Kazan Federal
UNIVERSITY

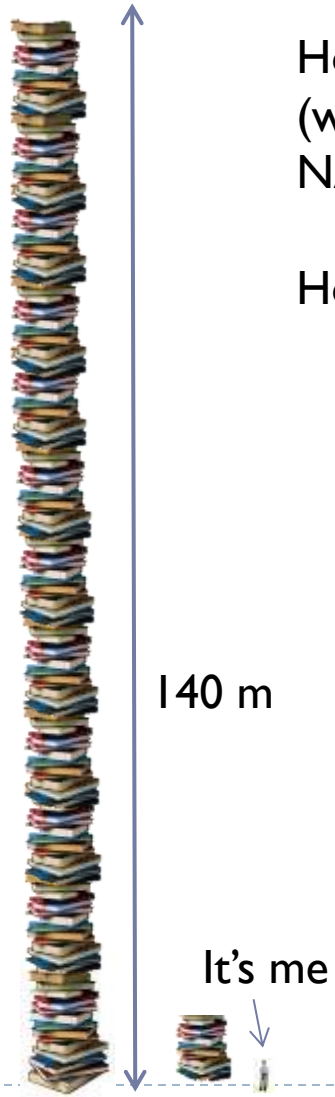
CHEMICAL STRUCTURE REPRESENTATION AND DATABASING

Dr. Timur I. Madzhidov
Kazan Federal University, Department of Organic Chemistry

Chemical Databases: Why?

>120 000 000 organic and inorganic compounds in the largest database
(> 6 millions added each year)

Acetone
Benzene
DMSO
Aspirine
Pyridine
Clorohexidine
Hexane
Heptane
Nonane
Decane



How to STORE such amount of paper (we only calculated place required for NAMES of compounds)???

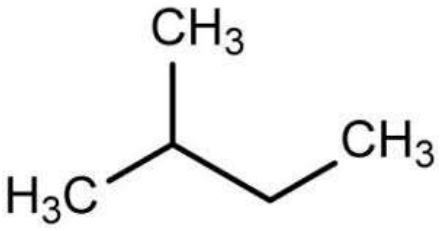
How to search required information???

1957 – Ray and Kirsch* reports invention of the first computer database and structure retrieval system
1966 – CAS finish to develop their own database (the greatest in the world)

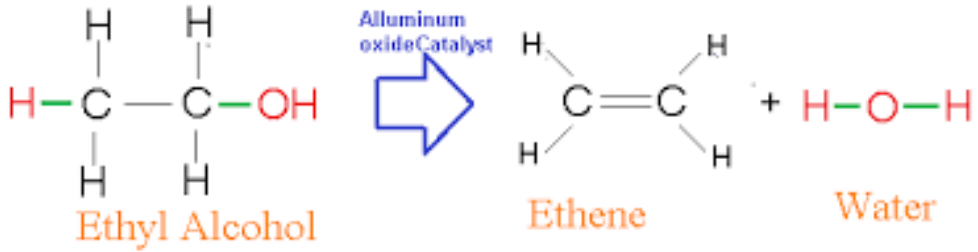
How to represent compounds?
How to search?

* Ray, L.C. and R.A. Kirsch, *Finding Chemical Records by Digital Computers.* Science, 1957. **126**: p. 814-819

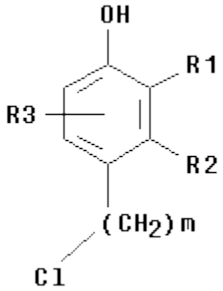
Data in Chemistry



Compounds

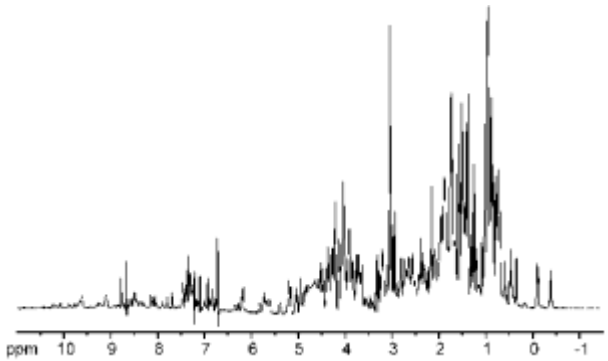


Reaction

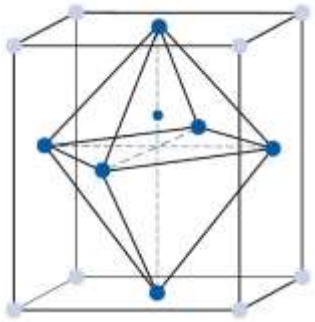


Substituent Variation:	R1 is methyl or ethyl
Homology Variation:	R2 is alkyl
Position Variation:	R3 is amino
Frequency Variation:	m is 1-3

