



 POLITECNICO DI MILANO

Dipartimento di
Elettronica e Informazione

Recent Advances in Chemoinformatics

D. K. Agrafiotis, D. Bandyopadhyay, J. K. Wegner, and H. van Vlijmen
Journal of Chemical Information and Modeling, vol. 47, no. 4, pp. 1279-1293

Lecturers:

Francesco Merlo

Antonio Miele

Problems and Approaches in Computational Chemistry - 2008

- Chemoinformatics at a glance
- Areas of application
 - Chemogenomics
 - Free energy and solvation
 - Pharmacophore discovery
 - Conformational analysis
 - Geometric algorithms and combinatorial optimization
 - Molecule mining
 - De-Novo and Fragment-Based design
 - QSAR
- Two approaches to automate the exploration of local/global QSAR
 - T-ANALYZE
 - T-MORPH
- Conclusions

- **What is?**
 - A large scientific discipline that deals with storage, organization, management, retrieval, analysis, dissemination visualization and use of chemical information.
- **Objectives**
 - *In silico* drug discovery and development
- **Issues**
 - High-throughput experimental techniques
 - Need to analyze very large data sets
- **Research areas**
 - In between of Chemistry, Biology and Computer Science

Conformational analysis and Pharmacophore discovery

QSAR

Free energy and solvation

Molecule mining

Chemoinformatics

Chemogenomics

De-novo and fragment based design

Geometric algorithms and combinatorial optimization

- **Objective**
 - Study of genomic response to chemical compounds
 - Rapid identification of novel drugs and drug targets by using multiple early-phase drug discovery technologies
 - (e.g., target identification and validation, compound design, chemical synthesis and biological testing)
- **Issues**
 - Advances in automation, liquid handling and data analysis have speeded up biological reaction analysis and improved their quality
 - Compounds can be tested at different concentrations differently from the past
 - Explain and generate SAR, and build a chemical genomic map
 - Understand the relationships between chemical compounds and protein families
 - Genomic data and relationship between chemical structure and biological target are public but pharmacological data are private
 - Critical need to organize and analyze this information and apply at virtual screening
 - Novel methods consider 3D data organization and fingerprint-based approaches to analyze interactions between protein and small-molecule complexes

- **Objective**
 - Accurately predict binding free energy by considering solvation effects and their impacts on electrostatic interactions
 - Electrostatic interactions are required to analyze protein-ligand binding
- **Issues**
 - Existing models of solvation effects are too slow for large scale virtual screening even if very accurate
 - Generalized Born (GB)
 - Poisson-Boltzmann (PB)
 - Reducing the scenario under inspection or using approximations may help in computation
 - Precalculate computational heavy matrixes
 - Eg.: Gaussian Surface-Generalized Born (GSGB) allows the computation of electrostatic solvation forces at every point in space by approximating electron density with gaussian functions and by using analytical gradients on gaussian surfaces
 - Results are variable; sometimes not always realistic
 - Binding free energy modeling presents similar accuracy problems

- **Objective**
 - Identify molecular frameworks that carry the essential features responsible for a drug's biological activity
 - The pharmacophore is the spatial arrangement of electronic features able to trigger biological response of biomolecular target
- **Issues**
 - Current approaches use manual curation or consensus to remove active compounds that have different binding modes from those identified by sampling the conformational space
 - Automated ligand-based drug design fails when some compound has a different binding from the rest
 - Requires to compare N molecules with M conformations
 - PharmID applies a statistical sampling comparing each molecule to a model of the active conformation and its key features

- **Objective**

- Identify which crystal structures (of proteins or new compounds) are bioactive
- Analysis tools are based on stochastic 3D models that provide samples of the different (and possible) conformations for:
 - Protein docking
 - Pharmacophore modeling
 - 3D QSAR

- **Issues**

- Stochastic 3D modeling techniques are very sensitive to:
 - Starting configurations
 - Random number effects
- As a consequence, results are difficult to reproduce
- This reproducibility problem affects systematic methods used by well-known softwares as well (Corina, Omega, Catalyst and Rubicon)

