Future Directions in KD/KM

Terry Brunck

Daylight CIS

Daylight

Chemical
Information
Systems, Inc.

Enterprise-level cheminformatics.

Copyright 2005, Daylight CIS, Inc.

Knowledge Discovery vs Knowledge Management

- Discovery
  - “How one understands and uses one’s data” (ACM)
  - “Each problem I solved became a rule, which served to solve other problems” (Descartes)
  - Issues of representation and abstraction
- Management
  - Information capture, integration, distribution, and application
  - Databases, ontologies, taxonomies

Knowledge Is Hierarchical

- One persons result is another person’s data

  Raw data -> result = data -> result

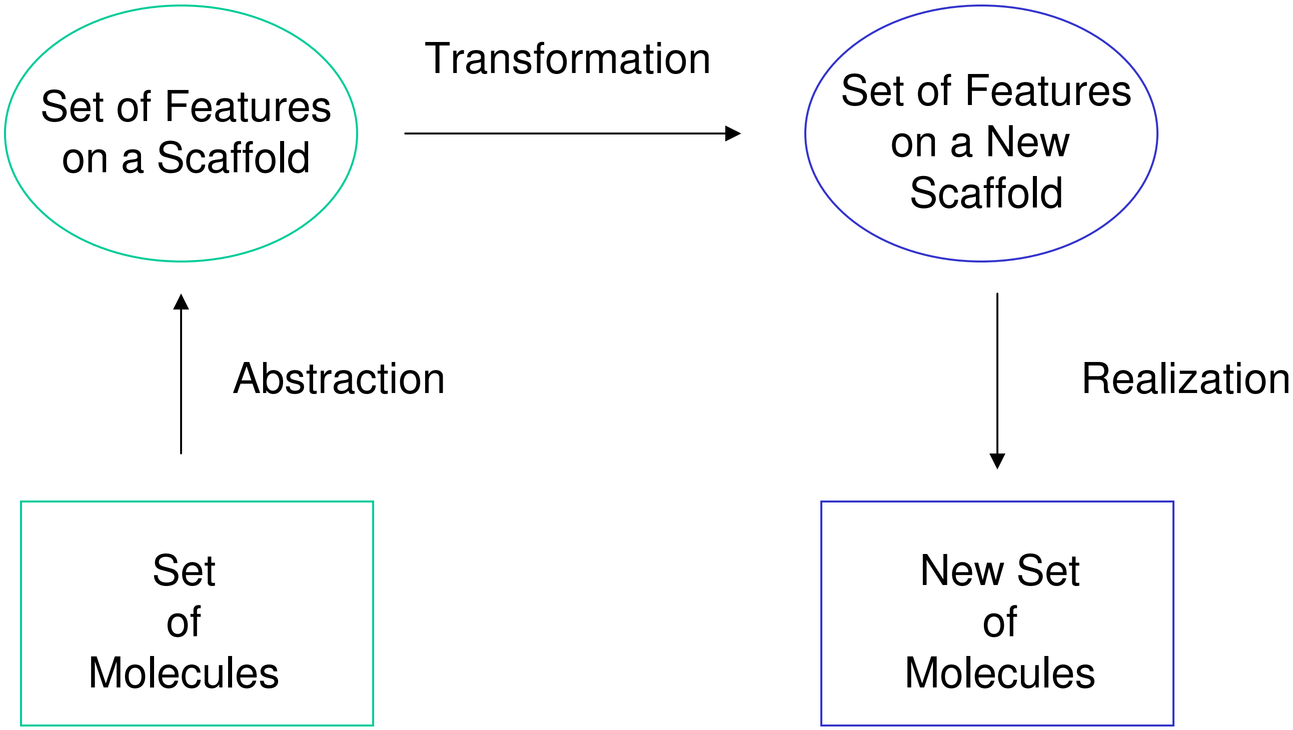
- Results are more abstract than the data from which they are derived

