# Drug Informatics for Chemical Genomics

...

## *An Overview*

First Annual

ChemGen IGERT Retreat

Sept 2005

# Topics

- ChemGen Informatics
- The ChemMine Project
- Library Comparison

# ChemGen Informatics

A Multidisciplinary Symbiosis

- Cheminformatics
- Drug Informatics
- Network Analysis & Modeling
- Structural Informatics
- Bioinformatics

# Cheminformatics

- Structure formats & format rendering
- Property descriptor generation
- Similarity searching
  - Substructure & similarity
  - Property searches
  - Model-based searches
    - Pharmacophore & QSAR
- 3D conformer calculation
- Library design & analysis
  - Library comparison
  - Structure & descriptor clustering
  - Diversity analysis
- Data management: databases

# Drug Informatics

- Screening database
- Bioactivity data
- Lead prioritization
  - QSAR analyses
  - Pharmacophore modeling
- Structural informatics
  - Docking
  - Rational drug design
  - Virtual screening

# Bioinformatics and Network Analysis

- Bioinformatics
  - Sequences & mRNA/protein profiling
  - Protein-protein interaction data
  - Pathways & Ontologies
  - Mutant analysis
- Network Analysis
  - Clustering (HC, KM, PCA, MDS, NN, etc )
  - Systems modeling
- Statistics for everything

# Dilemma for Academic Institutions

**Current infrastructure**

- Very limited open-source resources
- Dominance of commercial software

**Decision: commercial vs. public**

- Black boxes for a lot of $
- Focus on public resources with internal development time

**Advantage of Public Resource Approach**

- Higher educational value
- Transparent, modifiable and shareable
- Higher public impact

# ChemMine: Chemical Genomics DB

Compound mining and screening database for drug and chemical genomics discovery

- Current functionality
  - Activity- & property-based searching
  - Structure-based searching
  - Online clustering
  - Upload and retrieval of screening data
- Long term goals
  - Central depository of internal and external screening data
  - Online service for mining of drug-like compounds
  - Ontologies for bioactives and screening data
- *In silico* discovery
  - Training set for predictive approaches
    - Property-focused libraries
  - Systems approaches
    - Predictive process modeling

Publication: *Plant Physiol* (2005) **138**, 573-577

# Functionality Overview

URL: http://bioweb.ucr.edu/ChemMine/

Over 2.5 million structures

- Screening libraries
- Bioactives & natural CMPs
- Metabolic compounds

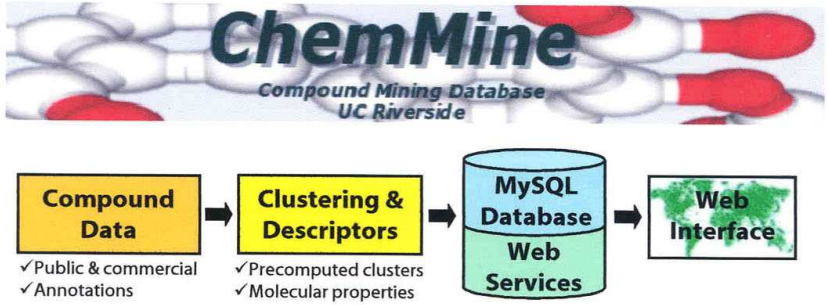
Activity information

- Screening data
- Target proteins
- Literature

Web services

- Clustering
- Chemical descriptors
- Structure formats

(A)

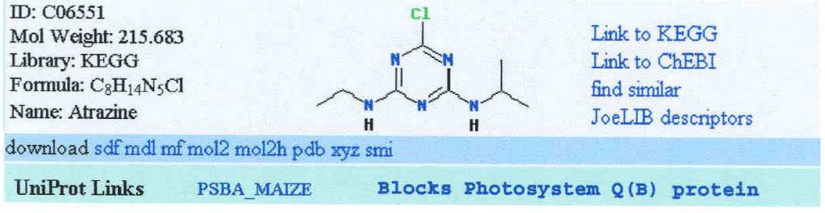



(B)

Descriptor Table (partial view)

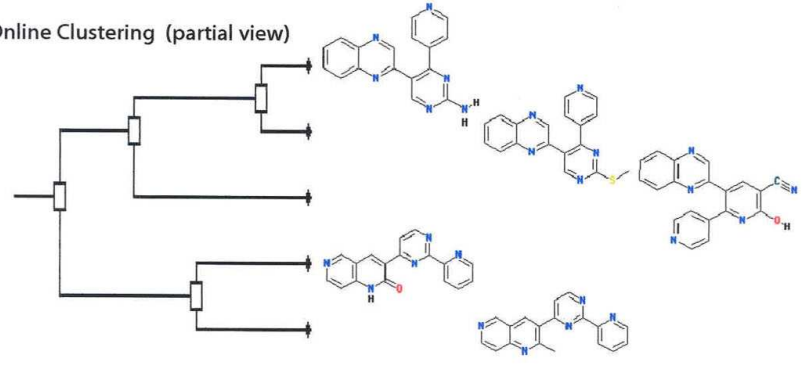
| | ID | Mol Weight | FOR | LGP | HA1 | HA2 | HD1 |
|---|---|---|---|---|---|---|---|
| ☑ | 1234 | 193.244 | $C_{14}H_{11}N$ | 5.1797 | 12 | 1 | 0 |
| ☑ | 12340 | 612.505 | $C_{32}H_{16}O_8N_6$ | 9.5864 | 22 | 2 | 0 |
| ☑ | 123400 | 409.432 | $C_{23}H_{23}O_6N$ | 8.2873 | 29 | 6 | 1 |
| ☑ | 123401 | 372.458 | $C_{21}H_{28}O_4N_2$ | 10.3585 | 33 | 5 | 2 |

Annotation Page (partial view)

