Current strategies for searching through structure and chemical compound databases

Grzegorz FIC*, Mariusz SKOMRA, Barbara DĘBSKA – Faculty of Chemistry, Rzeszow University of Technology, Rzeszów, Poland

Please cite as: CHEMIK 2016, **70**, 8, 410–418

1. Introduction

Modern databases which contain chemical compound structures are characterized by a large increase of information. Table 1 shows the increase of data in Pubchem database over the last years. This base originated in 2004 and is managed by the National Center for Biotechnology Information (NCBI) at the US National Institutes of Health (NIH). Currently, it has the largest free of charge dataset of chemical structures in the world. Pubchem consists of three bases which contain information about small molecules (less than 1000 atoms and bonds). PubChem Substance contains information about substances (such as mixtures, extracts, and complex compounds) from many other databases, PubChem Compound contains information about chemical structures in PubChem Substance, and PubChem BioAssay contains information about screening results for bioactivity (sets of tested substances are between one and several hundred thousand).

A majority of large chemical databases contain information compiled from other datasets. For example, PubChem Substance contains information from almost 400 databases. Some of them provide millions of records (e.g., Aurora Fine Chemical LLC has over 33 million records, while ZINC has 25.7 million), other (like the ones held by laboratories or small research groups) – only a few or even only one record. It is worth noting that these bases are compatible with each other, what enables the development of large searching systems. For example, the Entrez system [1] provides resources from 30 chemical, biological, and related databases. All these databases can by searched by formulating and entering only one search query using the global interface of Entrez.

Table 1

Increase of PubChem database in years 2007–2016 [mln]. Values were obtained by the use of searching query: all[filt]

	29.10.07	16.09.08	11.11.09	19.02.10	18.03.13	6.11.14	9.02.16
PubChem Substance	21.3	44.6	61.2	62.5	116.8	178.4	216.9
PubChem Compound	18.1	19.3	25.7	26.1	46.7	62.0	87.2
PubChem BioAssay	0.000667	0.001188	0.001917	0.002153	0.649147	1.112105	1.154427

The aforementioned features of chemical databases, foremost the large amount of gathered data, impose the necessity of developing sophisticated tools for searching and recovering information. In next sections, we discuss selected issues related to the creation of searching queries (i.e., sets of conditions which information must fulfil) and strategies of searching through databases. We pay special attention to the newest solutions in this area, such as strategies based on fuzzy substructures, 2D and 3D chemical similarity, ontology strategies, and structure-properties similarities.

Corresponding author:
Grzegorz FIC – PhD., (Eng.), e-mail: gfic@prz.edu.pl

2. Strategies for searching through chemical compound databases

Most generally, strategies which are currently used can be divided into:

1. **bibliographical**, which use fields like: name, author, producer, properties etc.
2. **structural, which use fields containing structural information** such as SMILES or MOL. In this category, several methods can be distinguished:
   - searching for structures which are identical with the one defined in the query (*Exact*), including tautomers (*All Tautomers* in ChemSpider database),
   - searching for structures containing the substructure defined in the query (*Substructure*),
   - searching for substructures of the structure defined in the query (*Superstructure*) – this strategy is useful during computer assisted organic synthesis (search for *building blocks*),
   - searching for similar structures which fulfill user-defined conditions of structural similarity; in conjunction with *Substructure* strategy it allows to create sets of structures for screening tests (for example in the process of drug design),
   - searching for precursors – structures from which the defined substructure can be synthetized (for example the *Precursor* strategy in BioPath database of metabolic transformations),
   - searching for all possible structures which contain the defined structural skeleton (for example *Flex*, *Flexplus* in ChemIDplus, *Same Skeleton* in ChemSpider),
   - searching for all isomers (for example *All Isomers* in ChemSpider).
3. **hybrid**, where the query is composed of conditions of different type (structural, textual, numerical, and logical) which define the needed and/or forbidden properties of substances
4. **ontological**, which are used mainly in the chemical and biological databases, require some ontologies, and allow to search and browse bases by terms which characterize individual substances
5. **structure-properties type**, searching for compounds which have similar properties to these of the defined structure/substructure