Markush



Spectra



Crystal data

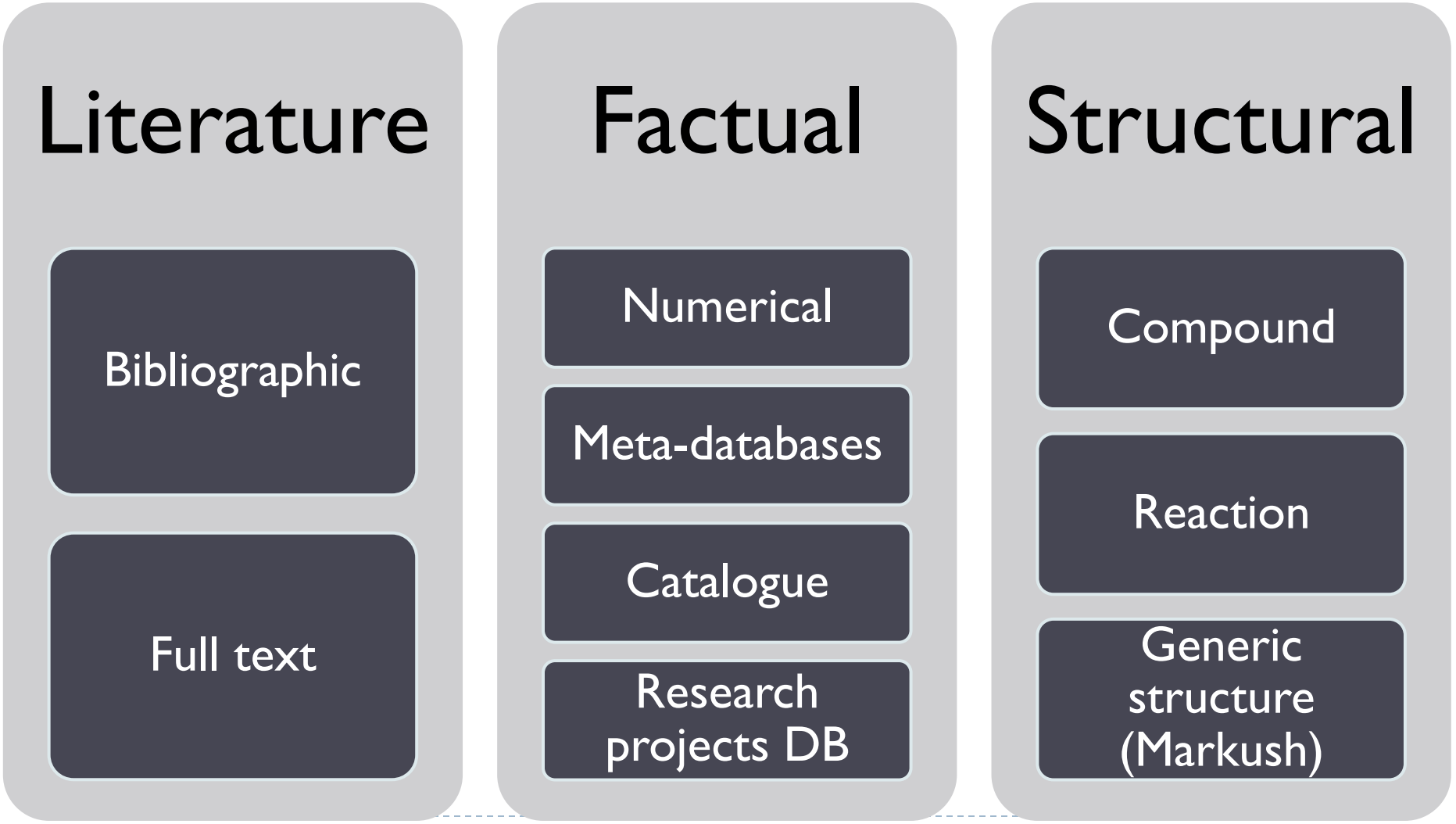
Aspirine

Names

MP 254°C

Properties

Types of databases in chemistry



- Gasteiger, T. Engel (Eds.), *Chemoinformatics. A textbook*. 2003, Weinheim: John Wiley & Sons, 680 p

Types of computer representation of objects

Possible molecule representation in computer

Image

- Very chemist-friendly
- Absolutely not computer-friendly
- Detailed
- Not capacious

Alphanumeric strings (text)

- Chemist-friendly
- Not computer-friendly
- Detailed
- Capacious

Numeric strings

Bit string

- Absolutely not chemist-friendly
- Very computer-friendly
- Loss of information
- Very capacious

Decimal number string

- Not chemist-friendly
- Computer-friendly
- Loss of information
- Capacious

Tables

- More or less chemist-friendly
- More or less computer-friendly
- Detailed
- Not capacious

One can look for such a MEANINGFUL numbers that would be important for description of some properties of molecules (DESCRIPTORS)

Molecule representation should be…

- Computer-readable (no comments)
- Easy to operate with (there should be an algorithms that efficiently handle representation, e.g. image is bad)
- Capacious (to store the information)
- Unique (to store and find information)
  - One molecule → one representation
- Univocal
  - One representation → one molecule
- Invertible
  - Molecule ⇔ Representation

benzene

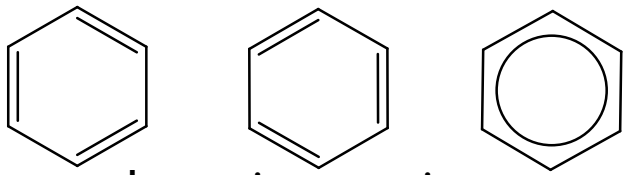
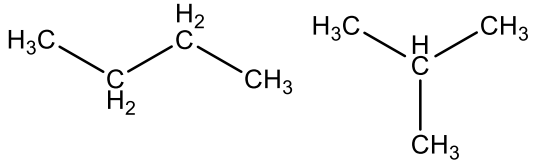


Image is not unique

C_4H_{10} - not univocal

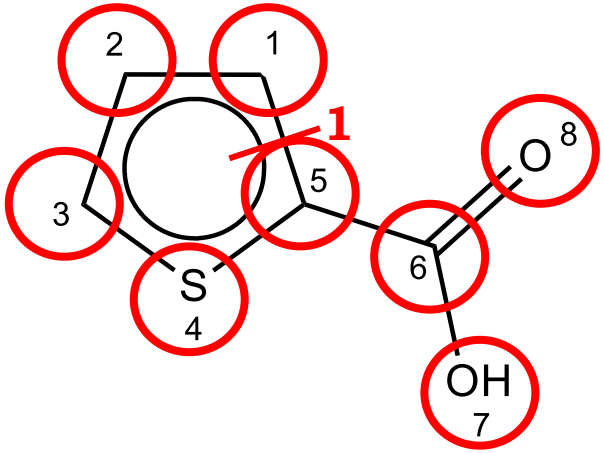


Linear notations

- Hill's formula : C6H6 , C6H6O3S , C10H8NO2
- Chemical name
  - Trivial or trade name : Vitamin B12, Cyanocobalamine
  - Systematic name : (Cyano-κC)[(1R,2R,3R,4R,6Z,8S,11Z,13S,14S,16Z,18S, 19S)-3,14,19-tris(2-amino-2-oxoethyl)-8,13,18-tris(3-amino-3-oxopropyl)-4-[3-({2-[({[(2R,3S,4R,5S)-5-(5,6-dimethyl-1H-benzimidazol-1-yl)-4-hydroxy-2-(hydroxymethyl)tetrahydrofuran-3-yl]oxy}phosphinato)oxy]propyl}amino)-3-oxopropyl]-1,4,6,9,9,14,16,19-octamethyl-20,21,22,23-tetraazapentacyclo[15.2.1.12,5.17,10.112,15]tricosa-5(23),6,10(22),11,15(21),16-hexaen-20-yl]cobalt
- SMILES : NC(=O)C[C@@]8(C)[C@H](CCC(N)=O)C=2/N=C8/C(/C)=C1/[C@@H](CCC(N)=O)[C@](C)(CC(N)=O)[C@@](C)(N1[Co+]C#N)[C@@H]7/N=C(C(\C)=C3/N=C(/C=2)C(C)(C)[C@@H]3CCC(N)=O)[C@](C)(CCC(=O)NCC(C)OP([O-])(=O)O[C@@H]6[C@@H](CO)O[C@H](n5cnc4cc(C)c(C)cc45)[C@@H]6O)[C@H]7CC(N)=O
- InChI : InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12) – for Aspirin
- SLN : CH3C(=O)OH – for acetic acid