  PMT current versus time -> reaction rate -> Ki -> QSAR

Chemistry Knowledge Discovery

- Units of reasoning and transformations
  - individual molecules
  - sets of molecules
  - reactions
  - general chemical transformations
- Use of experts to define units of reasoning and to exemplify reasoning processes

Reasoning



Set of Features on a Scaffold

Transformation

Set of Features on a New Scaffold

Abstraction

Realization

Set of Molecules

New Set of Molecules

Daylight's Historical Contribution to KD/KM

- SMILES
  - Represents individual molecules (or very small sets of stereochemically related molecules) and reactions
  - Unique identifier for database construction
  - Relationships - Synonyms, tautomers, isomers, parents, children, siblings, precursors, products
- SMARTS
  - represents abstract set of molecules - a slice of chemistry space
  - describes relationships - membership, substructure/superstructure
  - is used as a query to define sets over some collection of molecules
  - is an abstraction of a particular set of molecules

Daylight’s Historical Contribution to KD/KM

- SMIRKS
  - Transforms between sets of molecules
  - Describes relationships - reaction membership
- The languages, toolkits, and DayCart functions permit the construction of databases, including basic chemical taxonomies and ontologies

Additional Needs for KD/KM

- General representation of sets of molecules
  - Driven by use of sets by experts
  - Enumeration with SMILES
  - R-table enumeration
- General abstraction of sets of molecules
  - SMARTS -> set of molecules
  - SMIRKS -> set of reactions
  - But there are no tools for
    - set of molecules -> SMARTS
    - set of reactions -> SMIRKS

Example

- Vendor has collection of focused libraries for ~15 varied targets
- Each library has 200-1500 molecules
- What knowledge is embedded in these sets?
  - Does my corporate DB contain potential members of one of these sets?
  - Do they know something I don’t?

Queries as Abstractions

- Create an abstraction of a set by generating suitable queries
  - bounds of computable properties
  - substructures
- Use the queries to answer knowledge-based questions.
- How to search queries?

Approaches

- Commercial
  - Tripos SARNavigator, Charisma/Distill
    - Clustering, MCS
  - ChemAxon - MC(E)S, R-Tables
    - Similar to Tripos
  - Bioreason - ClassPharmer
    - MCS-based clustering, R-Tables, SAR
- Literature/In-house
  - Similarity - fingerprints or MCES from full or reduced graphs with appropriate measures
  - Clustering - JP, Ward's, CAST + many others most requiring adjustable parameters
  - Anaylsis - R-Tables, QSAR
- Mainly focused on hit-list analysis
  - How do you archive, search, or compare such knowledge?

Substructure Query Generation

1. Represent molecule as a set of shortest-linear-paths between atoms
   - analogous to adjacency matrix or distance matrix in DG
   - intuitively over-determined representation
   - include arbitrary atom/bond properties in the path definition (atom: charge, aromaticity, ring count, global topology etc.)
   - Subsequent manipulations become string operations

Substructure Query Generation

2. Extract the common set of shortest paths over a set
   - analogous to recursive MCS and adaptive, unbiased keys
   - potentially discontinuous common substructure
     - benzoic, phenyl acetic, phenyl propionic acids give benzene and carboxy as common paths

- Potential for querying
  - Substructures linked by boolean AND
- But sets may not be very homogeneous

Prototype Query Generation

- Requires some form of clustering
  - Fingerprint and cluster the set
    - 2048 bits
    - JP clustering 8/14
- Generate common set of shortest paths for each cluster
  - sort molecules by size
  - generate common paths iteratively based on paths of smaller molecules

