

Application of Graph Theory in Drug Design

Reihaneh Safavi-Sohi, Jahan B Ghasemi
Drug Design in silico Lab
Chem Faculty, K. N. Toosi Univ of Tech
Tehran, Iran

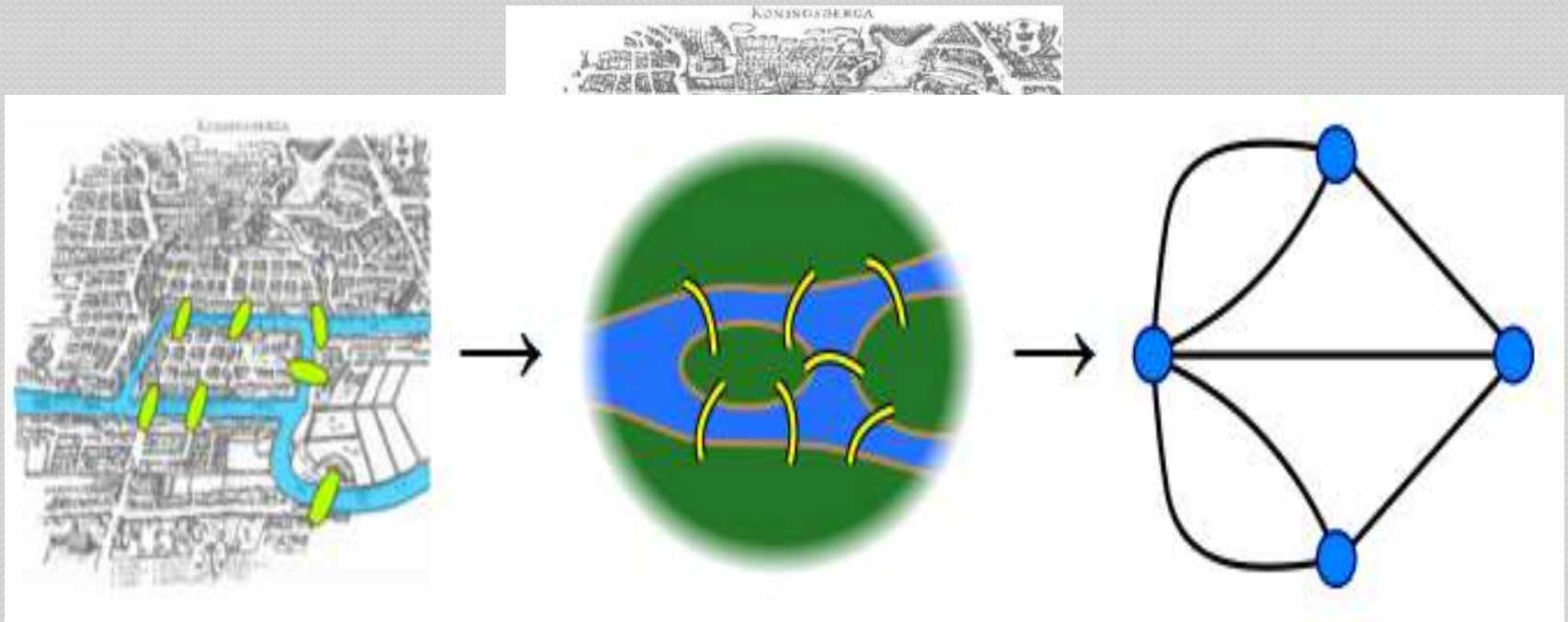
History

Graph theory is a branch of mathematics which studies the structure of graphs and networks.

Graph theory started in 1736, when Euler solved the problem known as the Königsberg bridges problem.

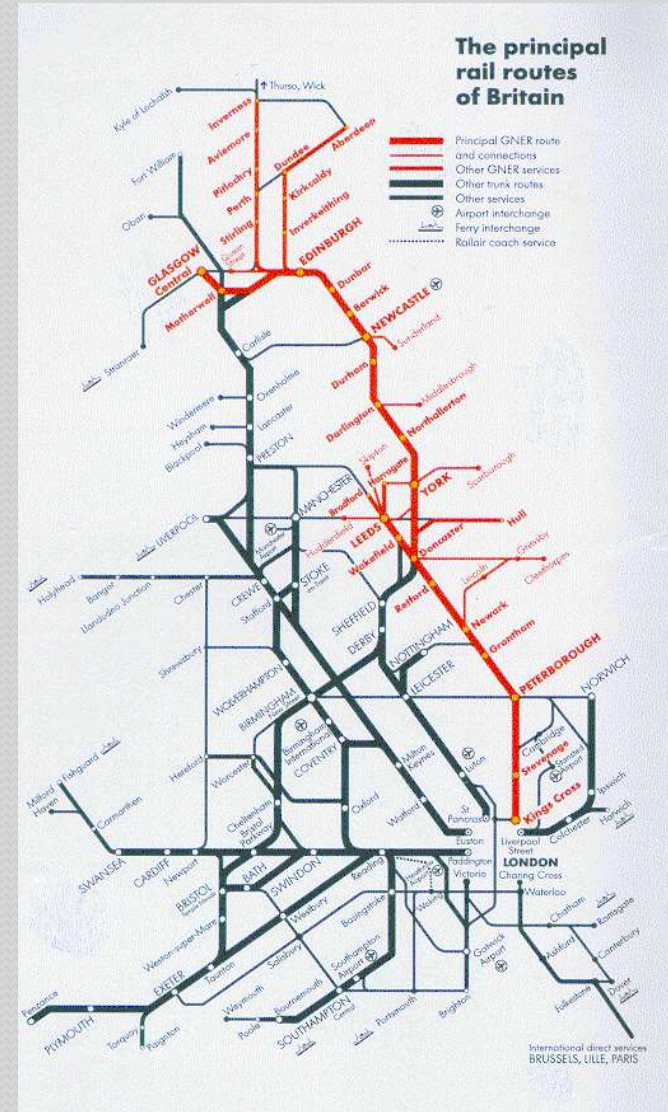
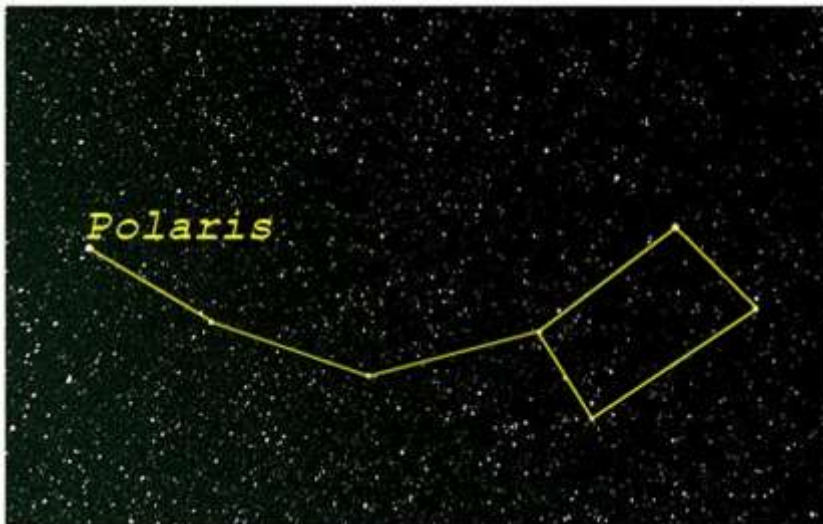
- Graph considers sets of objects, called nodes, and the relationships, called edges, between pairs of these objects.
- A collection of dots that may or may not be connected to each other by lines. It doesn't matter how big the dots are, how long the lines are, or whether the lines are straight, curved.
- **Neighbor:** The neighbors of a vertex are all the vertices which are connected to that vertex by a single edge.
- **The degree of a vertex** in a graph is the number of edges that touch it.
- The definition is completely general, allowing graphs to be used in many different application domains as long as an appropriate representation can be derived.