Geometric algorithms and Combinatorial optimization

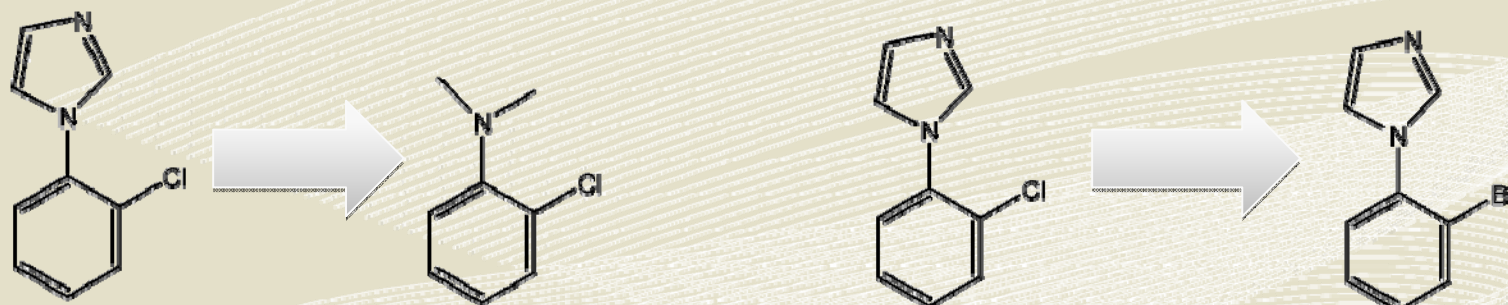
- **Objective**
 - Represent and analyze 3D models of drugs and protein structures in order to derive chemical characteristics
- **Issues**
 - Exploration of chemical and biological spaces is a needle-in-the-haystack combinatorial problem
 - Find suitable geometrical and combinatorial structures and methods to help facing such problem
 - Voronoi diagrams and Delaunay tessellations capture the nearest neighboring relationships among a set of points (atoms) in 3D
 - however, they are unstable in face of small perturbations
 - Approximate the largest common point set problem (LCP), which is NP-hard
 - Used to find the smallest set of features responsible for a biological effect
 - MultiBind approaches small problems (up to 100 proteins)

- **Objective**
 - Derive similarities in chemical/biological activity by considering suitable generic similarity measures for molecules
 - Base principle: a small change in the structure causes a small change in activity
- **Issues**
 - Define suitable coding schemes...
 - SMILES, InChI...
 - ... or molecule representations
 - Graph-based
 - Define suitable similarity metrics among schemes/representations
 - Graph edit distance
 - Best practices suggest that best similarity metrics should:
 - Include problem-specific expert knowledge (i.e., pharmacophore types)
 - Use multiple atom-and-bond or atom-and-feature properties

- **Objective**
 - Generation of novel chemical entities clearly distinguished from competitive products
 - Quick and effective means to access the uncharted chemical space
- **Issues**
 - How to "grow" molecules from smaller fragments
 - Mainly used to discover "molecular frameworks" that must be further refined
 - Assessing the right balance between novelty and complexity is an art
 - How to score identified molecules against desired objectives
 - There are no systematic studies on finding active molecules/compounds as raw output of such techniques
 - How to direct the discovery algorithms toward "interesting" areas
 - Common critics on the synthetic accessibility of resulting molecules

- **Objective**
 - Use of statistical methods (regression, pattern recognition or machine learning) to derive quantitative mathematical relationships linking chemical structures and biological activity
- **Different (yet complementary) approaches**
 - Global QSAR vs Local QSAR
- **Issues**
 - How to effectively automate the exploration of global/local QSAR
- R. P. Sheridan, P. Hunt, and J. C. Culberson, "Molecular transformations as a way of finding and exploiting consistent local QSAR", *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 180-192, 2006.

- A molecular transformation is common concept in chemistry
 - It consists in a small change to a chemical structure
 - Removing or replacing a substituent or an atom
 - Imidazole by Dimethylamino
 - Chlorine by Bromine



- Two ways for representing transformations
 - Substructure descriptor difference vectors
 - Set of atoms remaining once a maximum common substructure is eliminated
- Two algorithms for
 - Displaying set of closely related compounds
 - Proposing changes to a more active one