SMILES

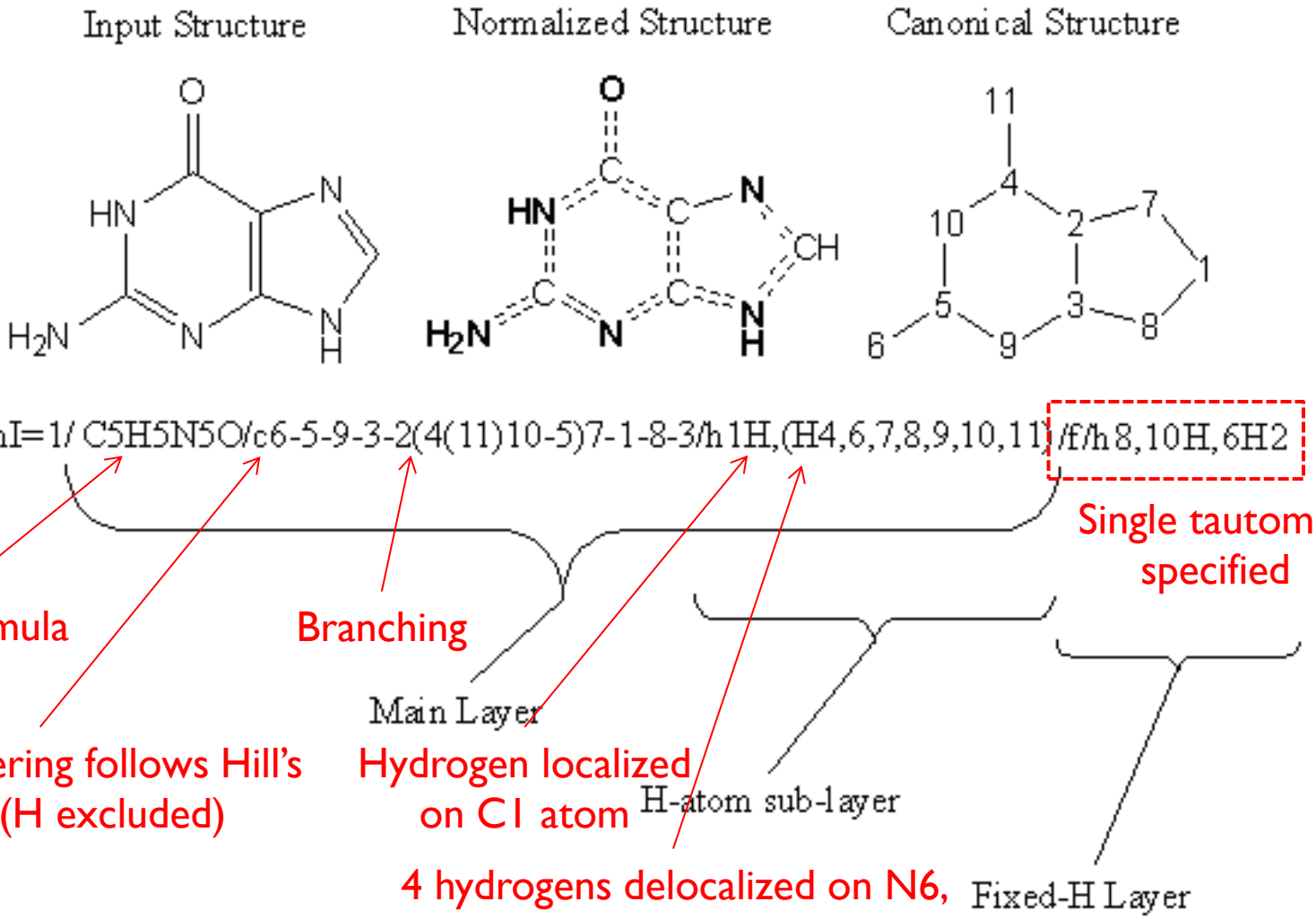
- Atoms: as their symbols, aliphatic – upper case, aromatic – lower case letters, hydrodens hide
- Order of atoms: the order in the detour
- Bonds: single – not specified, double =, triple #
- Branching: in brackets
- Cycles: bond cleavage, its number is written just after atoms



c

- http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

InChI

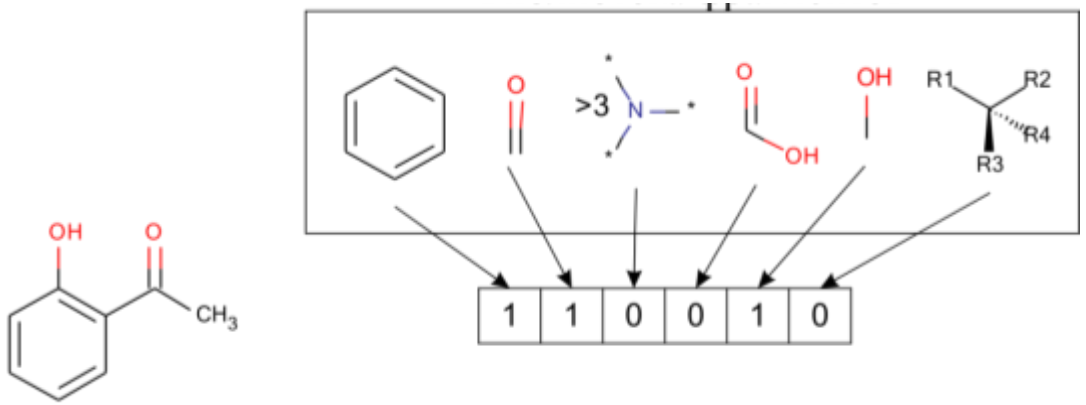


- http://www.inchi-trust.org/fileadmin/user_upload/html/inchifaq/inchi-faq.html#4

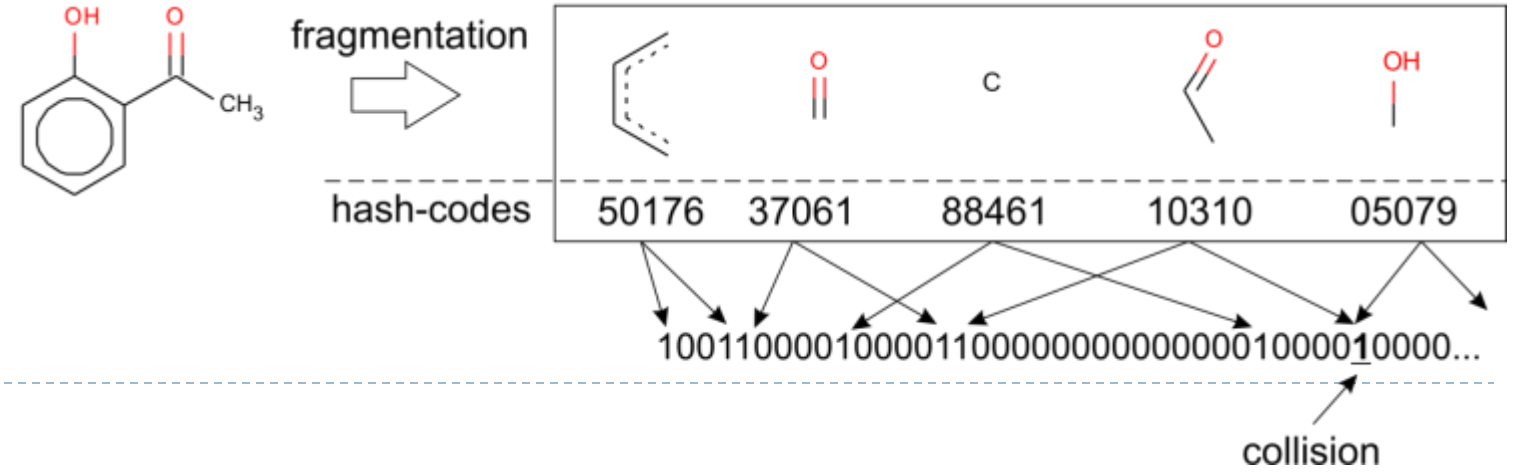
Bitstrings

Every bit in the string (digit of binary number) means the existence (1) or absence (0) of some substructure

- Structure keys – fragments are predefined in fragment library



- Hashed fingerprints - fragment generated on the fly, and the position of unity (1) in the bit string is defined by special *hashing* algorithm that return hash-code that defines address.