Seven Bridges of Königsberg



The problem was to find a walk through the city that would cross each bridge once and only once. The islands could not be reached by any route other than the bridges, and every bridge must have been crossed completely every time.

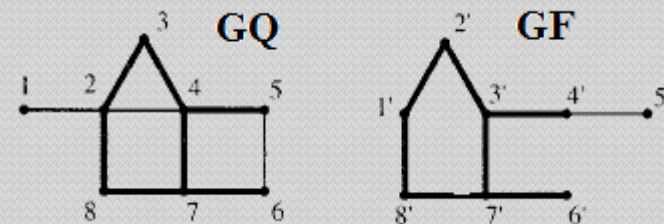
Another Examples Of Graphs



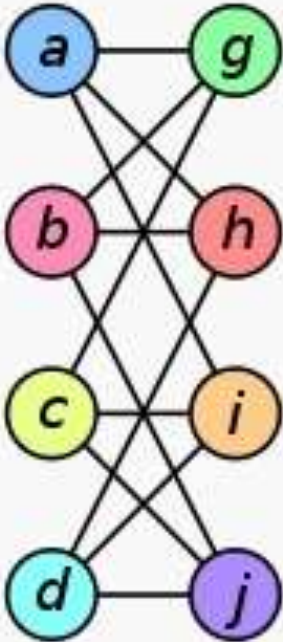
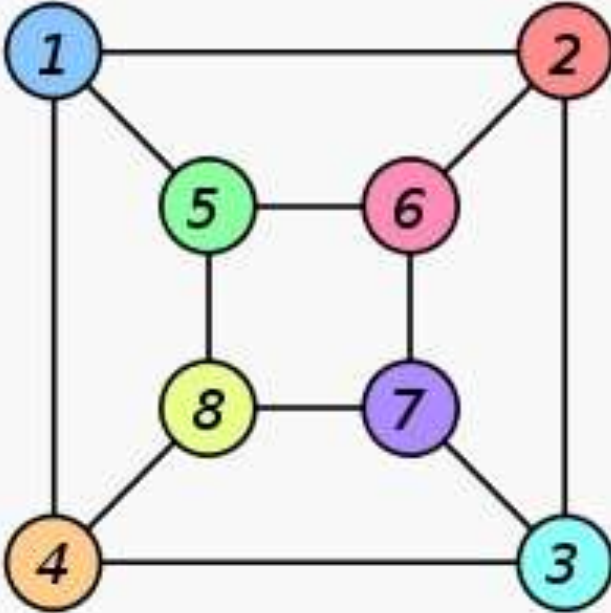
Isomorphism Algorithm

- Isomorphism procedures enable the comparison of pairs of graphs .
- To detect whether two graphs, or parts of there, are identical.
- Maximal common subgraph (MCS) isomorphism enables the identification of the largest subgraph common to a pair of graphs.

A subgraph isomorphism exists if all the nodes (atoms) of one graph (G_Q) can be mapped to a subset of the nodes of the other graph (G_F) in such a way that the edges (bonds) of G_Q simultaneously map to a subset of the edges in G_F .



- The labels carried by the nodes and edges (atom type and bond type, respectively) must be identical if the nodes or edges are to be mapped to each other.
- Testing for subgraph isomorphism is an **NP-complete problem**.
- Trying every possible way of mapping each of the n_Q nodes in G_Q onto one of the n_F nodes in G_F ($n_Q < n_F$);
Each of the $n_Q!/(n_F - n_Q)!n_F!$ possible mappings must then be tested to see if any of them obeys the adjacency condition.
- Even for very small graphs, the number of possible mappings rapidly becomes unmanageable.
- The use of this approach has been limited due to its combinatorially explosive nature and it has not been possible to apply the technique to complex and large structural databases.

Graph G	Graph H	An isomorphism between G and H
		$f(a) = 1$ $f(b) = 6$ $f(c) = 8$ $f(d) = 3$ $f(g) = 5$ $f(h) = 2$ $f(i) = 4$ $f(j) = 7$

Molecular Similarity

Molecular similarity is a measure of the **degree of overlap** between a pair of molecules in some property space.

- It can be calculated for a wide range of molecular properties.
- An important tool in the generation of quantitative structure-activity relationships (QSAR's).
- QSAR aim to link systematically the chemical or biological properties of molecules to their structures (rational drug design)

Why similarity searching is important?

- ✓ Development of combinatorial chemistry and the ability to synthesise and to assay vastly greater numbers of molecules than even a very few years ago requires tools to rationalise the resulting structural and biological data.
- ✓ Simply increasing the throughput of the synthesis and test cycle in itself does not necessarily lead to more high quality lead compounds.
- Many of the techniques that have been developed are based on the concept of molecular similarity.
- Similarity methods have been used for many years and are typically applied early in the drug discovery process when little is known about the biological target.

The main aims and objectives

- 1. Similarity searching methods in chemical data bases** (Identification of new Cliques)
- 2. Exploring Binding Site Similarity** (to find similarities amongst various cavities)
- 3. Ligand Superposition** (Alignment)
- 4. Clustering compounds**

Similarity searching requirements

Definition of :

1. A chemistry space

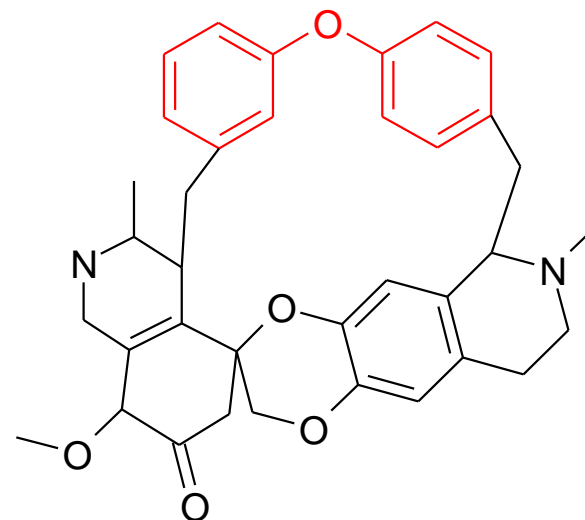
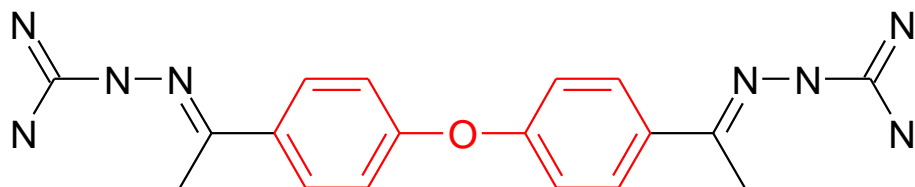
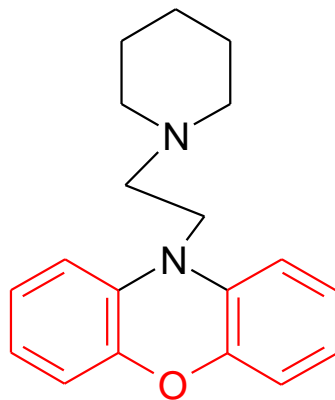
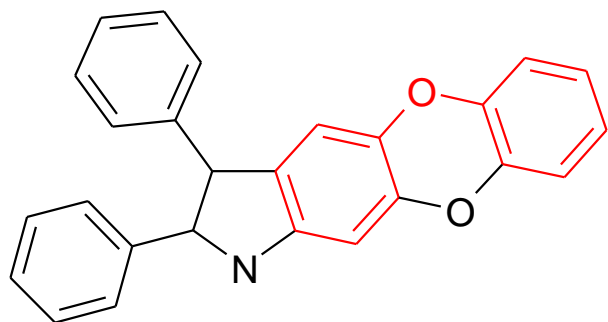
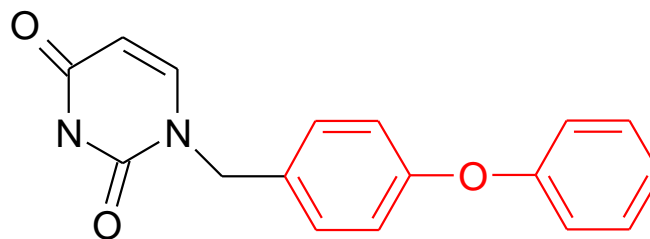
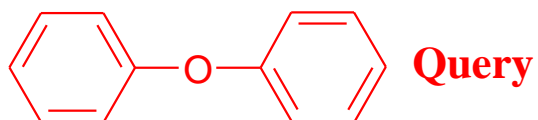
2. Molecular descriptors { 2D, 3D