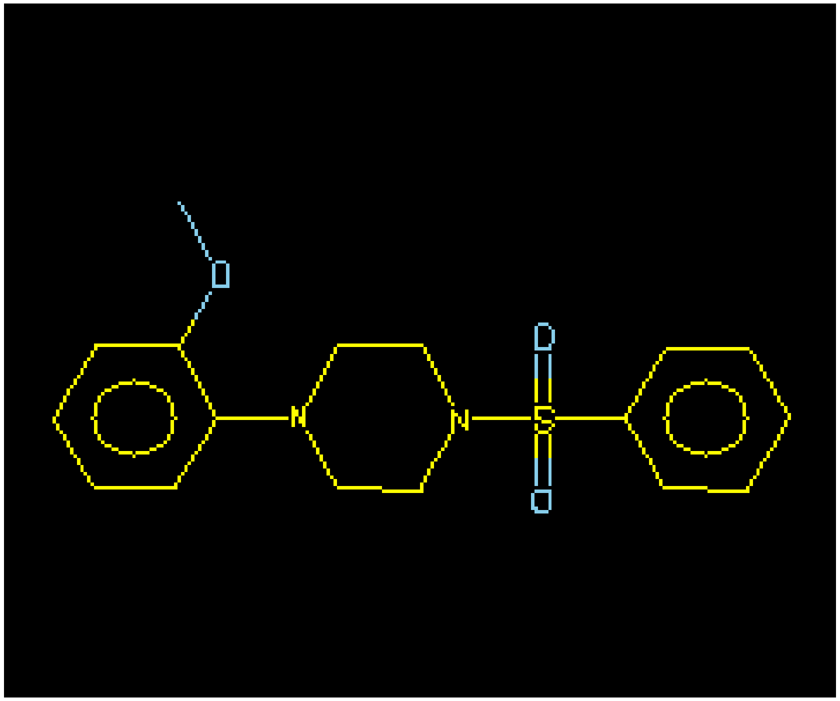
2-3-oxidosqualene-lanosterol cyclase inhibitors

- 1277 members in vendor library
- JP clustering
  - 80 clusters cover 1151 (90%) molecules
  - cluster size 117-> 2 members (53 clusters > 2 members)
  - 126 singletons
- smartsAbstraction - common shortest paths for each cluster

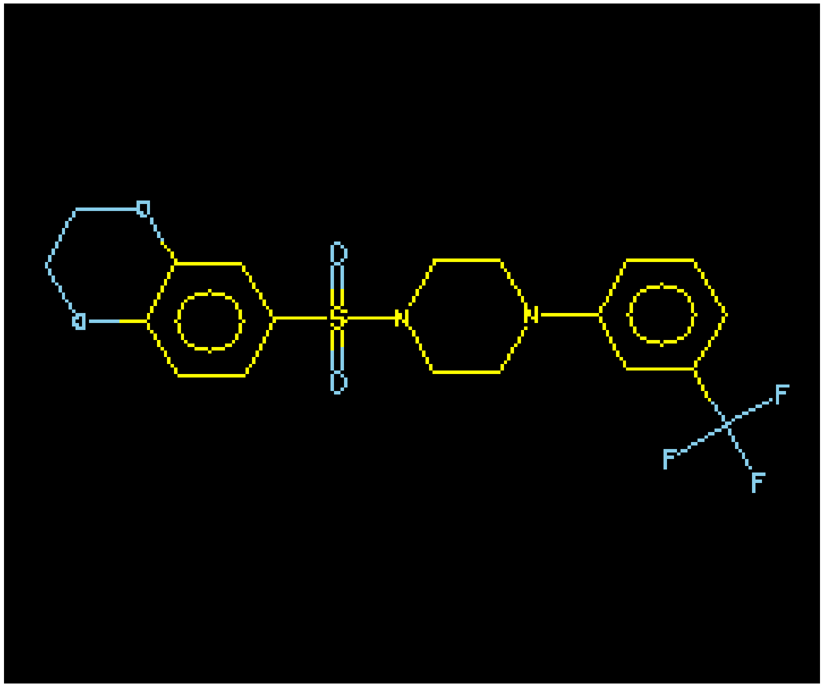
Cluster 0 (117 members)