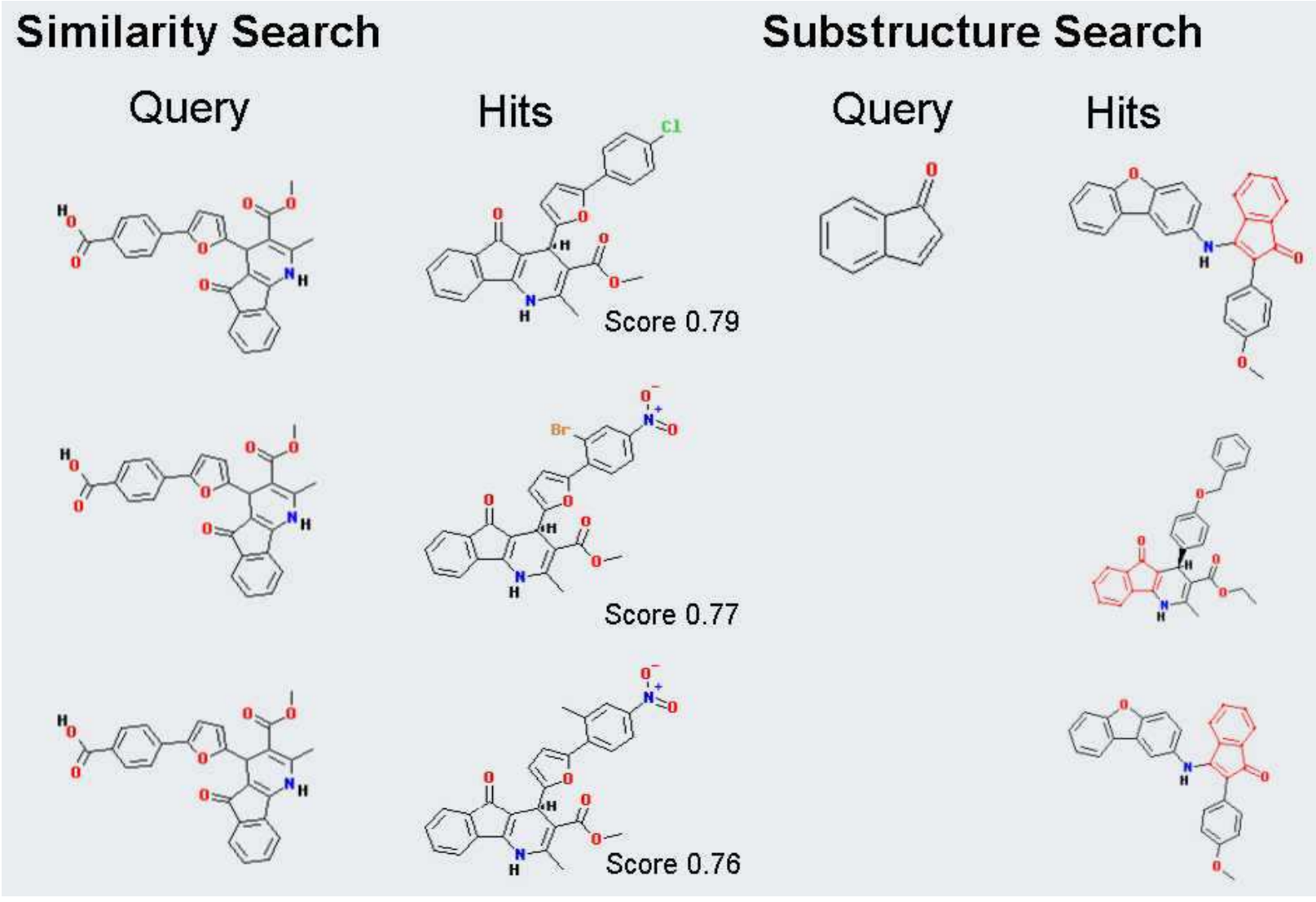



(C)

Online Clustering (partial view)



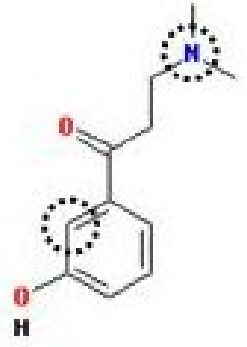
# Similarity Searching



# 2D Fragment Similarity Searching

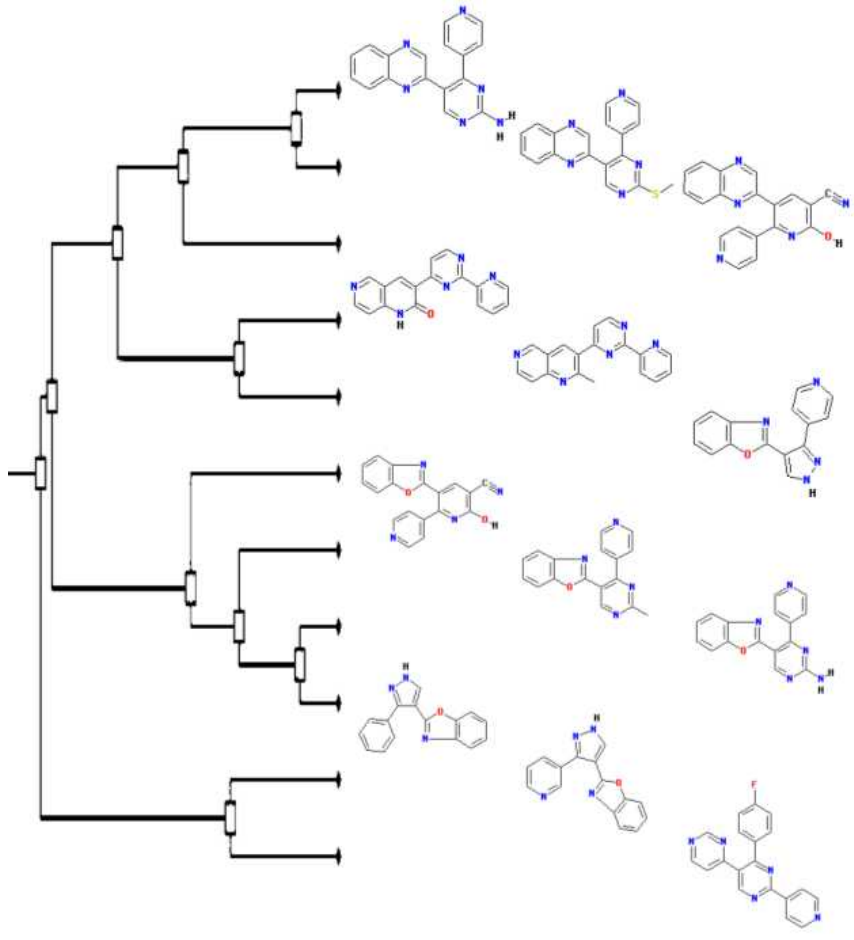
- Most commonly used for large database searching.
- 3D approaches used in pharmacophore searching.
- Advantages
  - Fast and accurate
- Involves 2 major steps
  - Structural descriptors
  - Similarity coefficients/measures
- Structural descriptors in similarity searching
  - Atom pairs: C12N03_06
  - Atom sequences: C12C13C13C02C02N03
  - Fingerprints: rules to enumerate all fragments in common structures
- Substructure searching
  - Similar to atom sequences