3. A similarity index

2D descriptors (Physicochemical properties)

- Molecular weight
- Octanol - water partition coefficient
- Total energy
- Heat of formation
- Ionization potential
- Molar refractivity
- ❖ **2D Fingerprints** : record the presence or absence of molecular fragments within a molecule.

2D Substructure Searching

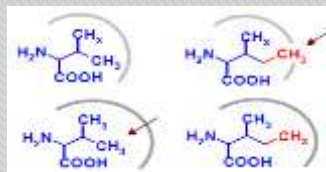


Problem

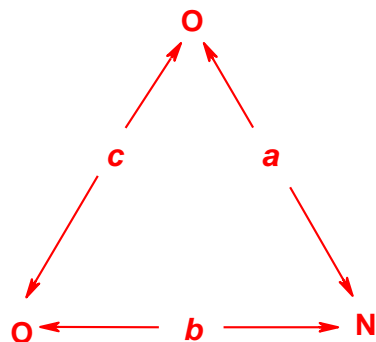
- Structurally similar molecules will tend to have the same properties.
 - Less effective at identifying compounds that have similar activities but that do not share the same structural skeleton.
-
- ✓ Similar Property Principle: Structurally similar compounds will exhibit
 similar physicochemical and biological properties
 - ✓ For lead discovery want a diverse space to locate all possible hits (actives) – called a diverse library

3D Similarity

- Distance-based and angle-based descriptors (e.g. inter-atomic distance)
- Field similarity
 - Comparative Molecular Field Analysis (CoMFA), CoMSIA
 - Electrostatic potential
 - Shape
 - Electron density
 - Any grid-based structural property
- Shape descriptors
 - van der Waals volume and surface (reflect the size of substituents)
 - Molecular Shape Analysis
 - WHIM descriptors (Weighted Holistic Invariant Molecular Descriptors)
- Receptor binding



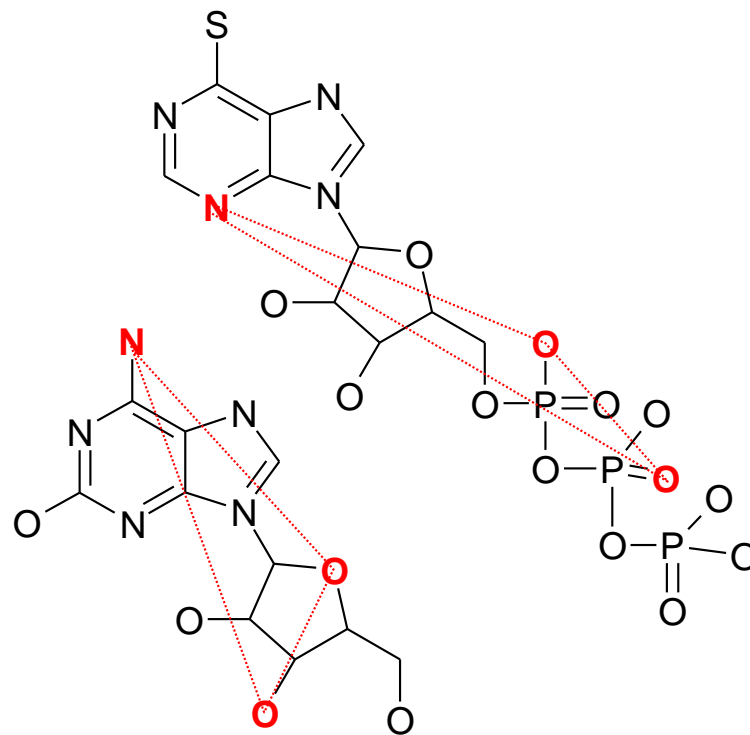
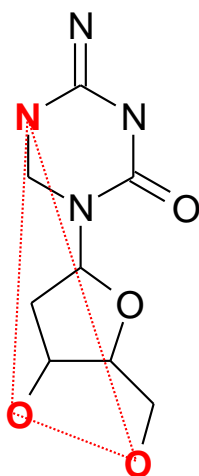
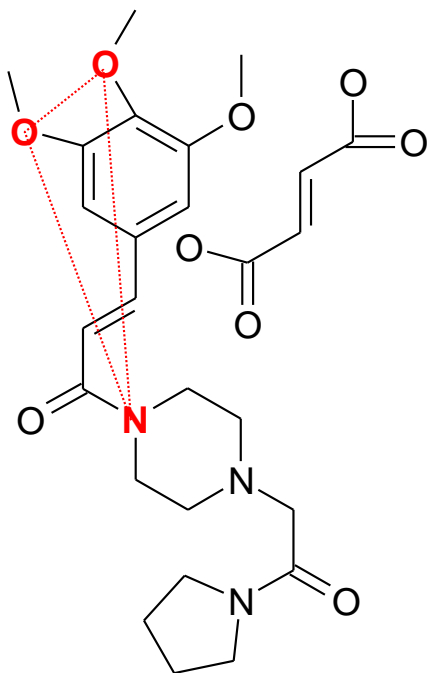
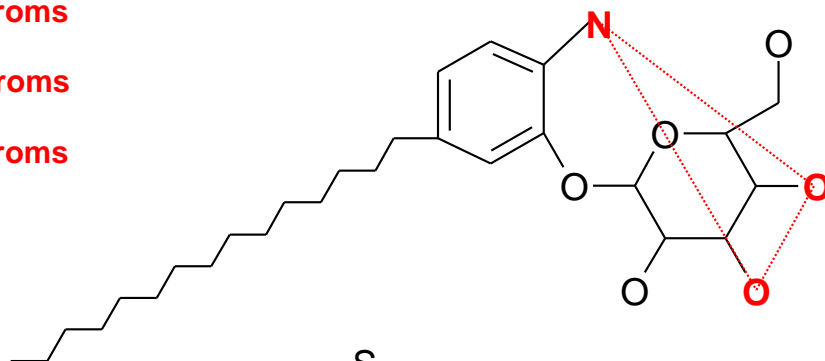
3D Substructure Searching



$a = 8.62 \pm 0.58$ Angstroms

$b = 7.08 \pm 0.56$ Angstroms

$c = 3.35 \pm 0.65$ Angstroms



3D descriptors

- 3D descriptors are highly selective than 2D descriptors
- Handling of conformational flexibility is a major problem !

