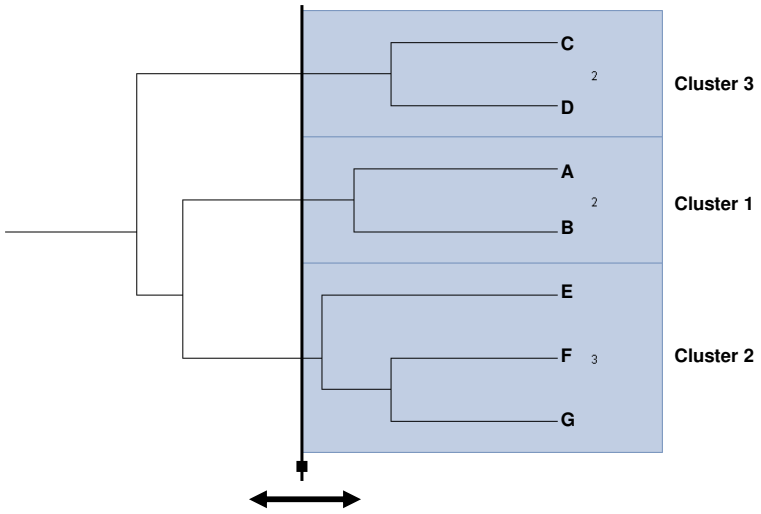
Chemical	5018	5035	5067	5081	5085	5292	5337	5400	5421	5432	5454	5497	5510	5568	5853	6509	6645	6883	6994	7132	12798	12942
A	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
B	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
C	1	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0
D	1	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0
E	0	1	0	0	1	0	0	0	0	1	1	1	1	0	0	0	1	0	0	0	0	0
F	0	1	0	1	1	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
G	0	1	0	1	1	0	0	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0

Hierarchical agglomerative clustering based on the fingerprints



Uses of the Euclidean distance and clusters using the average linkage joining rule

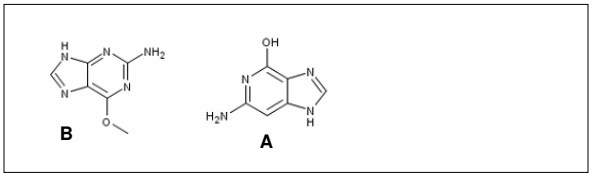
Generating clusters



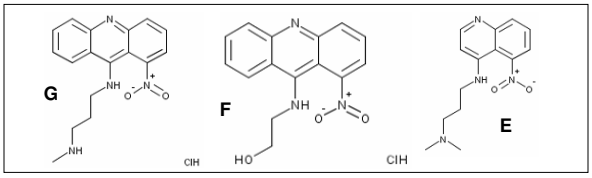
Adjusts the distance cut-off to change number of clusters

Cluster results at 0.7 cut-off

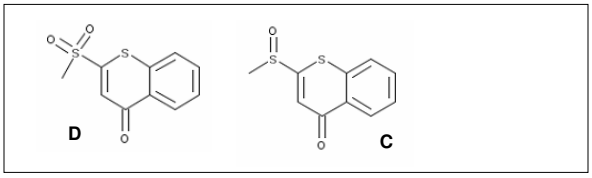
Cluster 1



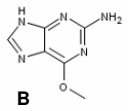
Cluster 2

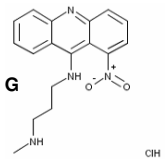


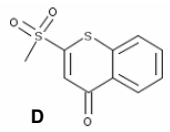
Cluster 3



Selecting a representative from each cluster







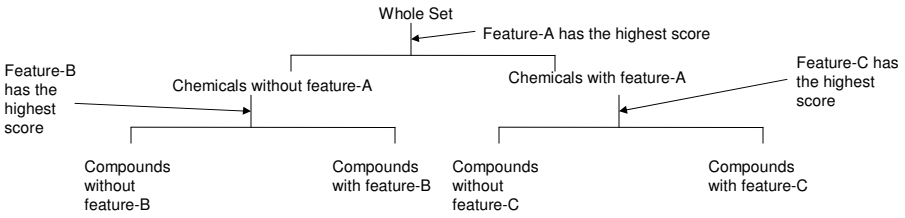
Analyzing HTS data

Use of decision trees

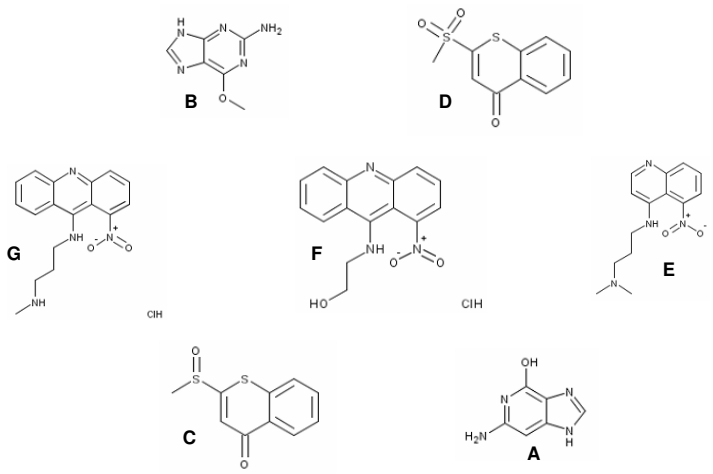
- Supervised learning approach
  - Partitions the set based on the descriptors
  - Uses the biological data to determine the groups
- Used to quickly identify biologically interesting groups of chemicals