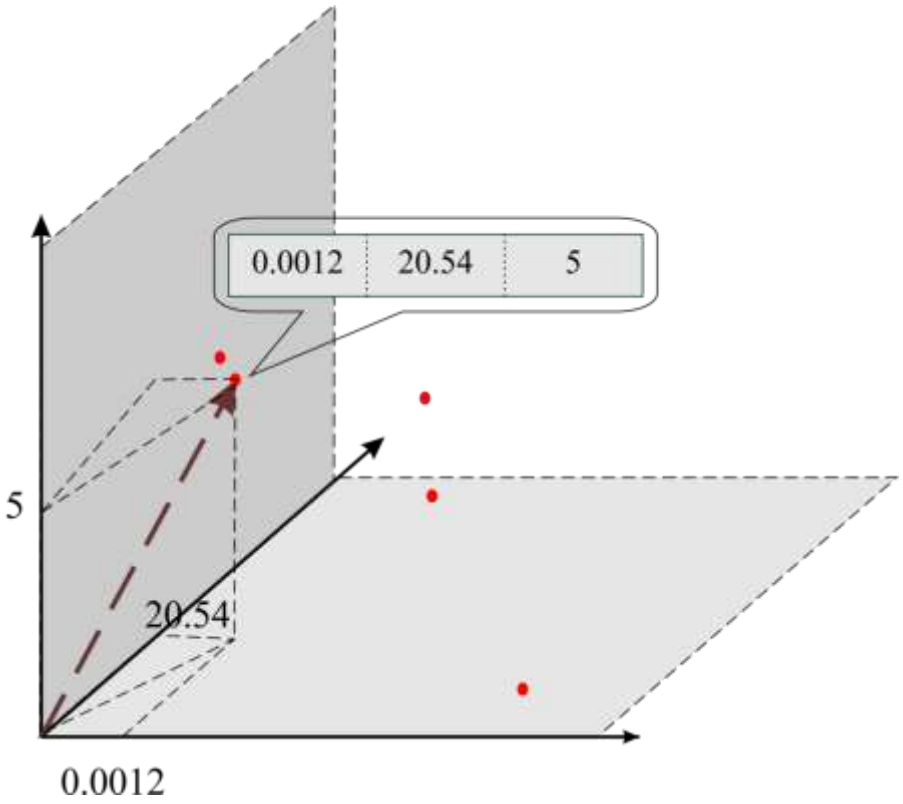


Decimal number string

It is a molecule representation that used for QSAR – **descriptor string**. Set of descriptors define chemical space.

There are more than 10 000 descriptors:

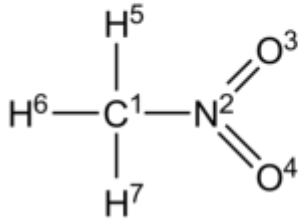
- Physicochemical descriptors
- Topological descriptors
- Fragment descriptors
- Pharmacophoric descriptors
- Constants of substituents
- Spatial (3D) descriptors
- Quantum chemical descriptors
- Molecular-mechanical descriptors
- Molecular fields descriptors
- Molecular similarity descriptors



Todeschini, R. and V. Consonni, *Molecular Descriptors for chemoinformatics*. 2009, Weinheim: Wiley-VCH. 967.

Tables

- Matrices (used for descriptor calculations)
- Connection tables (often used in databases)
- Cartesian coordinates and Z-matrices (for representation of 3D structures)



Atom list	
1	C
2	N
3	O
4	O
5	H
6	H
7	H

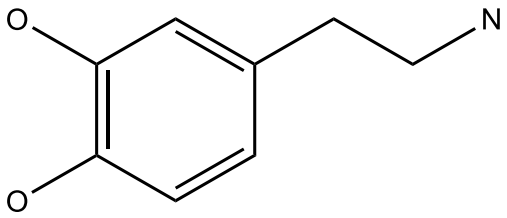
Bond list		
Atom 1	Atom 2	Bond order
1	2	1
2	3	2
2	4	2
1	5	1
1	6	1
1	7	1

+ Exhaustive and universal definition of the molecular structure
+ As concise as possible
+ Can be made unique (after canonicalization of atom numbering)
- The algorithms to handle connection tables are relatively slow
- Can hardly be adopted to be a field in database tables – requires special organization of data

- **Canonicalization of atom numbering (Morgan algorithm):** Morgan, H.L.. Journal of Chemical Documentation, 1965. **5**(2): p. 107-113.