Creativity is not about inventing something
totally new, it is about making new connections!
(**Albert Einstein**)

History

- A back-tracking graph algorithm for establishing a mapping between two structures was first described by Ray and Kirsch in 1957.
- The use of reduced graph representations of individual structures and the creation of hyperstructures encompassing all the structures in a database
- was first described by **Martin, Y.C., Peter Willett**
- University of Sheffield, UK, 1990.



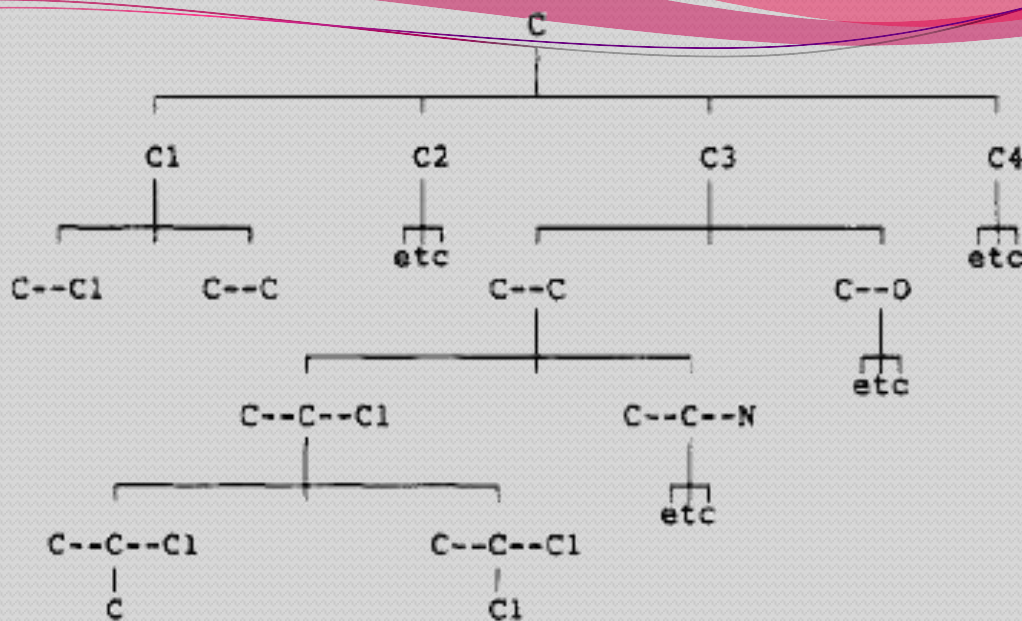
Central atom

Number of
Connections

First
Neighbour

Second
Neighbour

Third
Neighbour



Hierarchical fragment descriptions used in the CIS (Feldmann) substructure search system. Each level in the hierarchy enlarges the description of the fragment.

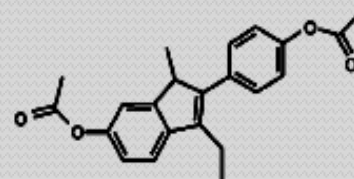
The HierarchicalTree Substructure Search (HTSS)

Each level in the hierarchical fragment tree is effectively part of an hierarchical classification of all the atoms in the database as a whole, initially by number of neighbors and atom type, and then by bonding pattern and atom type of neighbors.

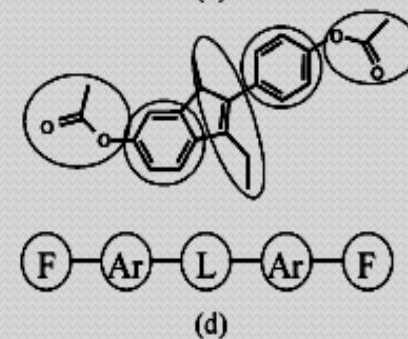
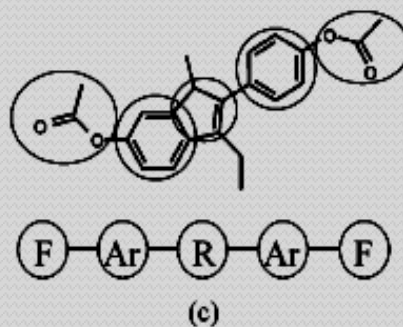
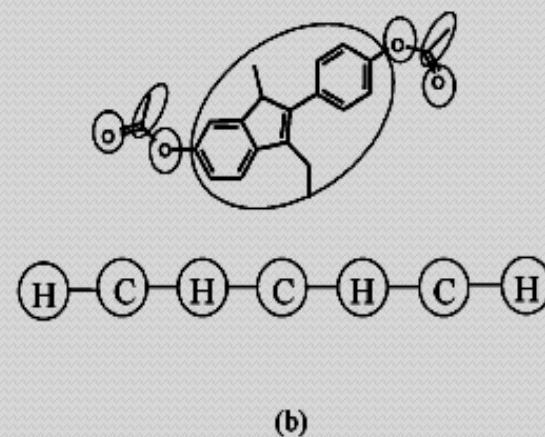
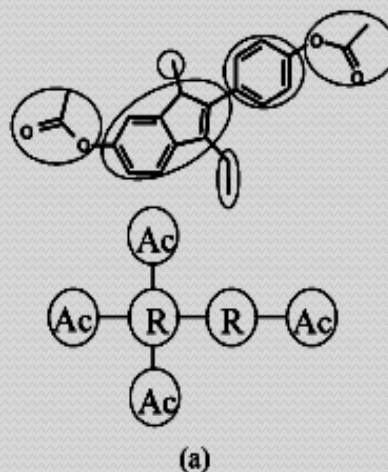
Similarity Searching using Reduced Graphs

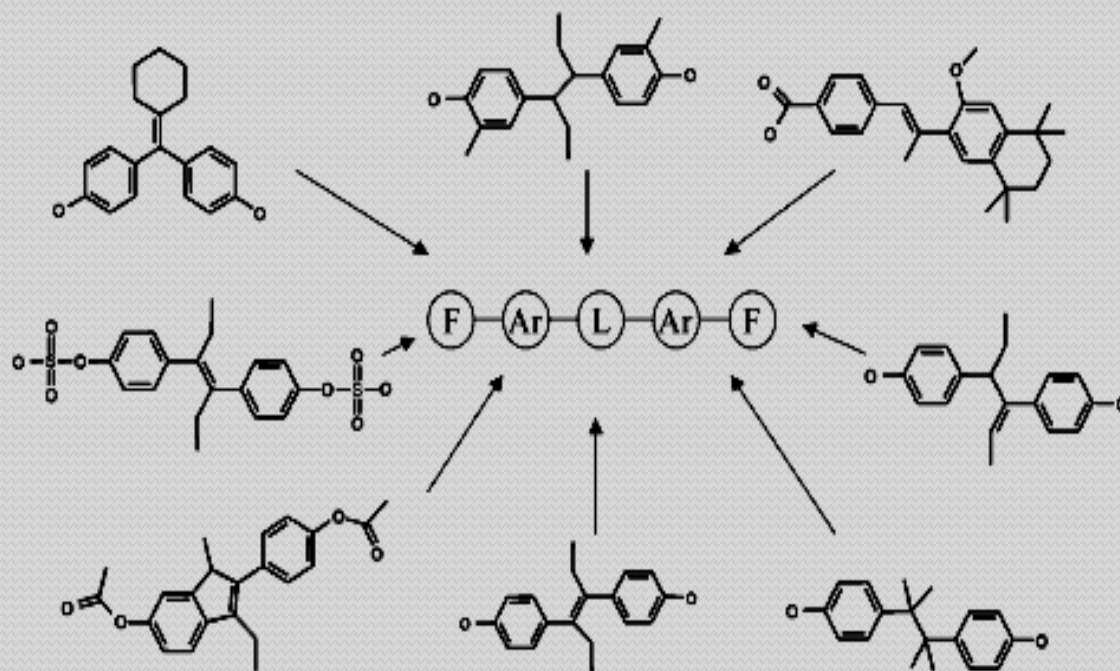
Reduced graphs are topological graphs that attempt to summarize the features of molecules that can result in drug-receptor binding while retaining the connections between the features.