Sample Structure



# Online Clustering

- Similarity-based
- Descriptor-based
- Hierarchical or binning clustering
  - All-against-all similarity
  - Distance matrix
  - Clustering



# Preclustered Libraries

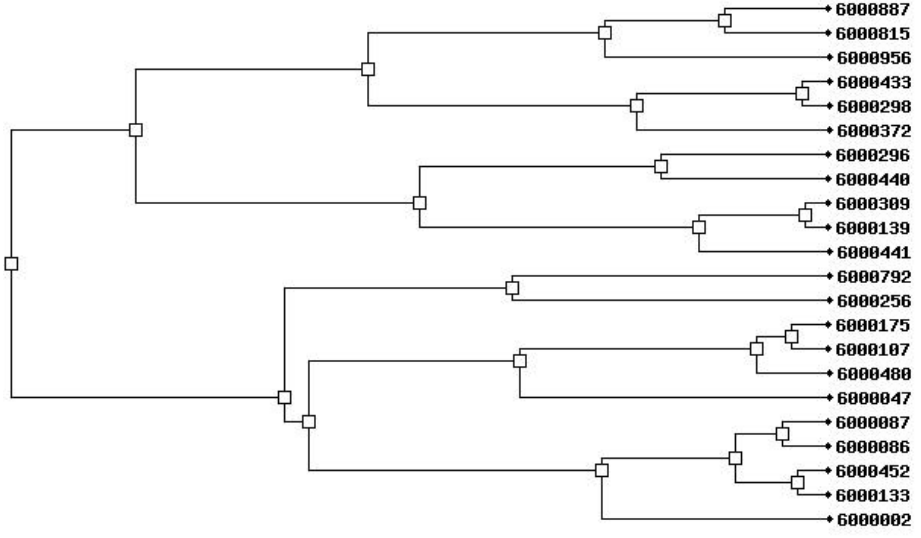
| PubChem: 711,361 compounds | | | | | | | |
|---|---|---|---|---|---|---|---|
| Similarity | Number of Clusters | Cluster Size >100 | >50 | >20 | >10 | >2 | Singlets |
| 90% | 110969 | 27 | 54 | 280 | 1199 | 109409 | 389901 |
| 80% | 116707 | 77 | 173 | 805 | 2579 | 113073 | 299187 |
| 70% | 88050 | 198 | 329 | 1520 | 3530 | 82473 | 169431 |
| 60% | 43730 | 95 | 148 | 667 | 1844 | 40976 | 80526 |

# Descriptor Clustering

| | ID | Mol Weight | FOR | LGP | HA1 | HA2 | HD1 | HD2 | ACG | AOH | BAG | FRB | ROT | AB | HCY | HPG | MOR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ☑ | 6000002 | 299.326 | $C_{19}H_{13}ON_3$ | 4.4841 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0.115 | 3 | 26 | 2 | 0 | 89.285 |
| ☑ | 6000047 | 297.435 | $C_{20}H_{27}ON$ | 5.0786 | 1 | 2 | 1 | 1 | 0 | 1 | 0 | 0.08 | 2 | 25 | 1 | 3 | 91.3358 |
| ☑ | 6000086 | 290.362 | $C_{18}H_{18}N_4$ | 3.7471 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0.08 | 2 | 25 | 3 | 1 | 85.6332 |
| ☑ | 6000087 | 267.329 | $C_{15}H_{17}N_5$ | 2.5788 | 4 | 5 | 0 | 0 | 0 | 0 | 0 | 0.043 | 1 | 23 | 3 | 2 | 75.6992 |
| ☑ | 6000107 | 265.353 | $C_{17}H_{19}N_3$ | 3.6399 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 0.086 | 2 | 23 | 2 | 1 | 80.5079 |
| ☑ | 6000133 | 292.335 | $C_{17}H_{16}ON_4$ | 2.5934 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0.08 | 2 | 25 | 3 | 0 | 82.6012 |
| ☑ | 6000139 | 366.339 | $C_{20}H_{13}N_4F_3$ | 5.4542 | 4 | 3 | 1 | 1 | 0 | 0 | 0 | 0.133 | 4 | 30 | 2 | 0 | 97.4957 |
| ☑ | 6000175 | 267.326 | $C_{16}H_{17}ON_3$ | 2.4862 | 3 | 2 | 1 | 1 | 0 | 0 | 1 | 0.086 | 2 | 23 | 2 | 0 | 77.4759 |
| ☑ | 6000256 | 332.786 | $C_{19}H_{13}N_4Cl$ | 5.0888 | 4 | 3 | 1 | 1 | 0 | 0 | 0 | 0.111 | 3 | 27 | 2 | 0 | 97.5037 |
| ☑ | 6000296 | 420.509 | $C_{26}H_{24}N_6$ | 4.192 | 4 | 5 | 1 | 1 | 0 | 0 | 1 | 0.108 | 4 | 37 | 3 | 0 | 122.344 |
| ☑ | 6000298 | 342.348 | $C_{21}H_{14}O_3N_2$ | 5.5734 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0.137 | 4 | 29 | 2 | 0 | 101.469 |
| ☑ | 6000309 | 334.322 | $C_{19}H_{12}N_4F_2$ | 4.7136 | 4 | 3 | 1 | 1 | 0 | 0 | 0 | 0.107 | 3 | 28 | 2 | 0 | 92.4097 |
| ☑ | 6000372 | 292.289 | $C_{17}H_{12}O_3N_2$ | 4.4202 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0.166 | 4 | 24 | 2 | 0 | 83.9634 |
| ☑ | 6000433 | 342.348 | $C_{21}H_{14}O_3N_2$ | 5.5734 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0.137 | 4 | 29 | 2 | 0 | 101.469 |
| ☑ | 6000440 | 338.362 | $C_{21}H_{14}ON_4$ | 4.7368 | 4 | 5 | 0 | 0 | 0 | 0 | 0 | 0.1 | 3 | 30 | 2 | 0 | 100.24 |
| ☑ | 6000441 | 316.332 | $C_{19}H_{13}N_4F$ | 4.5745 | 4 | 3 | 1 | 1 | 0 | 0 | 0 | 0.111 | 3 | 27 | 2 | 0 | 92.4517 |
| ☑ | 6000452 | 255.275 | $C_{13}H_{13}ON_5$ | 1.1791 | 5 | 6 | 0 | 0 | 0 | 0 | 0 | 0.045 | 1 | 22 | 3 | 0 | 68.1202 |
| ☑ | 6000480 | 258.402 | $C_{17}H_{26}N_2$ | 3.8445 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 0.047 | 1 | 21 | 2 | 0 | 78.2375 |
| ☑ | 6000792 | 442.592 | $C_{15}H_8O_4N_2ClI$ | 3.5978 | 5 | 6 | 1 | 1 | 0 | 0 | 0 | 0.08 | 2 | 25 | 2 | 0 | 86.1251 |
| ☑ | 6000815 | 455.462 | $C_{26}H_{21}O_5N_3$ | 5.9762 | 5 | 5 | 2 | 2 | 0 | 0 | 0 | 0.216 | 8 | 37 | 1 | 0 | 129.063 |
| ☑ | 6000887 | 350.325 | $C_{19}H_{14}O_5N_2$ | 4.3097 | 4 | 4 | 1 | 1 | 0 | 0 | 0 | 0.214 | 6 | 28 | 1 | 0 | 95.1981 |
| ☑ | 6000956 | 420.523 | $C_{21}H_{28}O_5N_2S$ | 4.203 | 7 | 7 | 1 | 1 | 0 | 0 | 0 | 0.3 | 9 | 30 | 0 | 0 | 113.197 |