Basic type of search in databases

Query



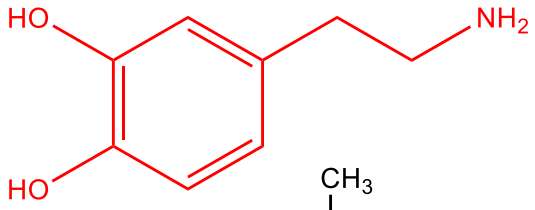
Structure search

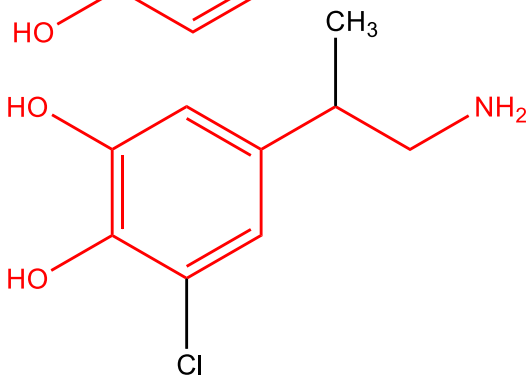
Substructure search

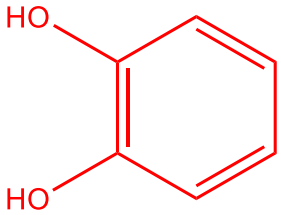
Superstructure search

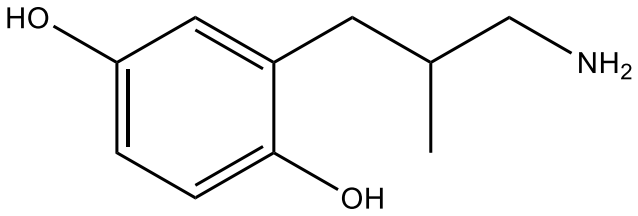
Similarity search

Retrieved structures







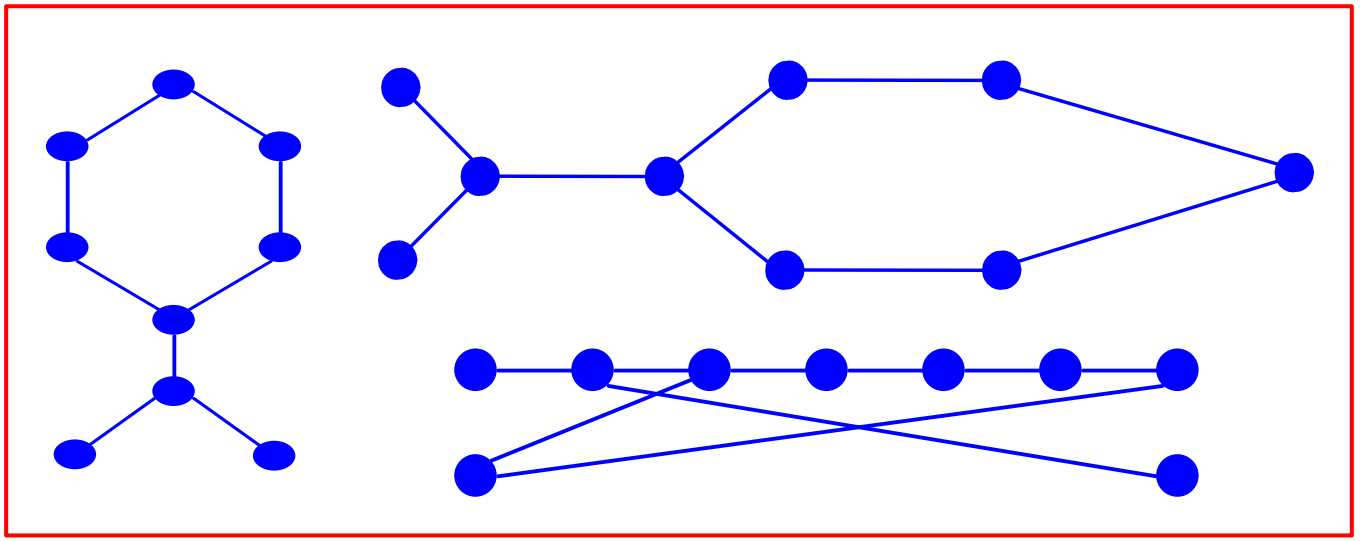


There is special types of search for Markush and reaction databases, special algorith used for them –
read our book (A. Varnek, I. Baskin, T. Madzhidov)

Molecule is a graph

- Graph is a set of nodes and edges
- Graphs are only about connectivity
  - spatial position of nodes is irrelevant
  - length of edges are irrelevant
  - crossing edges are irrelevant

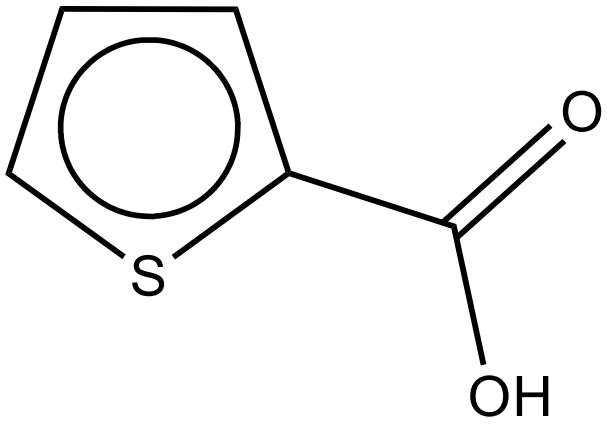
The same graph

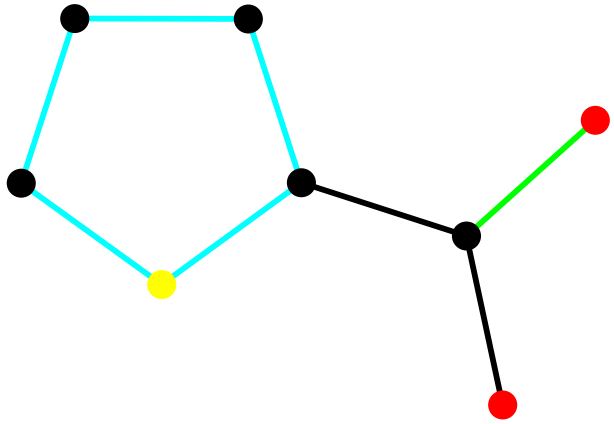


Molecule is a graph

- Vertices can be "colored" according to atom types
- Edges can be weighted according to bond type (or bond order)

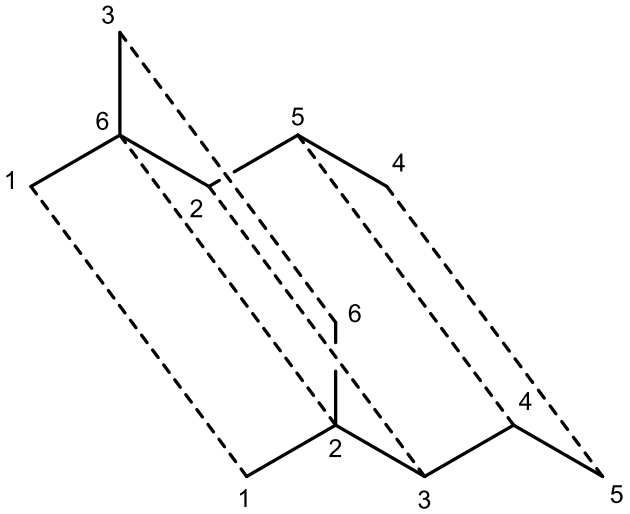
- Then well developed graph-handling algorithms of mathematics can be used in chemistry.





Structure search

Structure search – is the search of graph isomorphism



There is N! possible atoms renumeration

We need very efficient algorithms!

Query

Names, SMILES, InChI, connection table

Structure

Unique numbering algorithm

Canonicalized query

Only one comparison per structure!