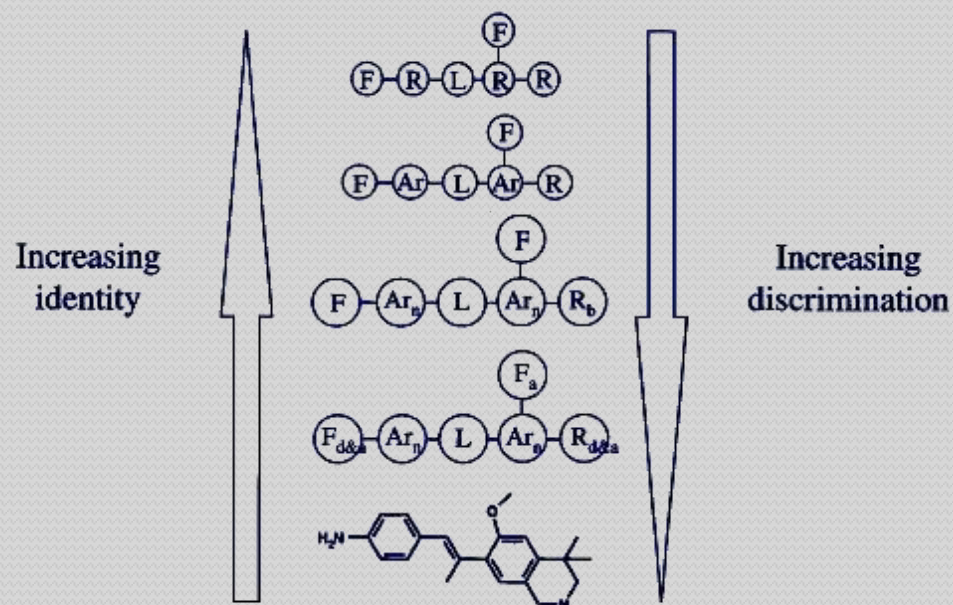
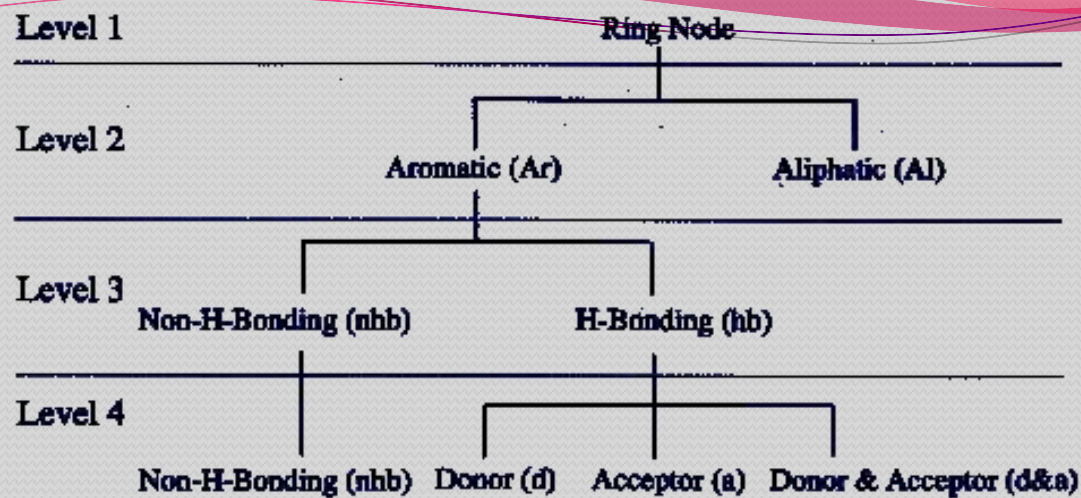
- ✓ potential to find structures that have the same binding characteristics with different carbon skeletons and may belong to different lead series.



- **R** ring systems
- **Ac** acyclic comp.
- **C** carbon
- **H** heteroatom
- **Ar** aromatic rings
- **F** functional group