- **Global QSAR**

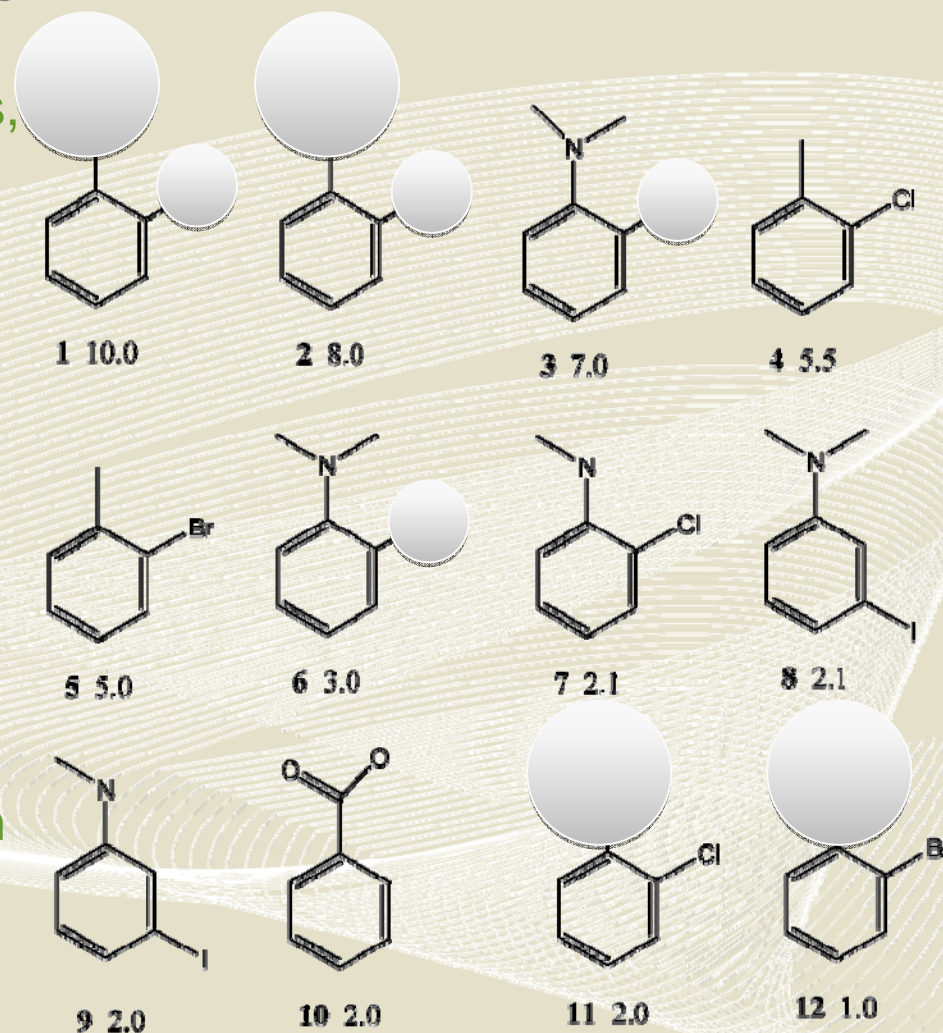
- Given a set of compounds, the global QSAR approaches summarize structure-activity over the whole set

- Influenced by most active and most inactive compounds

- **Local QSAR**

- Compares related compounds, pairwise or in small numbers, to identify beneficial changes

Imidazole



Carboxilate

- **Objective**
 - Take a large data set of compounds and organize it so that similar transformations are grouped together and trends in activities can be easily perceived

- **Procedure**
 1. Start with a set of connection tables and corresponding activities for a data set

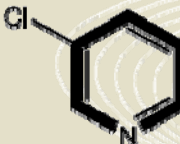


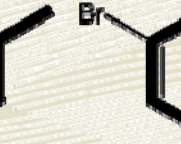
 2. Generate topological descriptors for the molecules
 - Each molecule is modeled as vector of substructure descriptors and their frequency
 - Atom Pair (AP)
 - Topological Torsion (TT)

3. Find pairs of molecules that are more similar than a user-defined cutoff
 - To perform meaningful comparisons, the transformation between the molecules must involve small changes compared to the size of the molecule
 - Similarity computed using Dice index on substructure descriptors
4. For each similar pair, define the transformation descriptor as the difference of descriptor vectors $A \rightarrow B$ and $B \rightarrow A$
 - Find the maximum common substructure (MCS) of A and B and find the atoms that remain when the MCS is deleted (RECS). This is a “MCS transformation”
 - Discard transformations with more than one separate fragment in RECS
 - Compute RECS hash string

	Cl		Br	
	A		B	
molecule	A	B	A→B	
activity	3.0	2.0	1.0	
C10Br1004	0	0	0	
C21Br1002	0	1	-1	
C21Br1003	0	2	-2	
C21Br1004	0	1	-1	
C21Cl1002	0	0	0	