Canonicalized structure

Structure search

1. Comparison of QUERY and molecule in database – **N** comparisons

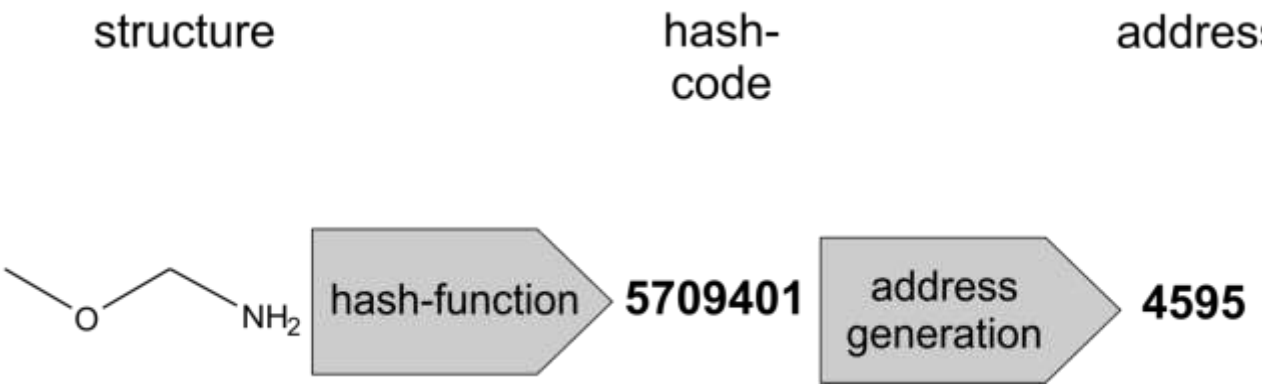
2. Bisection usage for search after sorting database – **log(N)** comparisons

Query: COCN

	Address	SMILES
1 no →	4591	CCCC
	4592	C(C)CC
	4593	CNCC
3 no →	4594	COCC
4 yes →	4595	COCN
	4596	CCCCC
2 no →	4597	COCCC

Solution there

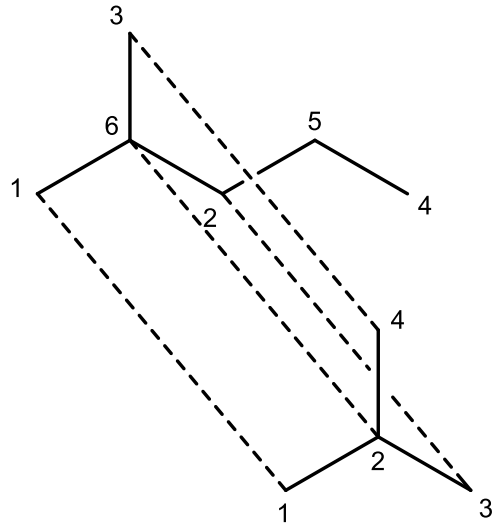
3. Hashing usage for search – only **1** comparison!



Hashing of chemical structures: Ihlenfeldt, W.D. and J. Gasteiger, *Hash codes for the identification and classification of molecular structure elements*. Journal of Computational Chemistry, 1994. **15**(8): p. 793-813.0

Substructure search

Substructure search – is the search of subgraph isomorphism



N atoms

M atoms

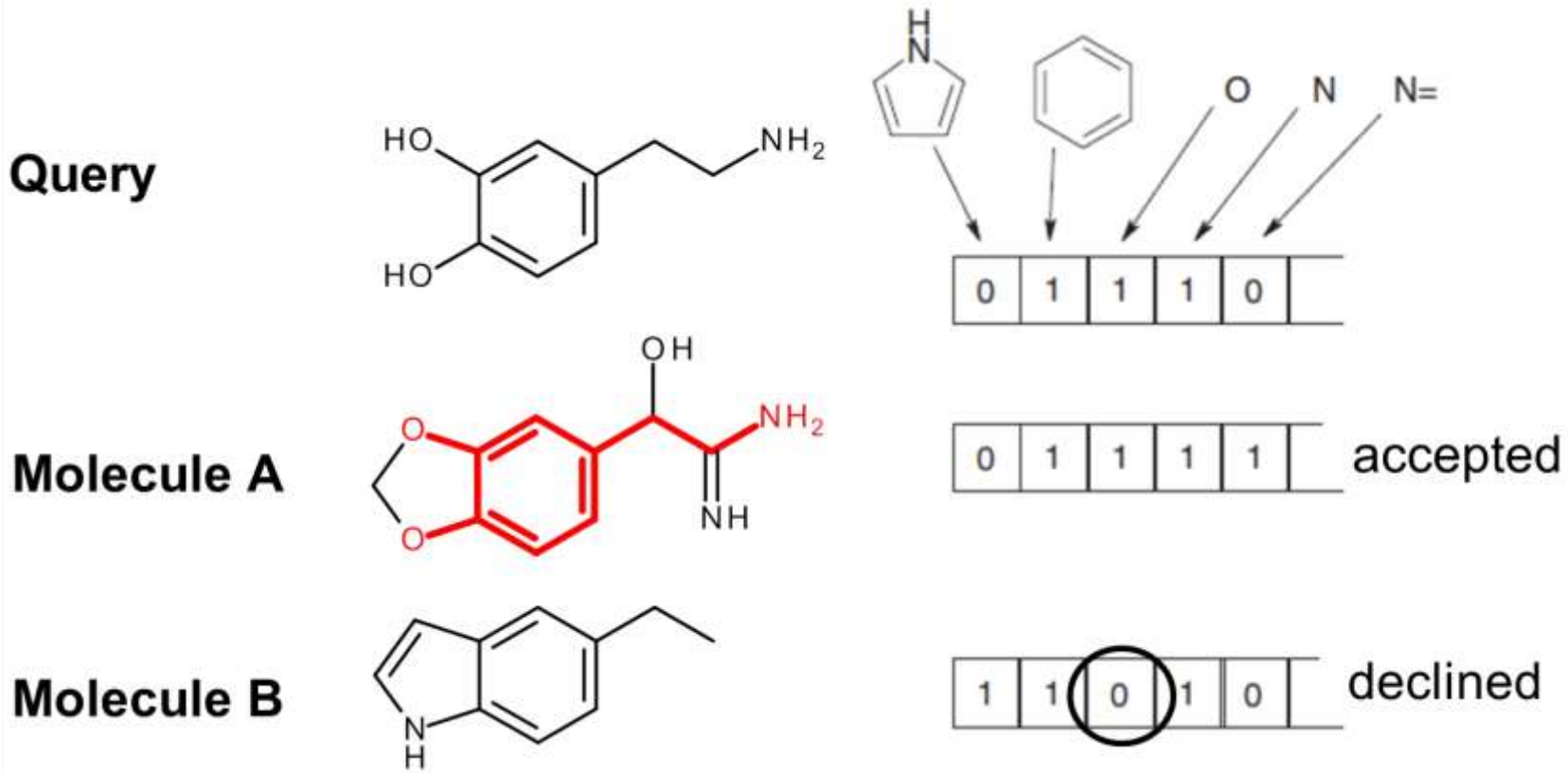
$$Number\ of\ mappings = \frac{N!}{(N-M)!}$$

2-step algorithm:

- Screening using bit string representation of molecules
- Subgraph isomorphism

Substructure search: screening

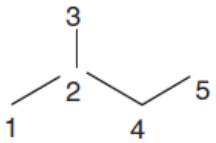
If graph of **query** is the subgraph of **molecule**, then all fragments of **query** MUST exist among fragments of **molecule.** All unities in the bitstring of query should match that of bitstring of accepted molecule



More than 90% of database should be declined for the second step

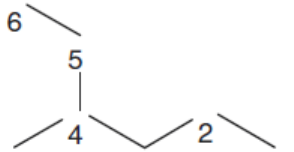
Substructure search: subgraph isomorphism

- The most popular algorithm is Ullmann's algorithm



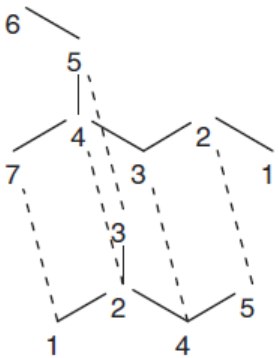
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

S



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

M



Matching matrix

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A

$A(AM)^T = S$

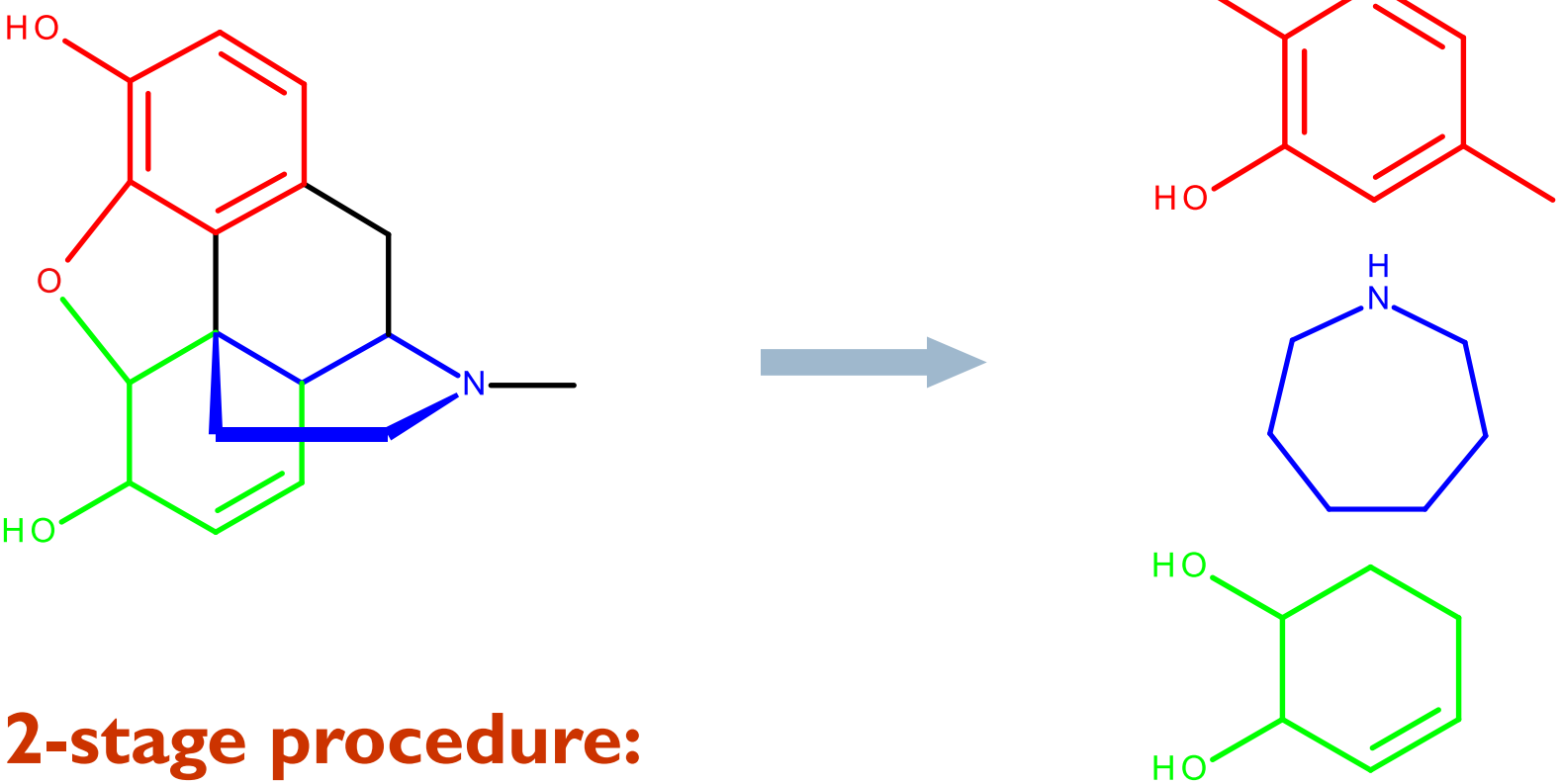
We look if matching matrix exist.

How **A** matrix can be found efficiently?

- Choose probe A matrix by heuristic rules on the basis of atom type and its connectivity,
- Look over possible probe matrices using back-tracking algorithm,
- Use *relaxation technique* – extend the information about vertex by iterative consideration neighboring atoms.

Ullmann J.R., *An Algorithm for Subgraph Isomorphism.* J.ACM, 1976. **23**(1): p. 31-42.

Superstructure search

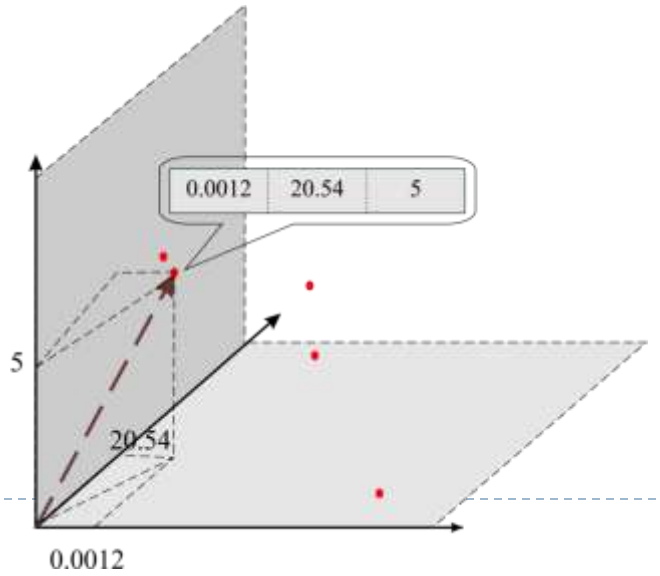


2-stage procedure:

- fingerprints fit
- graph theory algorithms

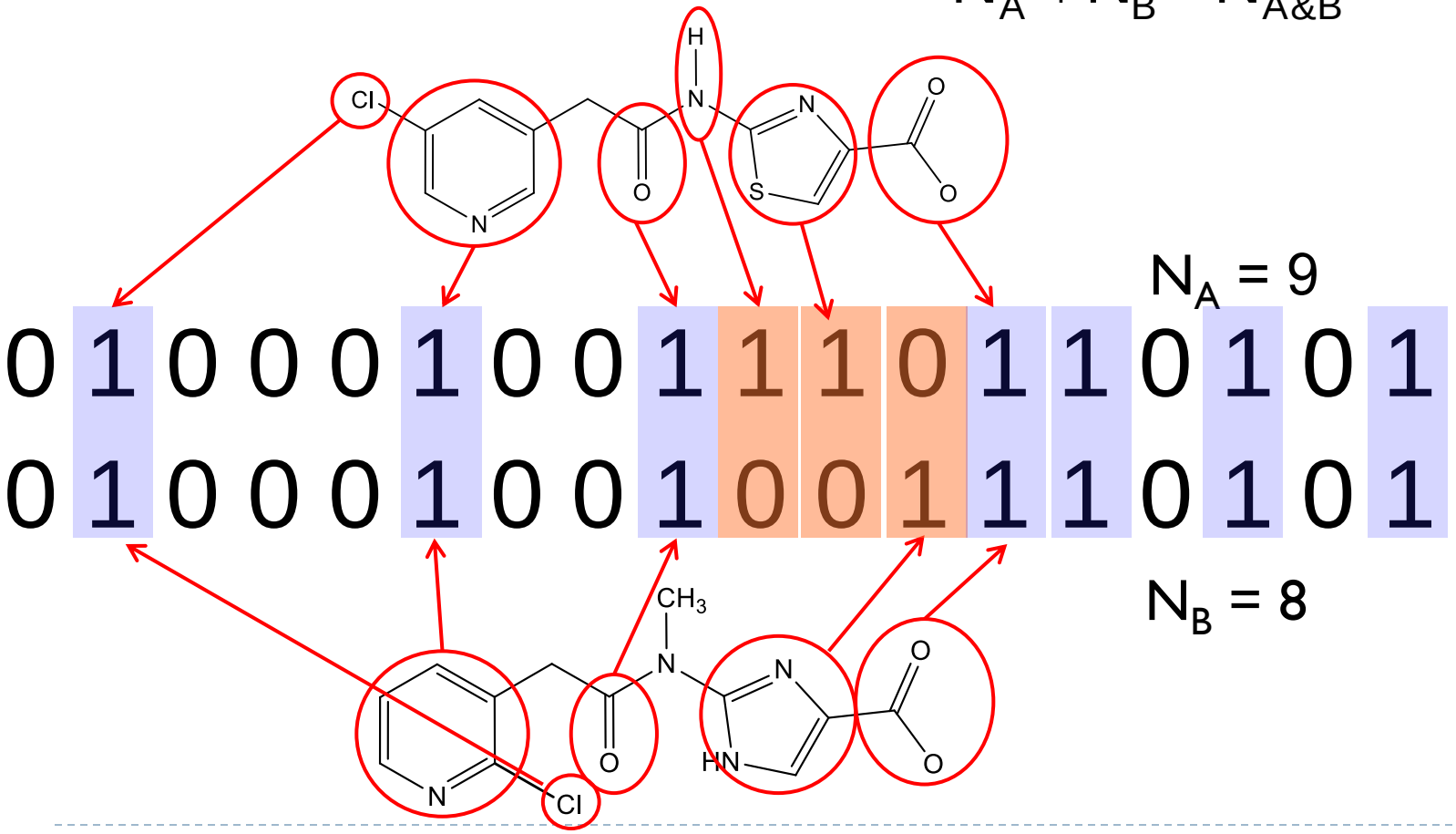
Similarity search

- Can be based on descriptor or bitstring representation of molecules.
- Similar molecules are the ones that are close in chemical space.
- There are number of different metrics of similarity. The most popular are Euclidian or Manhattan distance (for descriptor representation of molecules) or Tanimoto index (for bitstring)



Similarity search

Tanimoto (Jaccard) coefficient: $$Tc = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}$$



$N_A = 9$

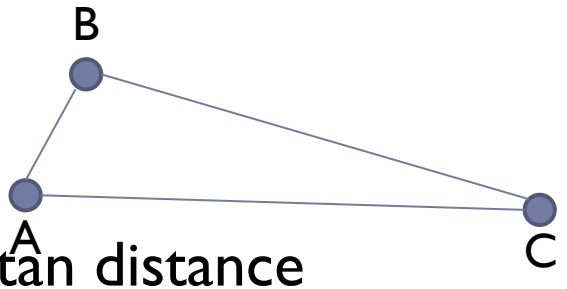
$N_{A\&B} = 7$

$N_B = 8$

$Tc = 0.7$

Effective similarity search

- In proposed formulation similarity search requires N comparisons. Complexity is O(N)
- However there are algorithms (*k-d tree, R tree, vp tree, BK tree etc.*) of similarity search whose complexity is **O(logN)**. They require that distance between molecules should be metric:
  - D (A,B) ≥ 0
  - D(A,A) = 0
  - D (A,B) = D (B,A)
  - Triangle rule: D (A,C) ≤ D (A,B) + D (B,C)



Metrics: Sörgel (1-Tc), Euclidian and Manhattan distance

- **k-d tree:** Bentley J.L., *Multidimensional binary search trees used for associative searching.* Commun. ACM, 1975. **18**(9): p. 509-517

CAS/SciFinder

- Consist from number databases:
  - Chemical Substances - CAS REGISTRY (>120 million organic and inorganic substances, >66 million sequences)
  - References – CAplus (>43 million records)
  - Reactions – CASREACT (>91.6 million reactions)
  - Regulated chemicals – CHEMLIST (>297,000 chemicals)
  - Chemical suppliers – CHEMCATS (millions commercially available products)
  - Chemical Industry Notes - CIN (>1.7 million records)
  - Markush – MARPAT (>1,134,000 Markush structures, >469,000 patent records)
- SciFinder is retrieval system

- Commercial
- The greatest database in the world
- Searchable
- Can't be downloaded in useful format

Reaxys

- Reaxys
  - >74.9 millions of compounds
  - >40.7 millions of reactions
  - >500 millions of facts (properties of compounds, reaction conditions, references etc)
  - >16,000 articles related to chemistry
- Reaxys Medicinal Chemistry
  - >6,2 millions of compounds
  - >30,5 millions of bioactivities
  - ~13,600 biological targets
  - >346,000 articles and >108,000 patents

- Commercial
- Reaxys Medicinal Chemistry database exist separately
- Searchable
- Data can be downloaded in useful format (SDF, RDF)

PubChem

- >91,7 millions of compounds
- >223 millions of substances
- >1,2 bioassays
- >2,2 millions of tested substances
- >231 millions data on bioactivity

- Non-Commercial and Open
- Biggest database of substances
- Searchable
- All data could be downloaded

Further reading