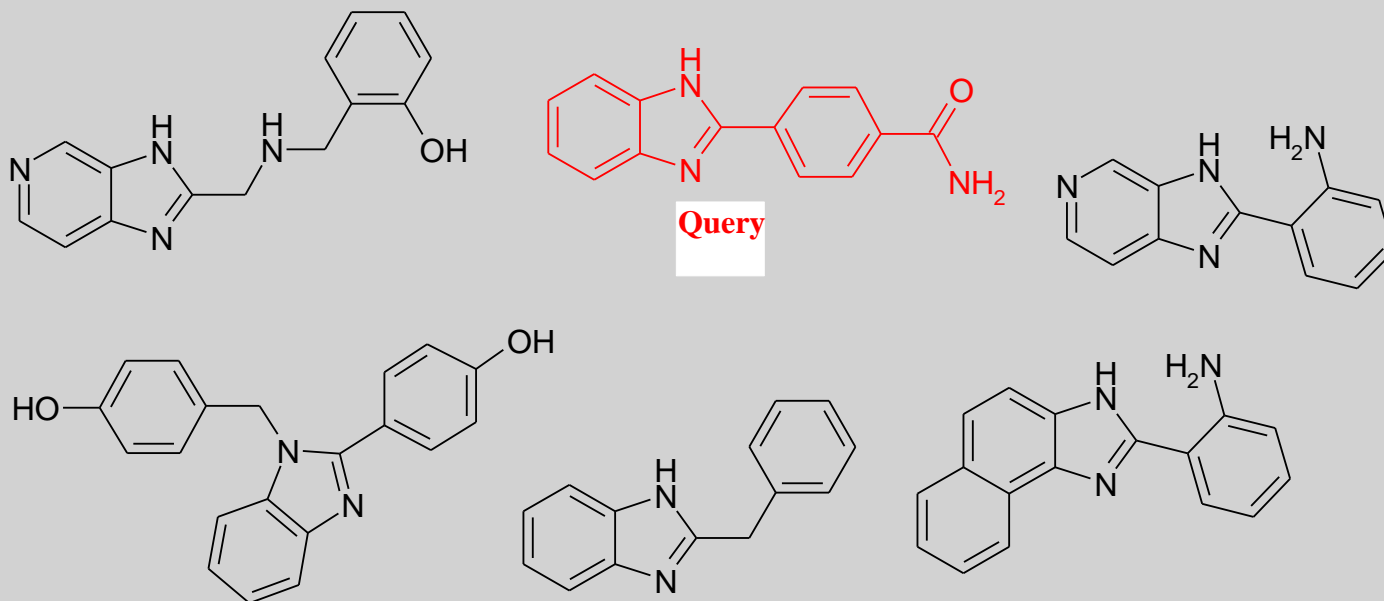




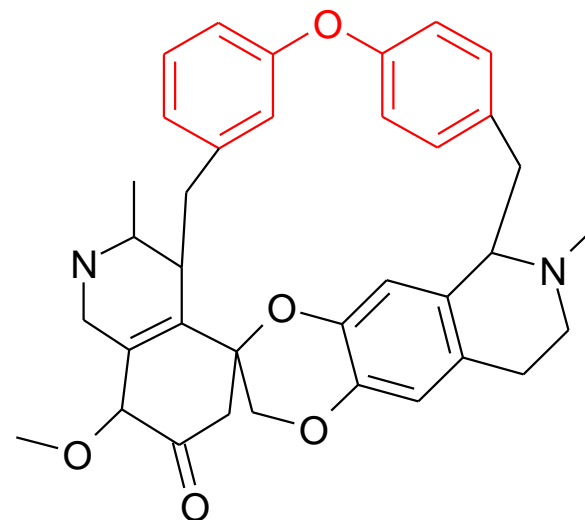
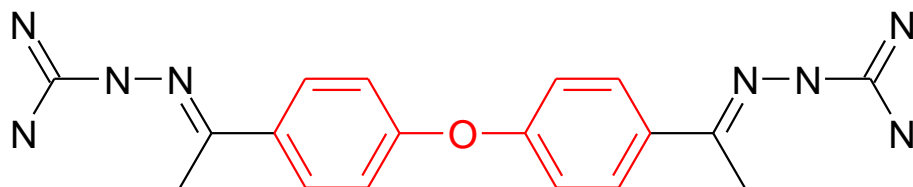
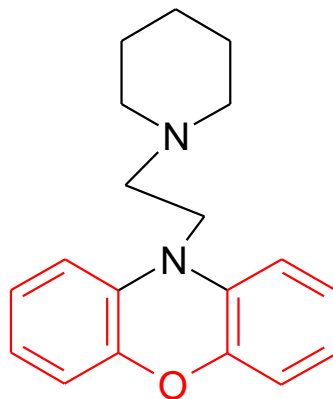
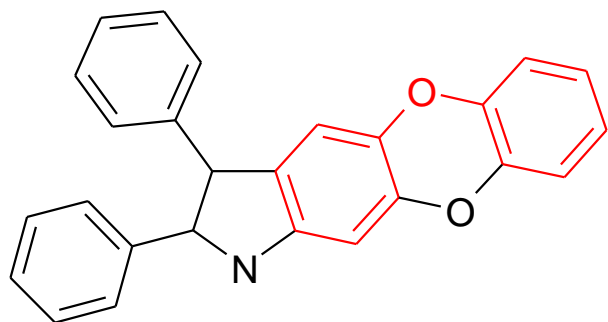
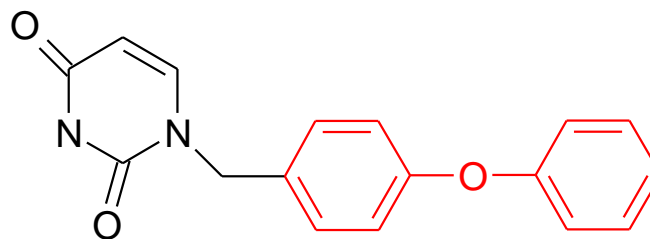
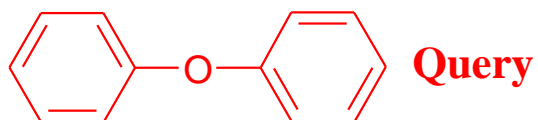
A hierarchy of reduced graphs exists.

- Many ways of computing the similarity between two molecules
 - Different representations
 - Different similarity coefficients

2D Similarity Searching



2D Substructure Searching

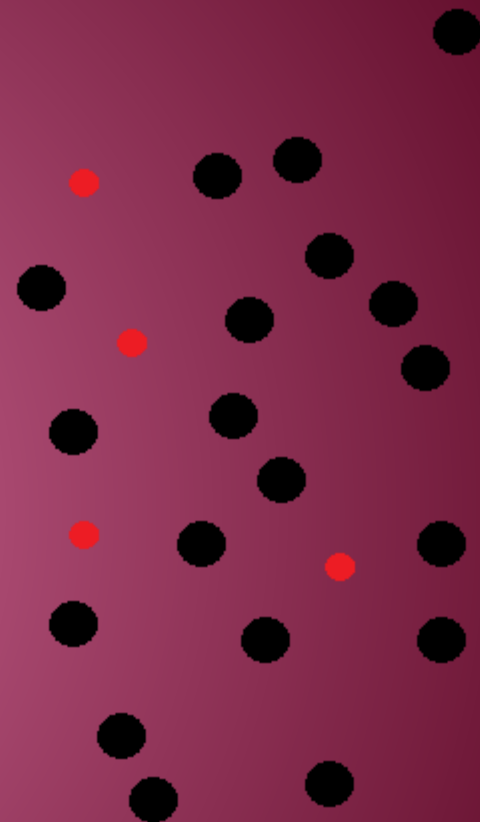


Similarity indices

1.	Jaccard/Tanimoto	$\frac{a}{a+b+c}$	10.	Sokal/Sneath(3)	$\frac{a+d}{b+c}$
2.	Dice	$\frac{2a}{2a+b+c}$	11.	Baroni-Urbani/Buser	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
3.	Russell/Rao	$\frac{a}{n}$	12.	Ochiai/Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$
4.	Sokal/Sneath(1)	$\frac{a}{a+2b+2c}$	13.	Kulczynski(2)	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$
5.	Kulczynski(1)	$\frac{a}{b+c}$	14.	Forbes	$\frac{n \times a}{(a+b)(a+c)}$
6.	Simple Matching	$\frac{a+d}{n}$	15.	Fossum	$\frac{n\left(a-\frac{1}{2}\right)^2}{(a+b)(a+c)}$
7.	Hamann	$\frac{a+d-b-c}{n}$	16.	Simpson	$\frac{a}{\min(a+b, a+c)}$
8.	Sokal/Sneath(2)	$\frac{2a+2d}{a+d+n}$			
9.	Rogers/Tanimoto	$\frac{a+d}{b+c+n}$			

Similarity indices

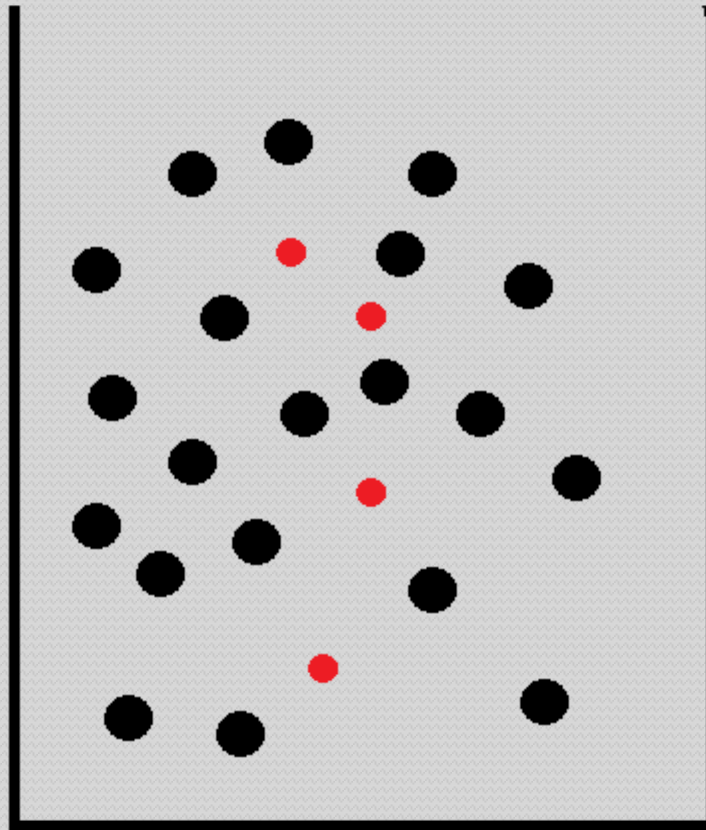
17.	Pearson	$\frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
18.	Yule	$\frac{ad - bc}{ad + bc}$
19.	McConnaughey	$\frac{a^2 - bc}{(a+b)(a+c)}$
20.	Stiles	$\log_{10} \frac{n \left(ad - bc - \frac{n}{2} \right)^2}{(a+b)(a+c)(b+d)(c+d)}$
21.	Dennis	$\frac{ad - bc}{\sqrt{n(a+b)(a+c)}}$