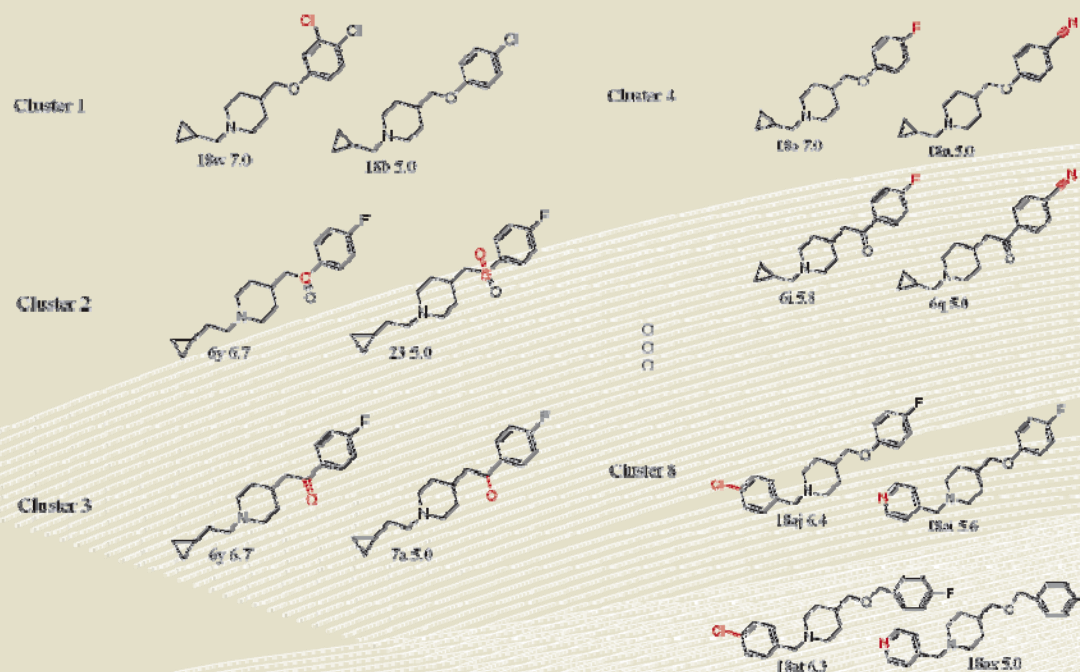
5. Compare pairs of transformations $A \rightarrow B$ and $C \rightarrow D$ to find those pairs that are “congruent”

- Two pairs are considered congruent iff:
 1. A, B, C and D are distinct molecules
 2. A, B, C and D are similar to each other (A, B and C, D are provided as similar from step 3; it has to be considered A vs C and D, and B vs C and D)
 3. The descriptor-based changes goes in the same direction
 4. $\text{Hash}(\text{RECS}(A \rightarrow B)) == \text{hash}(\text{RECS}(C \rightarrow D))$

						
			Cl Br			Cl Br
molecule	A	B	$A \rightarrow B$	C	D	$C \rightarrow D$
activity	3.0	2.0	1.0	2.5	1.0	1.5
C10Br1004	0	0	0	0	1	-1
C21Br1002	0	1	-1	0	1	-1
C21Br1003	0	2	-2	0	1	-1
C21Br1004	0	1	-1	0	1	-1
C21C1002	0	0	0	1	1	0

6. Cluster the transformations based on their congruency in step 5
 - Eliminate redundant clusters
 - Clusters are not overlapping
 - The Butina clustering algorithm is applied

7. Add activity data to the clusters. Sort the clusters and display to the user
 - Activity difference: (diff)
 - Average cluster activity difference: <diff>
 - Cluster Agreement: time fraction the transformations go in the same direction
 - Z-score = <diff> / standard deviation



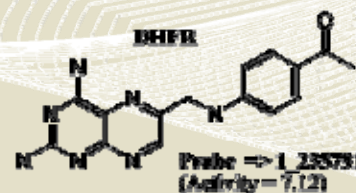
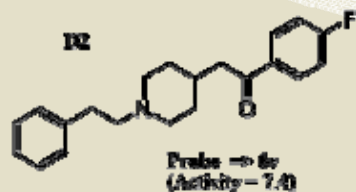
- Complexity = N^4 (comparison of pairs of pairs)
- The paper presents three different applications (Dopamine agonists from re 6, DHFR inhibition for the rat liver enzyme, ACE inhibitors)
- Performances: 36h for elaborating 20000 compounds on 80 processors

- **Objective**
 - Given a compound on hand, analyze and suggest the changes that can be made to improve the activity based on the transformations within the data set.