- “Centroid” (0.0491variance)
  - COc1ccccc1N2CCN(CC2)S(=O)(=O)c3ccccc3
- Outlier (0.1338 variance)
  - FC(F)(F)c1cccc(c1)N2CCN(CC2)S(=O)(=O)c3ccc4OCCOc4c3
- Longest shortest-path length 13
  - c:c:c:c-N-C-C-N-S-c:c:c:c
- 88 common paths (1-13 atoms)
  - overlaps not removed yet

Cluster 0 (117 members)



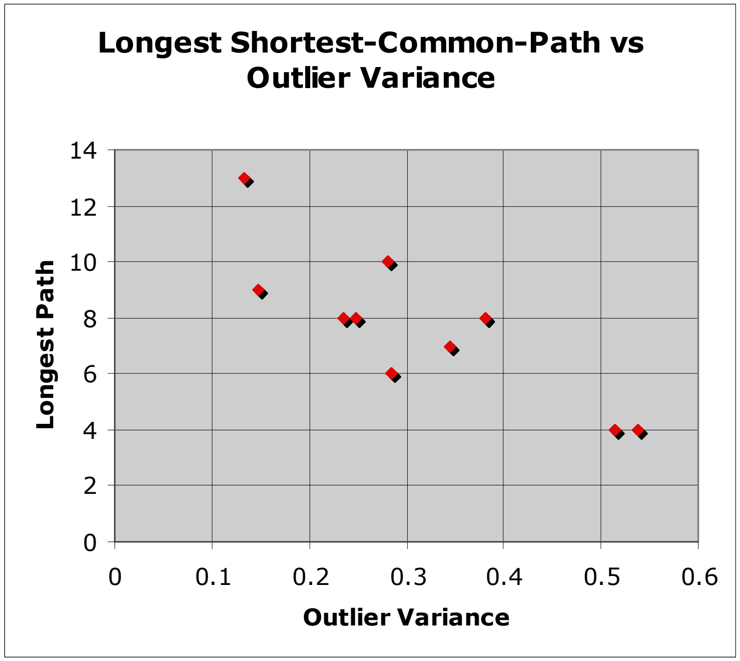
Centroid

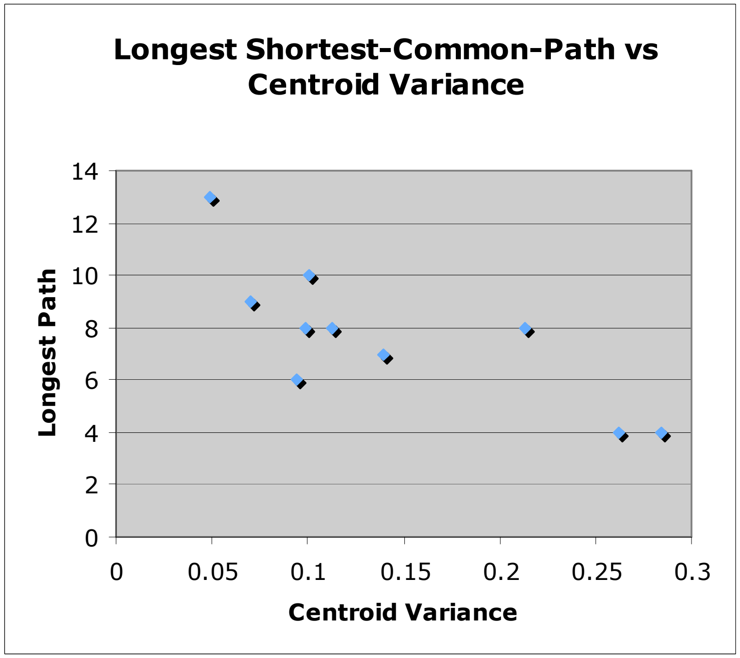