Calculation of similarity

—
—
—
—

—
—
—
—



Enrichment Factor

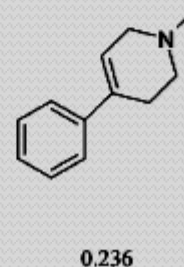
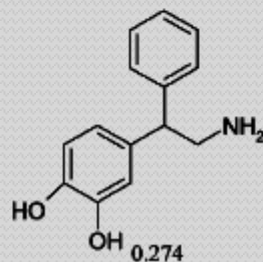
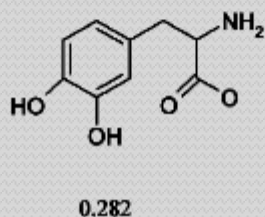
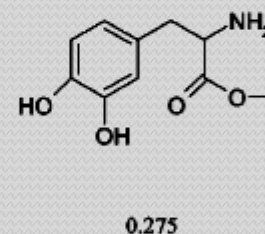
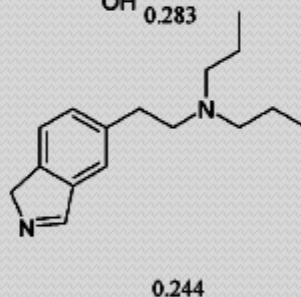
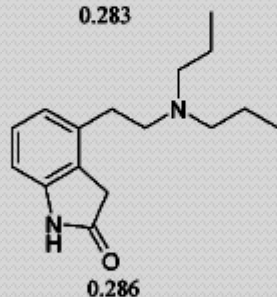
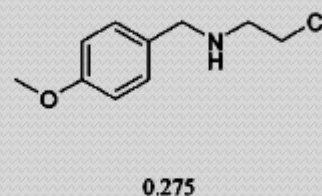
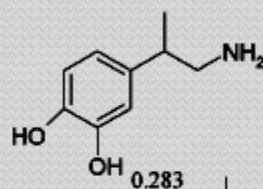
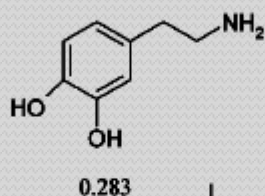
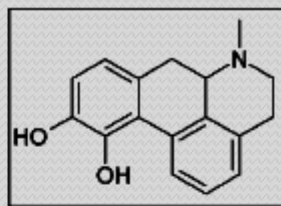
$$EF = \frac{na}{Na \times 0.1}$$

na : Number of actives, in the top of 10% of the ranked hit list

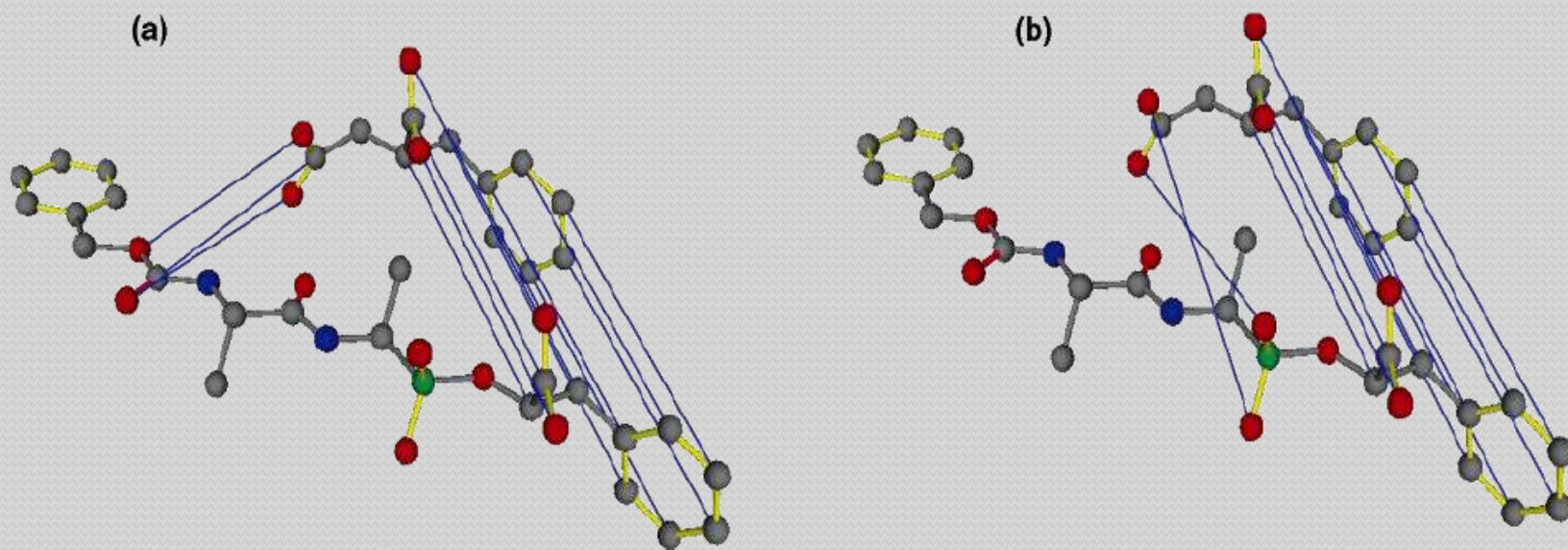
Na : Total number of actives

New Cliques Detection

- Discovery of novel bioactive molecules

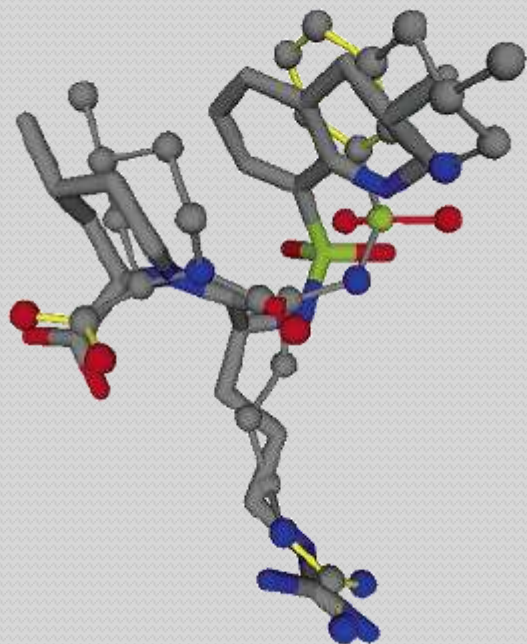


Graph-Based Molecular Alignment (GMA) was first carried out by J.Apostolakis

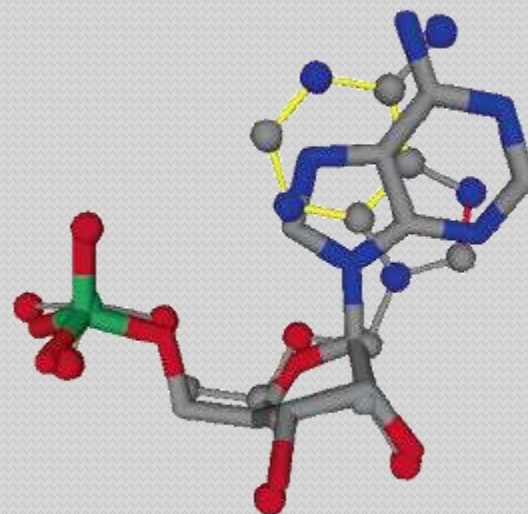


Comparison between aligned pose and crystal structure

Template is kept rigid, query molecule is flexible
By using torsion space optimization