Pages: 1

cluater all



# Future Updates

- Management of screening data
  - Upload of bioactivity information from external and internal screens
  - Diverse data types
    - Compound upload
    - Quantitative data forms
    - Image data
    - User-specific data?
  - Development of ontology for bioactives, screening data and phenotypes
  - Routines for publishing user data including mandatory curation system
  - Interoperability with other screeing projects: ChemBank, PubChem, etc.

- Chem/drug informatics utilities
  - QSAR analysis tools
  - Expansion of drug-likeness and descriptor predictions
  - Drug informatics R/BioConductor libraries
  - Size-insensitive similarity searching
  - Combinatorial searches, e.g. similarity plus descriptors
  - IUPAC/InChI names
  - 3D conformer generation

# Library Comparisons

Assemble collection of 50-100K diverse drug-like compounds

**Selection Criteria**

- Screening compounds, bioactives and natural products
- Minimum overlap within and between libraries
- Vendor diversity (J Chem Inf Comput Sci 44: 643-651)
- Majority drug-like: 'plant Lipinski rules' (Pest Manag Sci 58: 219-233)
- Elimination of undesirable side groups (filters)
- Recommendations (ICCB, etc.)
- Literature (e.g. Tudor Oprea: Curr Drug Disc Technol 1: 211-220)
- Resupply situation
- Price

# Analyzed Libraries

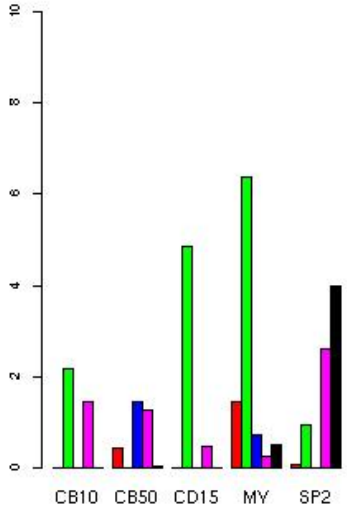
- Quantity conversions
  - First library: 0.1mg [20$\mu$l], 5$\mu$g/$\mu$l [10-20mM]
  - Final assay conc: 5-10ng/$\mu$l [10-40$\mu$M]
  - Number of screens: 100 screens with 100$\mu$l end volume
- Analyzed collections
  - ChemBridge: Microformat, 10,000 CMPs (0.1mg)
  - ChemBridge: DIVERSet, 50,000 CMPs (0.25mg)
  - Chemical Diversity: ICCB set, 15,000 CMPs (0.5mg)
  - Sigma/TimTec: MyriaScreen, 10,000 CMPs (0.25mg)
  - Microsource: Spectrum, 2,000 CMPs (0.25mg)
- On waiting list
  - ChemBridge: NOVACore (parallel DOS), 40,000 CMPs, 0.25mg
  - Biomol/TimTec: MaxiVerse, 9,600 CMPs, 0.25mg
  - World Molecules/MDD Inc: x CMPs
  - Enamine: 10,000-20,000 CMPs, 0.25mg
  - Analyticon Discovery: MEGAbolite & NatDiverse, 3000 CMPs, 0.25mg

# Clustering Methods

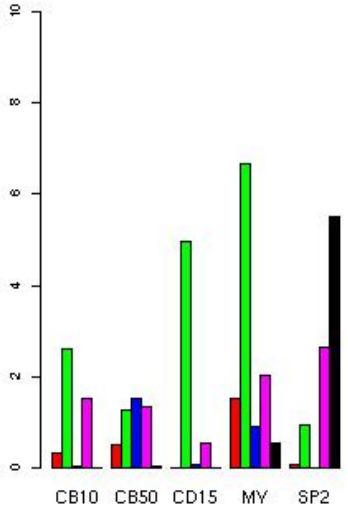
- Principal component analysis
  - Reduction technique of multivariate data to principal compoments to identify hidden variances
- Multidimensional scaling
  - Displays distance matrix of objects in spacial plot
- Single linkage binning
  - Uses provided similarity cutoff for grouping of items

