Chemical Databases: Encoding, Storage and Search of Chemical Structures

Dr. Timur I. Madzhidov
Kazan Federal University, Department of Organic Chemistry
Chemical Databases: Why?

73,000,000 organic and inorganic compounds in the largest database (~5-6 millions added each year)

How to search required information???

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

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)

110 m

How to represent compounds?

How to search?


Acetone
Benzene
DMSO
Aspirine
Pyridine
Clorohexidine
Hexane
Heptane
Nonane
Decane
Types of databases

- Literature Databases
  - Bibliographic
  - Full-text
  - Patent

- Factual Databases
  - Numeric
  - Metadata
  - Directory
  - Catalogs

- Structure Databases
  - Structure
  - Reaction
  - Patent

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

- **Bit string**
  - Not chemist-friendly
  - Computer-friendly
  - Loss of information
  - Capacious

- **Decimal number string**
  - More or less chemist-friendly
  - More or less computer-friendly
  - Detailed
  - Not capacious

- **Tables**
  - Detailed
  - Not capacious

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

Image is not unique

\[ C_4H_{10} \text{ - not univocal} \]
Linear notations

- **Hill’s formula**: $C_6H_6$, $C_6H_6O_3S$, $C_{10}H_8NO_2$
- **Chemical name**
  - **Trivial or trade name**: Vitamin B12, Cyanocobalamine
- **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@@]((N1[Co+]C#N)[C@@H]7/N=C(C\(C=3/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](n5nc4cc(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

InChI

Hill’s formula
Atom numbering follows Hill’s formula (H excluded)
Branching
Hydrogen localized on C1 atom
4 hydrogens delocalized on N6, N7, N8, N9, N10, O11 atoms

InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)/f/h8,10H,6H2

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.

![Diagram of bitstring representation of a chemical structure](image-url)
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

Tables

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

+ 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

Basic type of search in databases

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

Canonicalized structure

Only one comparison per structure!
1. Comparison of QUERY and molecule in database – $N$ comparisons

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

Query: COCN

3. Hashing usage for search – only 1 comparison!

Substructure search

Substructure search – is the search of subgraph isomorphism

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.

---

**Query**

```
\[
\begin{array}{c}
\text{HO} \\
\text{HO} \\
\text{NH}_2 \\
\end{array}
\]
```

**Molecule A**

```
\[
\begin{array}{c}
\text{HO} \\
\text{O} \\
\text{NH} \\
\text{NH}_2 \\
\end{array}
\]
```

---

**Molecule B**

```
\[
\begin{array}{c}
\text{NH} \\
\end{array}
\]
```

More than 90% of database should be declined for the second step.
The most popular algorithm is Ullmann’s algorithm

\[ A(AM)^T = S \]

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.

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: \( T_c = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}} \)

\[ \begin{array}{cccccccc}
0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{array} \]

\[ \begin{array}{cccccccc}
1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
\end{array} \]

\[ \begin{array}{cccccc}
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
\end{array} \]

\[ \begin{array}{cccc}
N_A = 10 \\
N_B = 9 \\
N_{A\&B} = 8 \\
T_c = 0.73 \\
\end{array} \]
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 \( O(\log N) \). They require that distance between molecules should be metric:
  - \( D(A,B) \geq 0 \)
  - \( D(A,A) = 0 \)
  - \( D(A,B) = D(B,A) \)
  - Triangle rule: \( D(A,C) \leq D(A,B) + D(B,C) \)

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

# Chemical databases: compounds

<table>
<thead>
<tr>
<th>Name</th>
<th>Content</th>
<th>No comp</th>
<th>Link</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>SciFinder</td>
<td>Compounds, Reactions, Markush, Commercially available, Regulated Chemicals, Bibliographic</td>
<td>73 mln</td>
<td><a href="https://scifinder.cas.org">https://scifinder.cas.org</a></td>
<td>Commercial, Download is limited, Self-annotated, Chemical search supported</td>
</tr>
<tr>
<td>Reaxys</td>
<td>Compounds, Reactions, Bibliographic</td>
<td>19 mln</td>
<td><a href="http://www.reaxys.com">http://www.reaxys.com</a></td>
<td>Commercial, Download is limited, Self-annotated, Chemical search supported</td>
</tr>
<tr>
<td>SPRESI</td>
<td>Compounds, Reactions</td>
<td>5.6 mln</td>
<td><a href="http://www.infochem.de/products/databases/spresi.shtml">http://www.infochem.de/products/databases/spresi.shtml</a></td>
<td>Commercial, Can be provided in SDF and RDF</td>
</tr>
<tr>
<td>ChemSpider</td>
<td>Compounds</td>
<td>28 mln</td>
<td><a href="http://www.chemspider.com">http://www.chemspider.com</a></td>
<td>Non-commercial, Downloadable, Chemical search supported</td>
</tr>
<tr>
<td>CWM</td>
<td>Compounds</td>
<td>-</td>
<td><a href="http://www.akosgmbh.de/globalsearch/index.htm">http://www.akosgmbh.de/globalsearch/index.htm</a></td>
<td>Commercial, Metabase, Chemical search supported</td>
</tr>
<tr>
<td>SCRIPDB</td>
<td>Compounds from patents</td>
<td>-</td>
<td><a href="http://dcv.uhnres.utoronto.ca/SCRIPDB">http://dcv.uhnres.utoronto.ca/SCRIPDB</a></td>
<td>Non-commercial, Download of retrieved structures only, Patent download, Chemical search supported</td>
</tr>
<tr>
<td>Ambinter</td>
<td>Commercially available</td>
<td>18 mln</td>
<td><a href="http://www.ambinter.com">http://www.ambinter.com</a></td>
<td>Non-commercial, Downloadable, Chemical search supported</td>
</tr>
<tr>
<td>BioinfoDB</td>
<td>Commercially available</td>
<td>?</td>
<td><a href="http://cheminfo.u-strasbg.fr">http://cheminfo.u-strasbg.fr</a></td>
<td>Non-commercial, Downloadable, Chemical search supported</td>
</tr>
<tr>
<td>ChemExper</td>
<td>Commercially available</td>
<td>?</td>
<td><a href="http://www.chemexper.com">http://www.chemexper.com</a></td>
<td>Chemical search supported, Download is limited</td>
</tr>
<tr>
<td>ZINC</td>
<td>- Commercially available</td>
<td>17 mln</td>
<td><a href="http://zinc.docking.org">http://zinc.docking.org</a></td>
<td>Non-commercial, Downloadable, Chemical search supported, Prepared for virtual screening</td>
</tr>
<tr>
<td>CoCoCo</td>
<td>- Commercially available</td>
<td>7 mln</td>
<td><a href="http://cococo.unibo.it/cococo">http://cococo.unibo.it/cococo</a></td>
<td>Non-commercial, Downloadable, Chemical search supported, Prepared for virtual screening</td>
</tr>
</tbody>
</table>
Consist from number databases:

- Chemical Substances - CAS REGISTRY (>73 million organic and inorganic substances, >64 million sequences)
- References – CAplus (>37 million records)
- Reactions – CASREACT (>67.1 million reactions)
- Regulated chemicals – CHEMLIST (>297,000 chemicals)
- Chemical suppliers – CHEMCATS (>68 million commercially available products, >21 million unique compounds)
- Chemical Industry Notes - CIN (>1.7 million records)
- Markush – MARPAT (>992,000 Markush structures, >408,000 patent records)

SciFinder is retrieval system

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