- **Procedure**
 1. Start with a set of connection tables and corresponding activities for a data set

 2. Generate topological descriptors for the molecules
 - Each molecule is modeled as vector of substructure descriptors and their frequency
 - Atom Pair (AP)
 - Topological Torsion (TT)

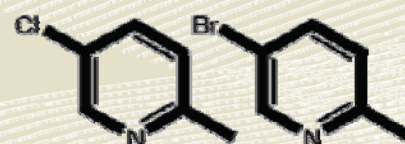
- Find pairs of molecules that are more similar than a user-defined cutoff.
 - To perform meaningful comparisons, the transformation between the molecules must involve small changes compared to the size of the molecule
 - Similarity computed using Dice index on substructure descriptors
- For each similar pair, define the descriptor transformations as descriptor difference vectors $A \rightarrow B$
 - $B \rightarrow A$ are computed online by negating the sign (save space and computation time)
 - MCS transformations are not used
- Calculate the descriptors for a probe molecule
 - The probe molecule is provided by the user together with a desired direction for activity change and an equivalence criterion



6. Identify all transformations in the database that could apply to the probe molecule

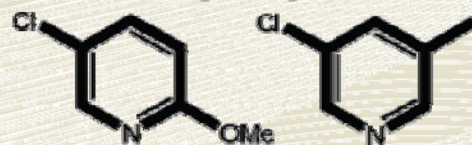
- Compare the frequency of all negative TT's in the stored transformation to the frequency of that descriptor in the probe molecule. If the probe contains all instance to be removed the transformation can be applied
- If the original transformation is not applicable, it is negated ($B \rightarrow A$)

molecules in the transformation



Transformation A→B

examples of probes

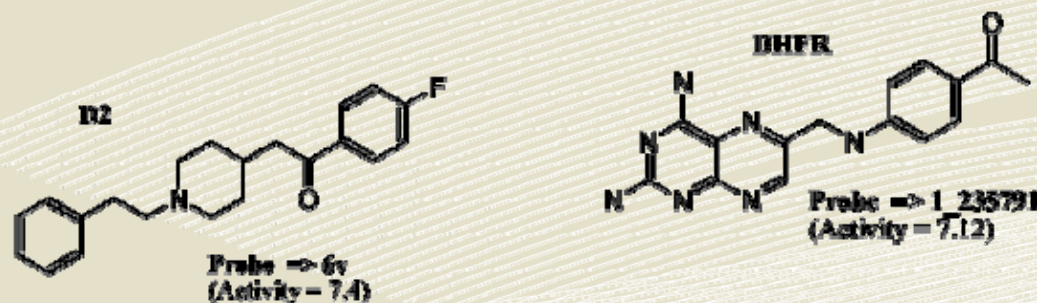


TT descriptors	A	B	A→B	E	F
Cl10C31C21N21	1	0	-1	1	1
Cl10C31C21C21	1	0	-1	1	0
Br10C31C21N21	0	1	+1	0	0
Br10C31C21C21	0	1	+1	0	0
C31C21N21C31	1	1	0	1	0
C31C21C21C31	1	1	0	1	0
C21N21C31C10	1	1	0	0	0
C21N21C31C21	1	1	0	1	0
N21C31C21C21	1	1	0	1	0
Cl0C31C21C21	1	1	0	0	0
N21C21C31C21	1	1	0	1	2

7. Cluster the applicable transformations.
 - A selfsimilarity matrix is computed by using normalized dot product of all pairs of vectors
 - A hierarchical clustering method (average linkage) is applied