# Similarity Clustering

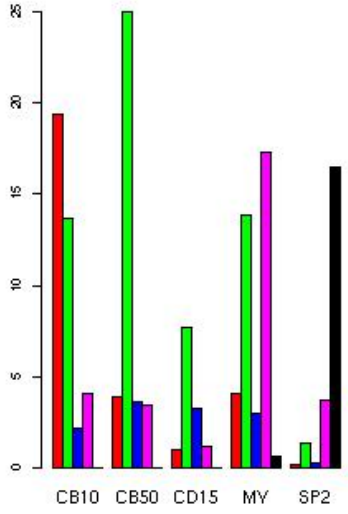
100% Ident APs



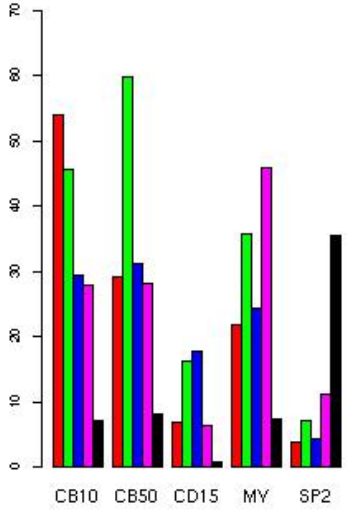

90% Ident APs



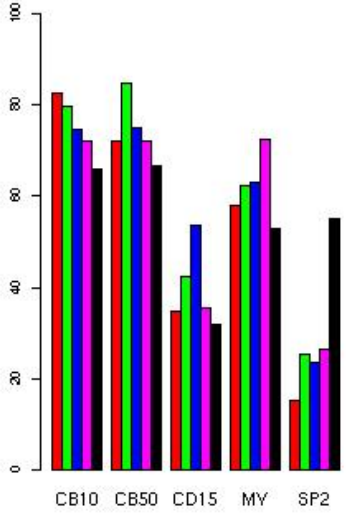

80% Ident APs




70% Ident APs



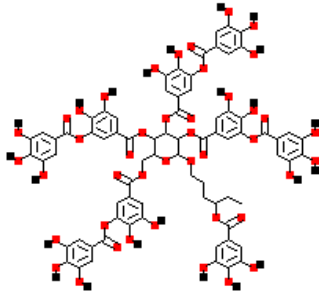

60% Ident APs



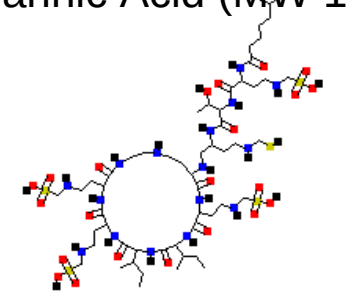

CB10 CB50 CD15 MY10 SP2

# Property PCA

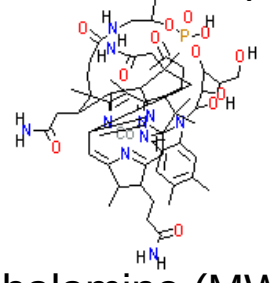
Diversity relative to 80 complex bioactives in MS Spec



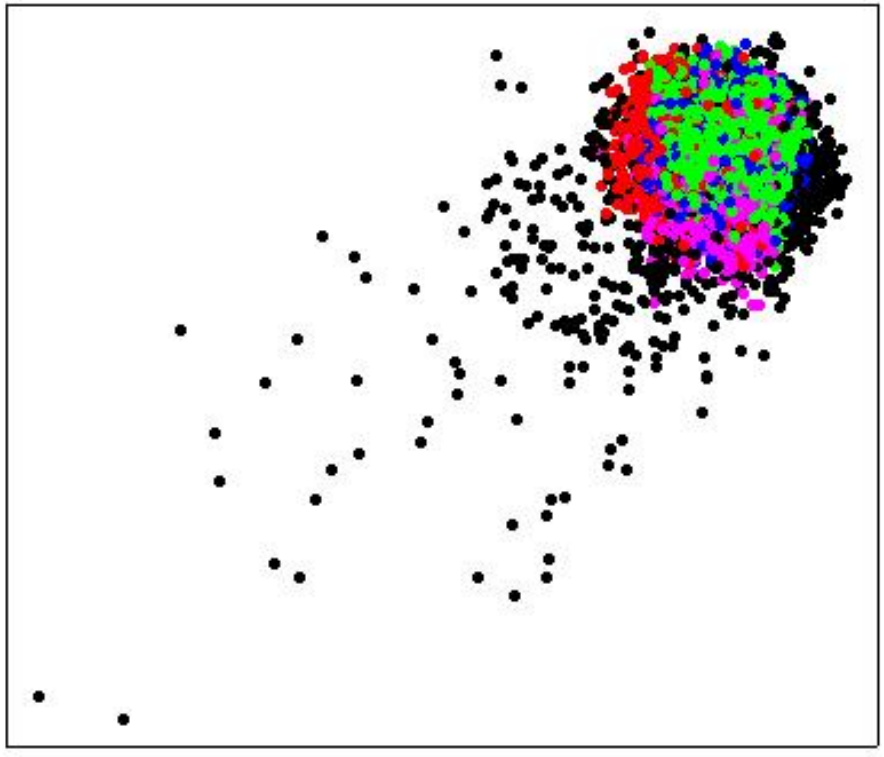
Tannic Acid (MW 1600)



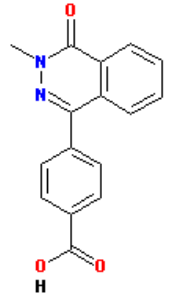
Colistimethate (MW 1400)