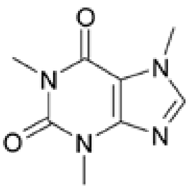
3. Methods of entering structural information

Information about chemical structure, which is essential for searching queries, can be entered in several ways using:

- **text identifiers of the structure** – these are strings which can represent chemical structures in queries, such as systematic name (IUPAC), vernacular name, id number (e.g., *CAS Registry Number*, *EINECS Number*), linear code SMILES [2], InChI vectors (*IUPAC*

International Chemical Identifier) [3], or InChIKey (hash version of InChI) [4]. InChI is made up of several layers of information separated by the sign "/" and a prefix (Fig. 1): 1S denotes the InChI version, chemical formula comes next, "/c" denotes the beginning of the atom's connection table, "/h" – sublayer which describe hydrogen atoms connected with other atoms. There can be more sublayers related to charge or stereochemical and isotopic properties. InChIKey (Fig. 1) is a vector with constant length of 27 characters containing information unreadable for humans. It is an unambiguous structure representation (but there are some exceptions [5]). It is created in a hash-process of InChI information by SHA-256 algorithm. It consists of three blocks, separated by dashes. First of them is characteristic for compounds containing a defined structural skeleton, while the second differentiate isomers. They are used during searching through structure databases. SMILES is a result of the linearization of the chemical structure process, i.e., of "cutting" one bond in every ring. The beginning and the end of the chain are represented by a pair of identical numbers after atoms' symbols (in Fig. 1. these are N1 and C1 atoms and C2 and N2 atoms).

- **graphical structure information** – drawn structure/substructure in a coded form (such as SMILES) is send to the database server (search engine) by a special software which can have a form of an Internet browser plug-in or of an applet which is send by the server to the user's computer
- **file structural information** – uploaded file which contains information about structure in one of the available formats (such as mol, cml, smi, pdb). File can be saved on a computer or in the Internet
- **file graphical information** – uploaded file which contains the image of the structure (such as a handmade picture or scan of figure from an article) saved in one of the formats accepted by the search engine (such as gif, jpg, tif, pdf); this technology is currently available in ChemSpider.com and Chemical Structure Lookup Service 2008.



1,3,7-trimethylpurine-2,6-dione

InChI=1S/C8H10N4O2/c1–10–4–9–6–5(10)7(13)12(3)8(14)11(6)2/h4H,1–3H3

RYYVLZVUVIJVGH-UHFFFAOYSA-N

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

58-08-2

200-362-1

Fig. 1. Selected caffeine text identifiers. From above: IUPAC name, InChI, InChIKey, SMILES, CAS Registry Number, EINECS Number

4. Fuzzy substrcutures

The so-called "substructure" is an important element for many searching queries. Substructures are fragments of chemical structures. They can be defined in a "sharp" way (when every attribute which describes them has exactly one value) or in a "fuzzy" way (Fig. 2). Markush structures, which include varying fragments, are ancestors of fuzzy substructures. The idea of fuzzy substructures was introduced in the purpose of providing consistent representation of classes and groups of compounds with similar structures. In fuzzy substructures [6] the attributes which describe atoms' properties (such as type, number of neighbors, number of free electrons, position in a ring, aromatic system, aliphatic chain etc.) and/or bonds (bond type, position), can have more than one value. Every attribute has a set of allowed and forbidden values. Therefore, every fuzzy substructure represents not one, but many substructures. Information about fuzziness is entered into the search engine's interface by a special form or by SMARTS notation (Fig. 3) [7], which belongs to SMILES notation family.