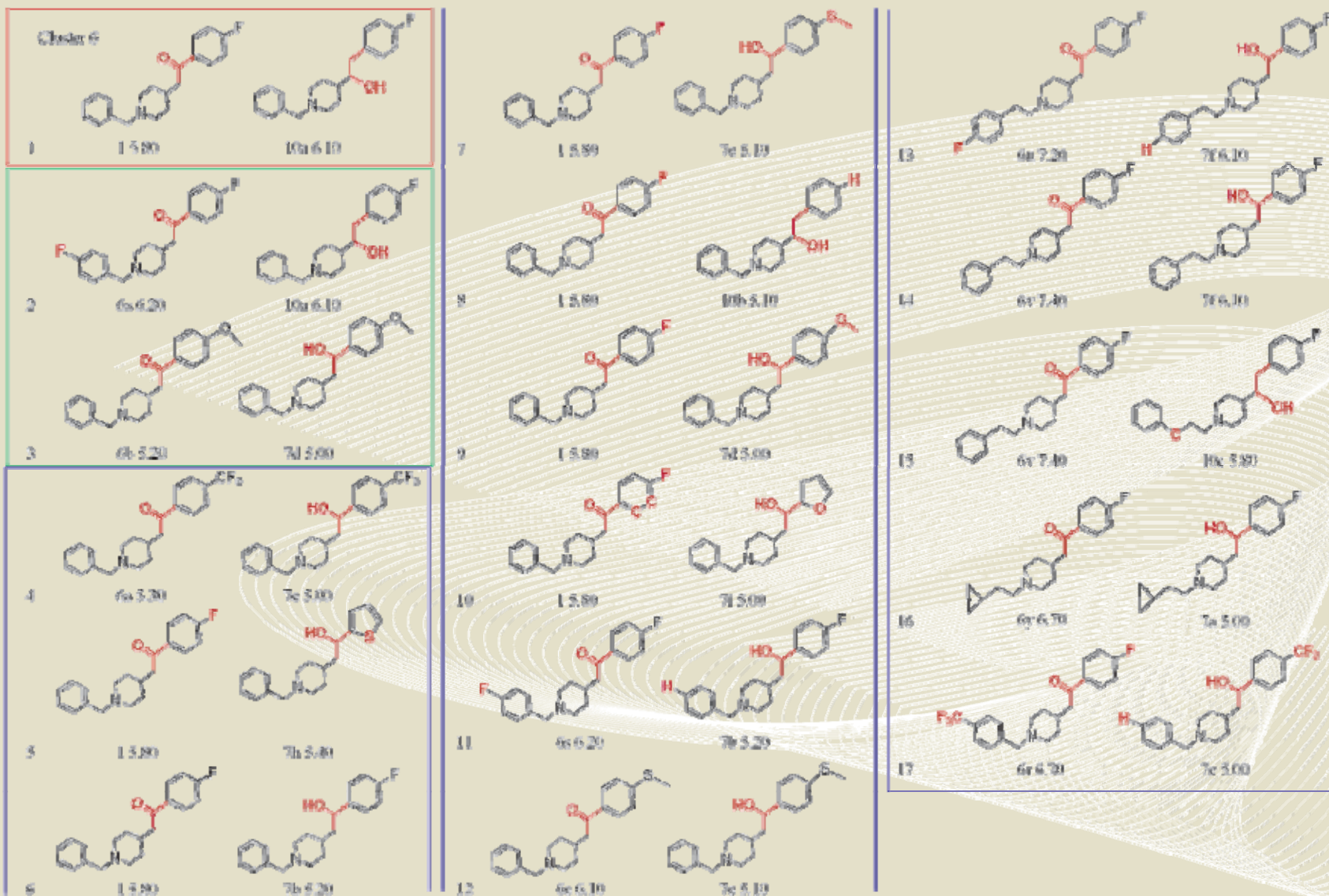
8. Add the activity data, sort the transformations within each cluster, and present to the user
 - Clusters are ordered for decreasing size
 - Transformation with largest desired effect is the cluster representative
 - Shown data: compound identifiers, their activities, the count of increasing/decreasing activity compounds (by using the equivalence criterion)
 - Transformation are sorted by Diff and ordered by direction (other criterions can be used)

- Complexity = N^2
- The paper presents two applications with two probes in different datasets (Dopamine agonists from re 6, DHFR inhibition for the rat liver enzyme)



- Performances:
 - 2h for analyzing a dataset of 6371 compounds and building transformation database (121743 transformations)
 - 15 min for searching and clustering a set of 1800 transformations

T-MORPH – Results



- T-ANALYZE and T-MORPH are a methodology that mimicks local QSAR but makes the process automatic and more systematic
 - The methodology aims only at organizing existing data
- T-ANALYZE helps to understand local QSAR in a large unfamiliar data set
- T-ANALYZE is a complement to global QSARs
- T-MORPH helps seeing which transformations are applicable and their effects