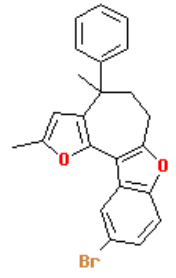
Cobalamine (MW 1100)



CB10 CB50 CD15 MY10 SP2



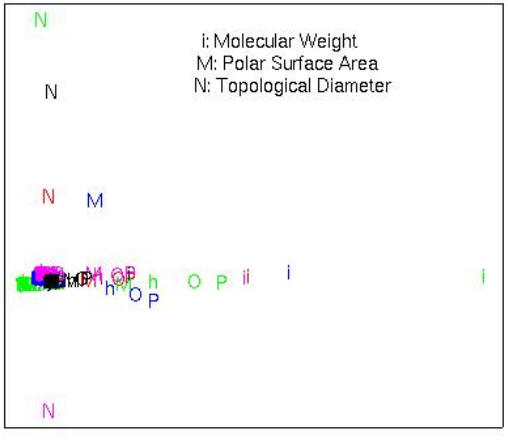
56421 (MW 270)



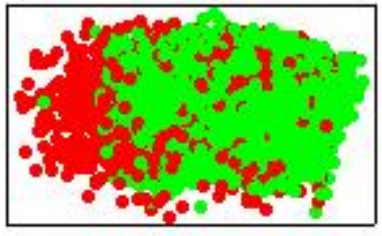
19737 (MW 390)