Outlier

Longest shortest-path(s) highlighted

Largest 10 Clusters

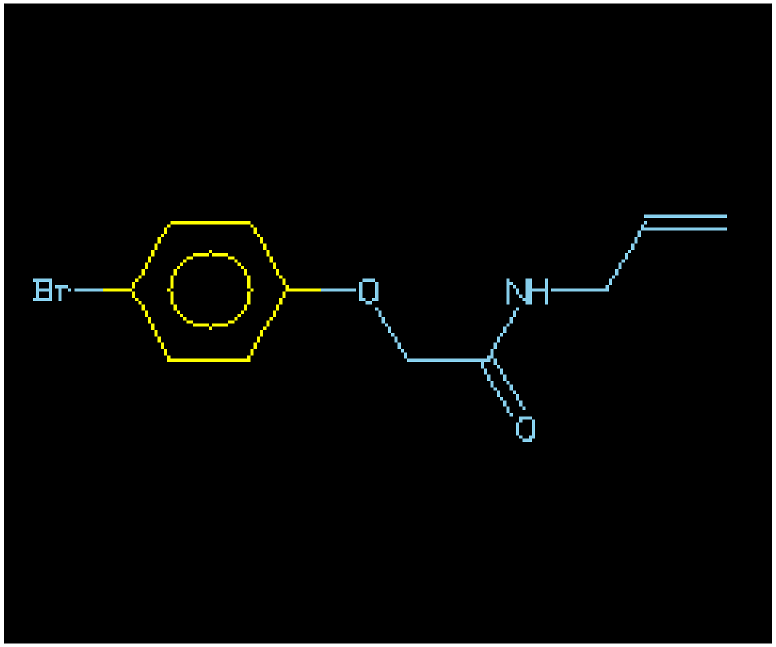


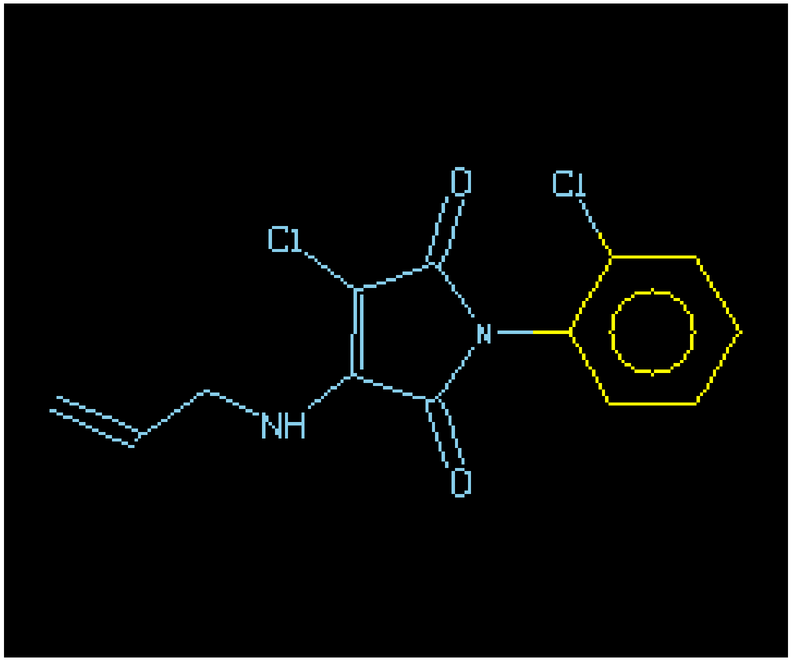


Worst Case Cluster 8

- Centroid (0.2846 variance)
  - Brc1ccc(OCC(=O)NCC=C)cc1
- Outlier (0.5384 variance)
  - ClC1=C(NCC=C)C(=O)N(C1=O)c2ccccc2Cl
- Longest shortest-path length 4
  - c:c:c:c
- 8 common shortest paths (length 1-4)