RMSD = 1.29



RMSD= 0.59

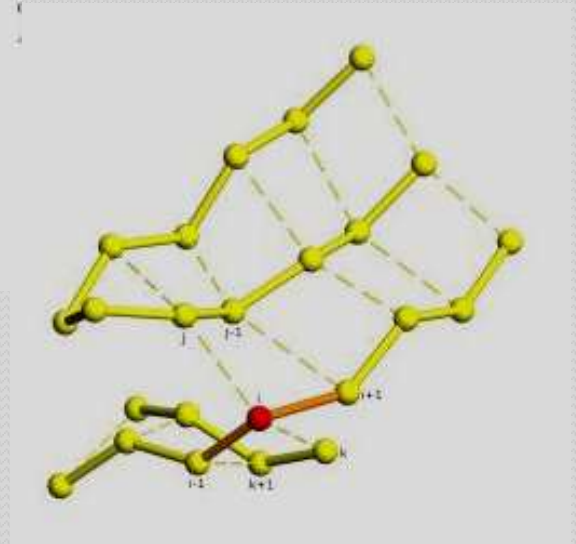
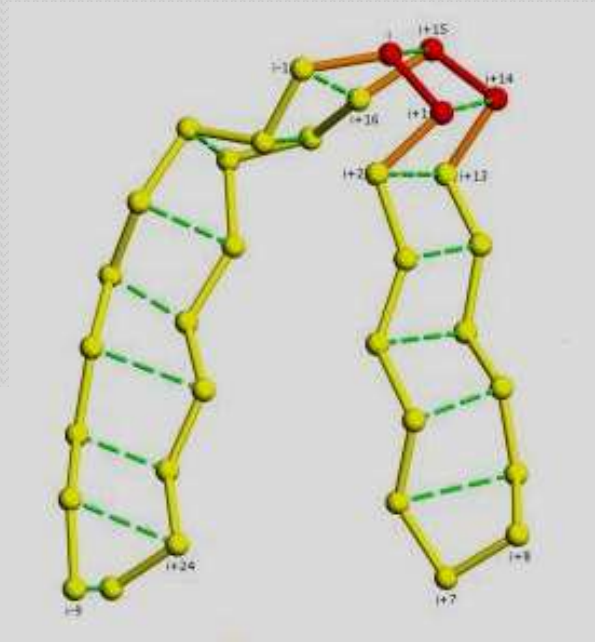
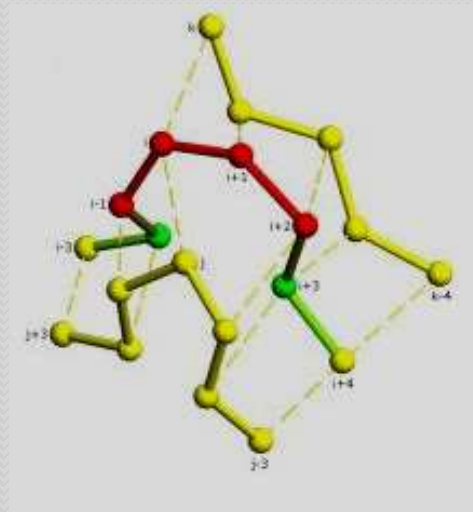
Searching 3D Protein Structures

- Extensive collaboration between Information Studies and Molecular Biology and Biotechnology to develop graph representations of proteins that can be searched with isomorphism algorithms analogous to those used for chemical structures
- Focus here on folding motifs (secondary structure elements) in proteins and protein amino acid side chains

Representation Of Protein Folding Motifs

- The helix and strand secondary structure elements (SSE) are both approximately linear, repeating structures, which can be represented by vectors drawn along their major axes
- The nodes of the graph are these vectors (individual amino acid side-chains) and the edges comprise:
 - The angle between a pair of vectors
 - The distance of closest approach of the two vectors
 - The distance between the vectors' mid-points
- PROTEP compares such representation using a maximal common subgraph isomorphism algorithm to identify common folds

Secondary Structure Elements (SSE)



The majority of PDB files include lines labeled HELIX and SHEET, detailing the residues involved in α -helices and β -sheets within the protein.

- This information is used to label each residue in the vector file with
- ‘h’ to indicate α -helix
- ‘s’ to indicate β -sheet
- ‘x’ for residues that are in neither of these.
- e.g. ASPh
- A minority of PDB files (about 7.5% in the search files) do not have secondary structure information

Data Fusion

- Improved performance can be obtained in many classification tasks by combining evidence from several different sources
 - Calculate the similarity between a user query and each of the documents in a database
 - Rank the documents in order of decreasing similarity
 - Repeat using several different representations, coefficients, *etc.*
 - Add the rank positions for a given document to give an overall *fused* rank position
 - The resulting fused ranking is the output from the search
 - Small, but consistent, improvements in performance over use of a single ranking

Ar/F

M2
M41
M23
M3
M9
M18
.
.
.
.

R/F

M6
M89
M234
M13
M90
M180
.
.
.
.

Fused Data

M8
M130
M257
M16
M99
M198
.
.
.
.

sum

sort

M8
M16
M99
M130
M198
M257
.
.
.
.

Fusion Of Chemical Similarity Coefficients

- Searches were carried out using different similarity coefficients, and the resulting rankings fused to give rankings corresponding to all combinations of 1, 2, 3.... 20, 21 coefficients
- The effectiveness of a combination was evaluated by the number of actives in the 10% top positions of the fused ranking

Advantages Data Fusion:

- Fusion of rankings can provide a small, but consistent, improvement in the effectiveness of searching if an *appropriate* combination of coefficients is chosen
- Data fusion of this sort provides a simple, but highly cost-effective, way of enhancing existing systems for chemical similarity searching

Advantages of using Reduced Graph

- Highly effective at finding compounds that share the same activity but that are based on different lead series.
- High-throughput screening
- Method is capable of analysing large data sets and can deal with noisy data.(High speed)

Structure is not the sole factor for biological activity

- Interactions with environment
 - Solvation effects
 - Metabolism
 - Time dependence
 - More...



References

- Searching Databases of Three-Dimensional Structures , Chapter 6, pp 213-263 Martin Y.C., Willett P.
- Applications of Graph Theory in Chemistry (J. Chem. InJ Comput. Sci. 1985, 25, 334-343)
- Substructure Searching Methods: Old and New (J. Chem. Inf. Comput. Sci. 1993, 33, 532-538)
- Similarity Searching Using Reduced Graphs(J. Chem. Inf. Comput. Sci. 2003, 43, 338-345)
- Searching for Patterns of Amino Acids in 3D Protein Structures (J. Chem. Inf. Comput. Sci. 2003, 43, 412-421)
- Graph-Based Molecular Alignment (GMA),(*J. Chem. Inf. Model.* 2007, 47, 591-601)
- Chemical Similarity Searching (J. Chem. Inf. Comput. Sci. 1998, 38, 983-996)
- Scaffold Hopping Using Clique Detection Applied to Reduced Graphs (J. Chem. Inf. Model. 2006, 46, 503-511)
- Further Development of Reduced Graphs for Identifying Bioactive Compounds (J. Chem. Inf. Comput. Sci. 2003, 43, 346-356)
- Maximum common subgraph isomorphism algorithms for the matching of chemical structures (J. Comp. Aided. Mol. Des, 16: 521–533, 2002)
- A Robust Clustering Method for Chemical Structures (J. Med. Chem. 2005, 48, 4358-4366)
- Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure (J. Chem. Inf: Comput. Sci. 1992, 32, 639-643)

Thank You