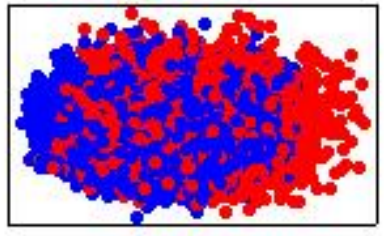
# Property Differences

Property Plot

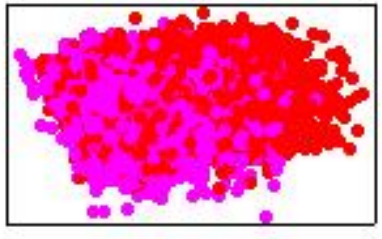


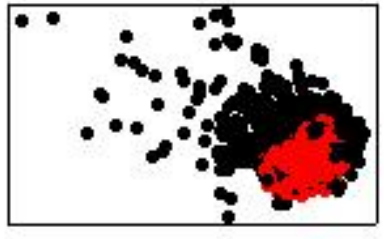

Compound Plots





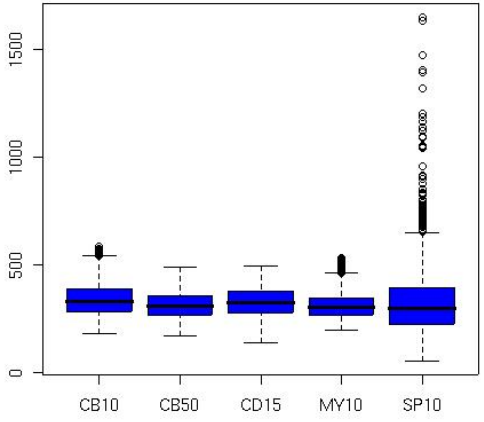
CB10 CB50 CD15 MY10 SP2



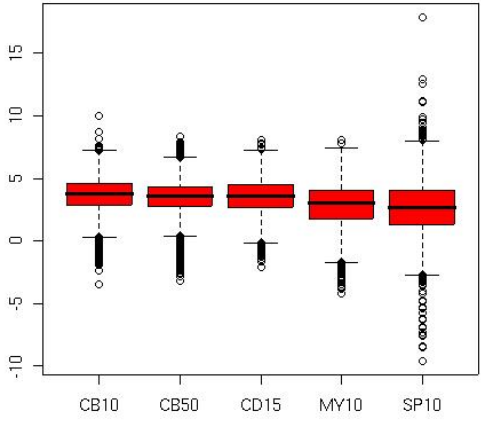


# Lipinski Descriptors

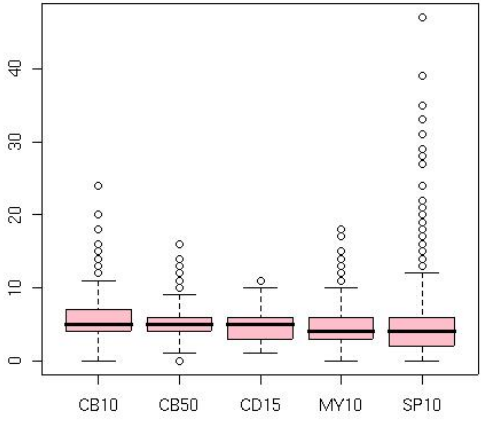
Molecular Weight



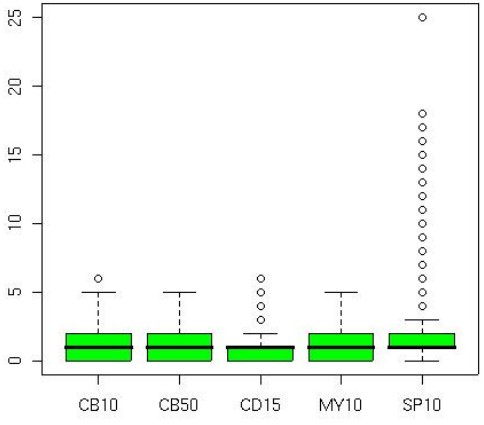

LogP



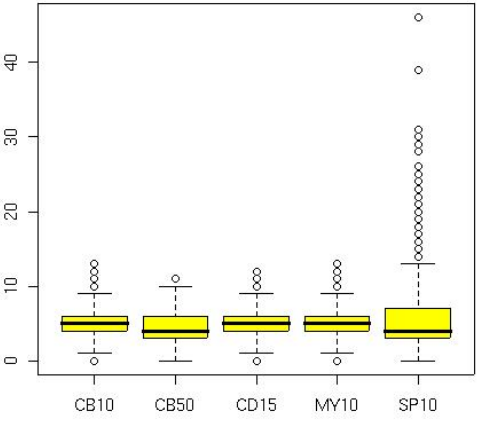

Rotatable Bonds




Hydrogen Bond Donors

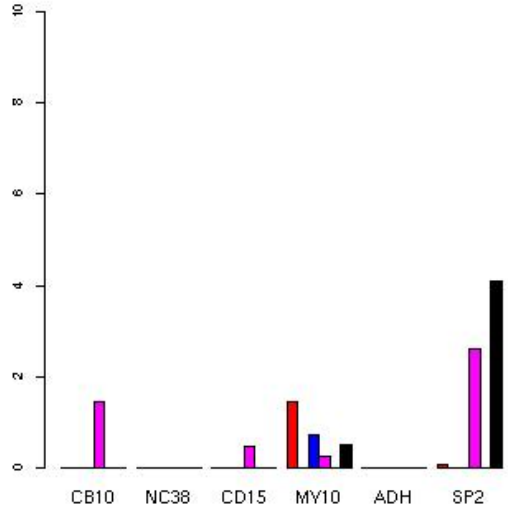



Hydrogen Bond Acceptors

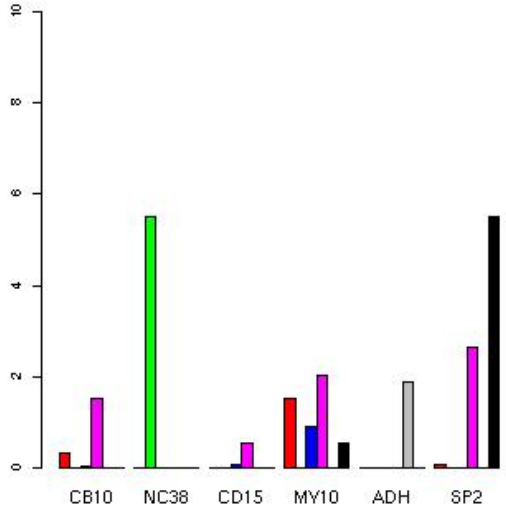


# Similarity Clustering #2

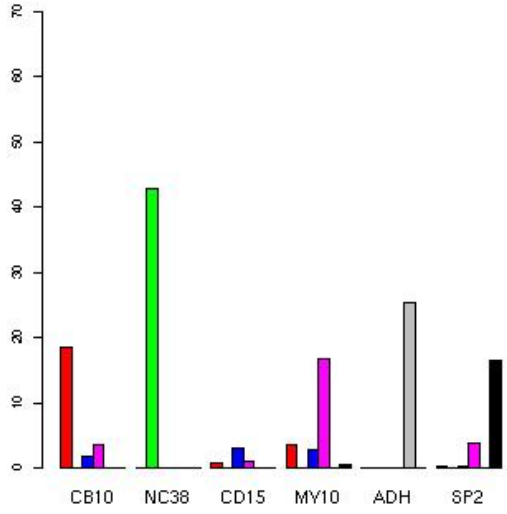
100% Ident APs



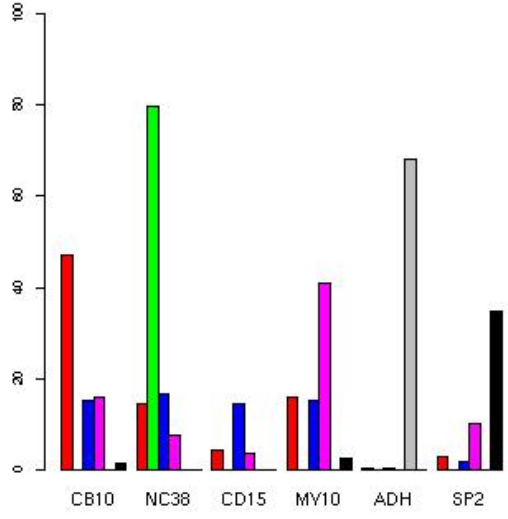

90% Ident APs



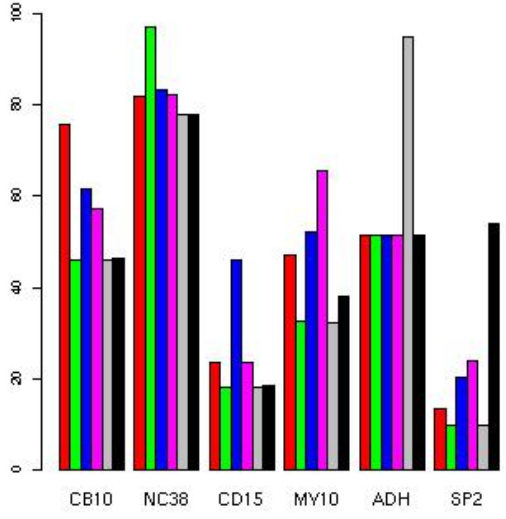

80% Ident APs




70% Ident APs




60% Ident APs

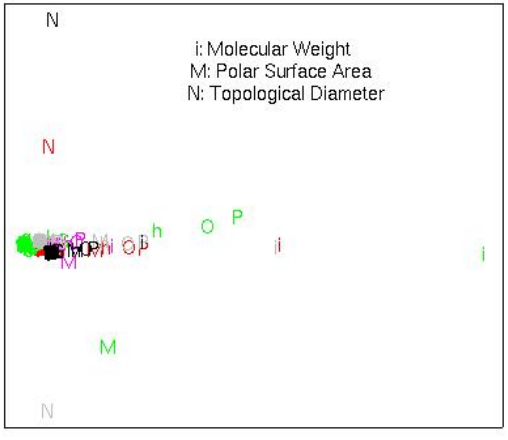



CB10 NC38 CD15

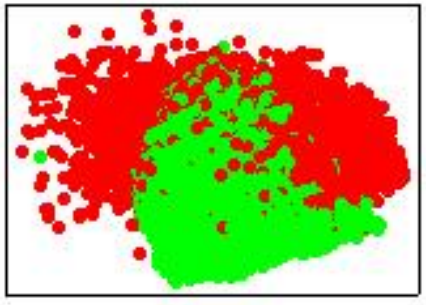
MY10 ADH1 SP2

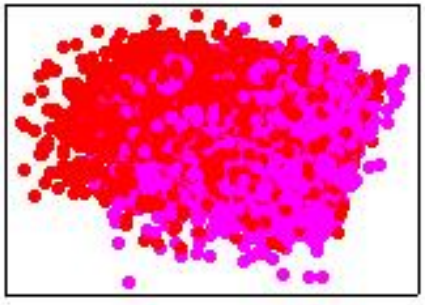
# Property Differences #2

Property Plot