Worst Case Cluster 8



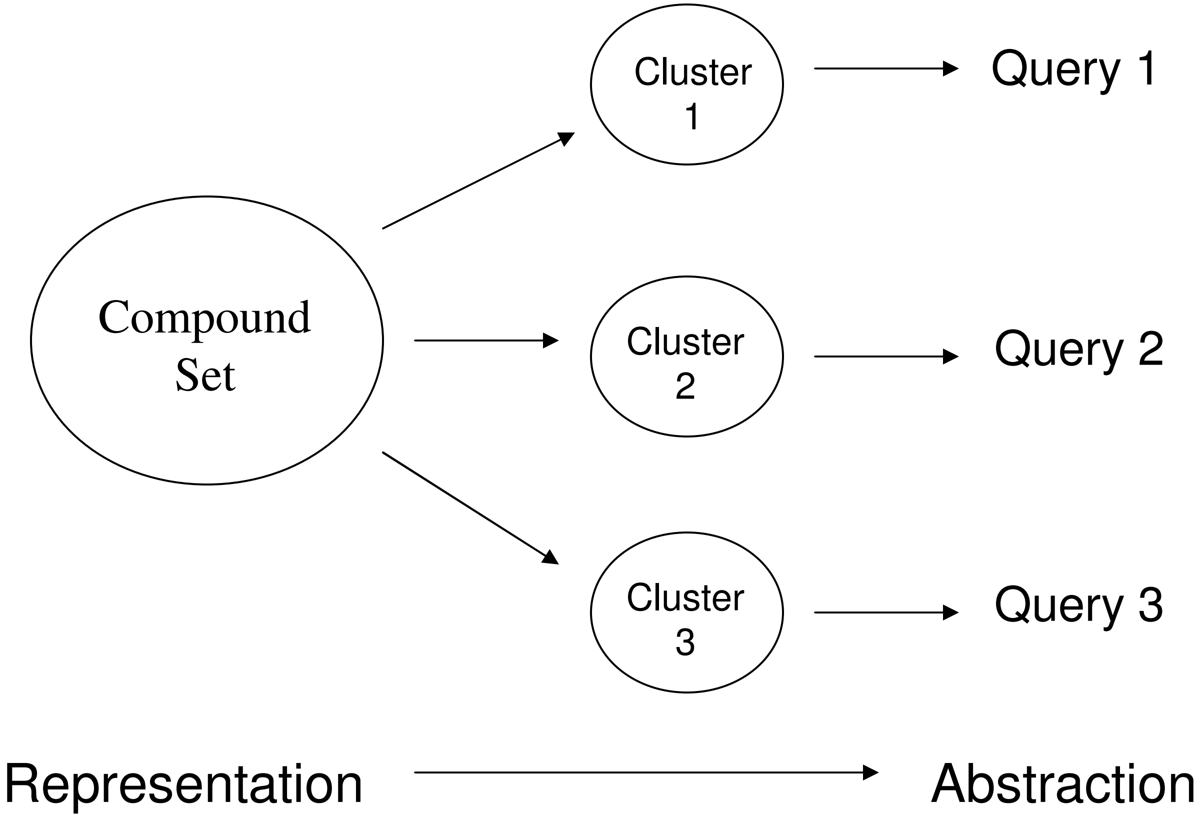


Centroid

Outlier

Longest shortest-path highlighted

Summary



Compound Set

Cluster 1

Cluster 2

Cluster 3

Query 1

Query 2

Query 3

Representation

Abstraction

Summary

- Sets of compounds are natural units of reasoning
  - Biology creates sets - e.g. toxicity, metabolism, activity
  - Chemists think in terms of sets (e.g. preferred scaffolds)
- A prototype method generates substructure queries from clustered sets of molecules as an abstraction of the set
  - substructure paths offer the potential to compare abstractions
  - better clustering will lead to better substructures
  - linear queries will facilitate querying of knowledge
- Query generation is knowledge discovery
- Queries must be searchable for knowledge management

Futures

- Queries formed by ANDing substructures and computable properties might be used to abstract molecule sets more completely
- Databases of such queries and their relationships may be useful for knowledge management
  - Ring system ontologies
  - QSAR database enhancement
  - Rule extraction from biological databases
- Vision: To build a layer of knowledge represented as queries on top of general chemical databases.