Generating decision trees

1. Find the most significant feature to split Feature-A
2. Partition the set according to those compounds containing the feature and those without
3. Partition each child node until a threshold is reached



Using 7 chemicals to illustrate

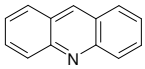
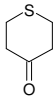


This type of analysis is usually performed on hundreds of thousands of chemicals

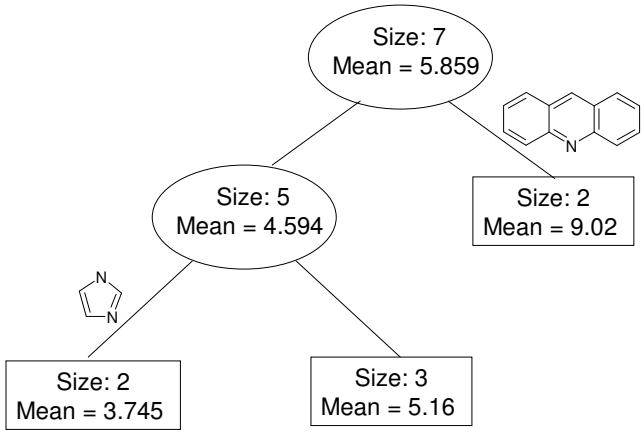
Potency data

Structure ID	Data
A	4.44
B	3.05
C	5.21
D	5.04
E	5.23
F	9.33
G	8.71

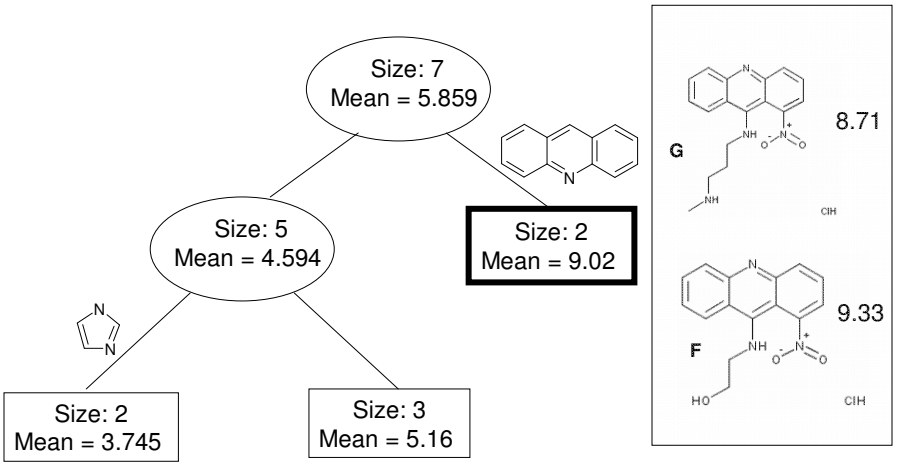
Fingerprints

							
A	0	0	0	1	1	0	1
B	0	0	0	0	0	0	1
C	0	0	0	0	0	1	0
D	0	0	0	0	0	1	0
E	0	1	1	1	0	0	0
F	1	1	1	1	1	0	0
G	1	1	1	1	0	0	0

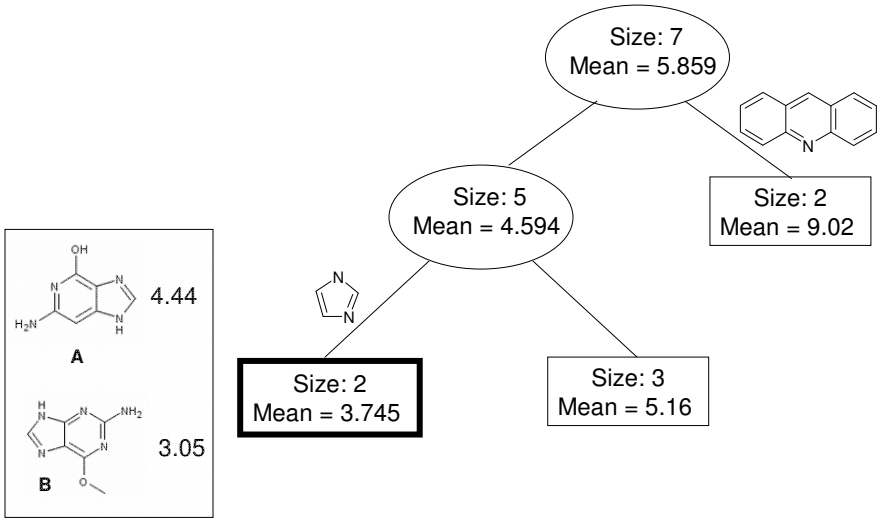
Decision Tree



Decision Tree



Decision Tree



Predicting chemical safety

Safety prediction example

- Prepare data
  - Integrate, normalize data, generate descriptors
- Prune descriptors
  - Remove constants, desciptors lacking in information
- Understand chemical space
  - Subset data
- Understand mechanisms of action
- Build and optimize model(s)
- Assess models
  - Evaluate quality
  - Combine models
  - Applicability domain
- Apply to untested chemicals, where chemical is within the domain of the model

Summary

- The chemical industry generates information about chemicals and its relationship to drug potency, safety, agrochemicals, …
- Data mining is used extensively to accelerate the development of new products
- Representing and describing chemicals is a large part of the challenge of data mining chemical information