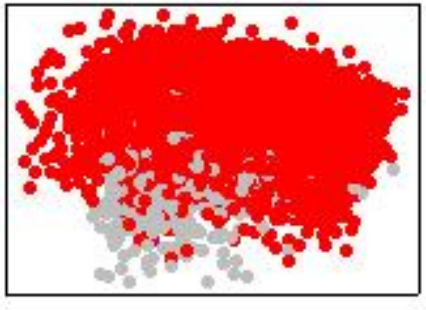


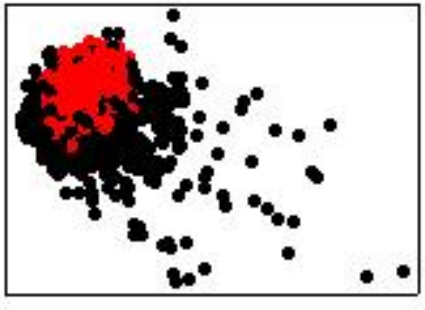

Compound Plots





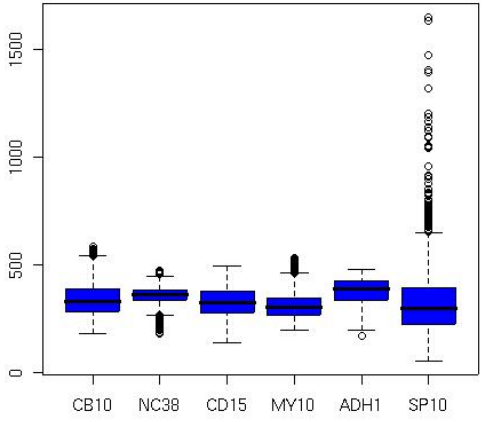
CB10 NC38 MY10 ADH1 SP2



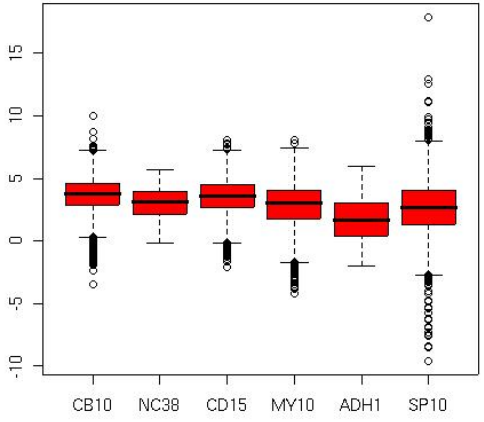


# Lipinski Descriptors #2

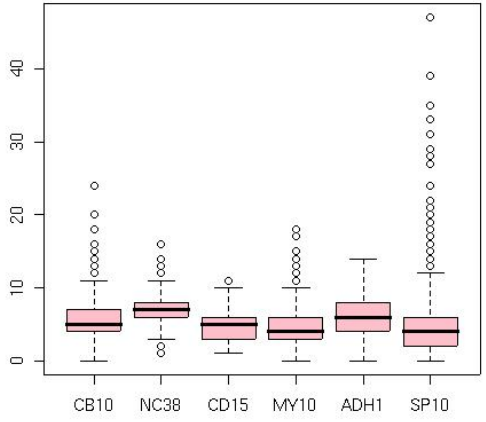
Molecular Weight



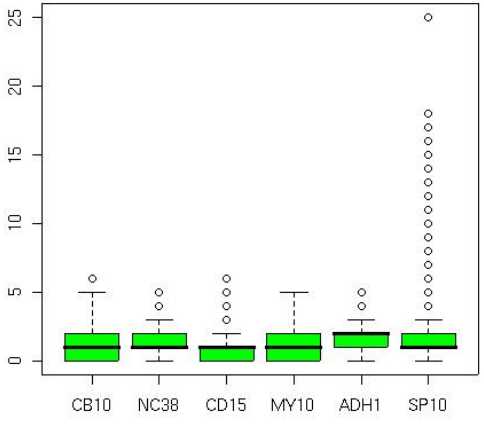

LogP



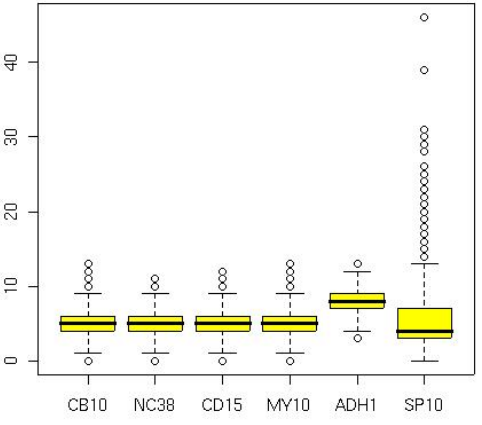

Rotatable Bonds




Hydrogen Bond Donors




Hydrogen Bond Acceptors



# Conclusions

- Low overlap in high similarity ranges (>80%)
- Strong overlap in low similarity ranges (<70%)
- Small differences in property spectrum
- Property spectrum clusters around known drugs
- Base decision on:
  - Price
  - Resupply
  - Vendor diversity

# Lab Members

**Center for Plant Cell Biology, UCR**

- Julian Krause, Undergraduate S. (CS)
- Josh Lauricha, Systems Admin (CS)
- Kevin Horan, Programmer (Math/CS)
- Li-Chang Cheng, Undergraduate (CS)
- Charles Jang, Graduate S. (CG)
- Colleen Knoth, Graduate S. (CG)
- Jack Cui, Temp (BCH/CS)
- Thomas Girke, AC & PR (BCH/MolB/BI)

# Questions

?