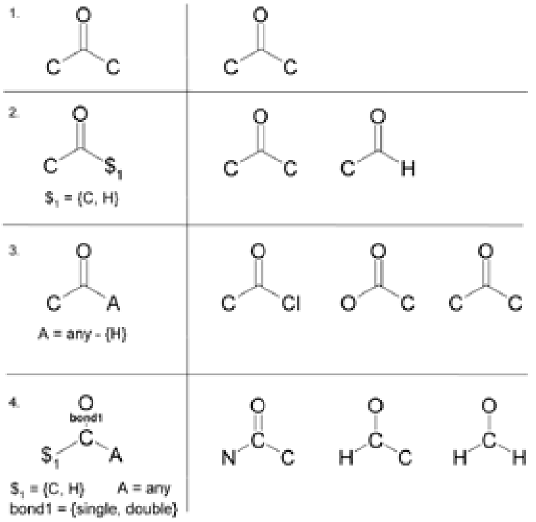


Fig. 2. Examples of fuzzy substructures. On the left – definition of fuzzy structure, on the right – selected substructures which fulfil this definition (no. 1: sharp substructure, without fuzziness)

[n;R]&*C(=O)O	Searching for all molecules with contain a carboxylic group and a nitrogen atom in some aromatic ring
[OH]c1ccccc1	Searching for all molecules which contain a benzene ring with a hydroxyl group
(C(=O)O).(OCC)>>C(=O)OCC.O	Searching for an intermolecular esterification reaction
(C(=O)O.OCC)>>C(=O)OCC.O	Searching for an intramolecular esterification reaction

Fig.3. Examples of searching queries in SMARTS notation (extension of SMILES notation which allow for searching with fuzzy substructures approach)

5. Chemical structure similarity

Different computer methods for evaluation of similarity between chemical structures have been proposed. Every one of them needs a proper structure representation method and a defined chemical similarity measure.

Structure representation methods. They consider only constitutional structure (2D chemical similarity) or full spatial structure (3D chemical similarity). There are three main types of 2D representation which are used during the evaluation of similarities:

- *fingerprints* – binary vectors with different lengths, where every element confirms (value 1) or excludes (value 0) the presence of some property in the compound. Many different *fingerprints* have been described. They differ by: (i) number of bits, (ii) meaning of bits and (iii) the selection of elements which describe the structure in 3D. The most common are PubChem fingerprint (881 structure features) [8], FP3 and FP4 (in OpenBabel, they accommodate 55/307 substructure types) [9], MACCS [10] – which has different forms: 166/322-bit *fingerprint* and 166/322-elements number descriptor (the elements of vector are numbers which define the amount of structural features that are present in the compound structure). A major disadvantage of this type of *fingerprint* is the impossibility of determination of all compound structural features. In effect, results of comparison are not fully objective. The degree of similarity depends on the used *fingerprint* – it is possible that two structures are very similar based on one *fingerprint* and dissimilar based on another. Hash *fingerprints* attempt to eliminate this disadvantage. In this case, the set of features is not defined in advance. The complete structure fragmentation process creates the set of all possible substructures, from one to n- atomic where n denotes the number of atoms in the molecule. Afterwards, the hash algorithm is used and a binary *fingerprint* is generated (it usually has length of 1024 or 512 bits). *Fingerprint* has always the same number of elements, independently of the number of substructures. As a result of hash process some bits can represent more than one structure element (so called "bits' collision"). That situation is allowed but rather inadvisable. The most commonly used hash *fingerprints* are: FP2 (OpenBabel) where structure fragmentation is limited to fragments containing 1–7 atoms, ECFP [11] – group of new generation hash *fingerprints* where the initial range of features is expanded and takes into account the influence of neighbors in another layers (which are measured by the number of bonds from initial state).

- molecular descriptors [12] – numerical values, each of which represents quantitatively one or more of the specified structural features of the molecule
- molecular eigenvalues – one numerical value represents the whole chemical structure, for example BCUT descriptors [13].

Similarity measures. They describe in a quantitative way the similarity between two chemical structures. Some of them, which are used to evaluate similarity by *fingerprints*, are shown in Figure 4. Tanimoto similarity (TS): *a* is the amount of [1,0] bit pairs (i.e., the amount of features represented in the *fingerprint* of one structure, but absent in the second structure), *b* – amount of [0,1] bit pairs, *c* – amount of [1,1] bit pairs. One needs to mention here that value TS=1 does not indicate that the compared structures are identical, but only states that they have the same *fingerprints*. TS values above 0.85 indicate high similarity of two chemical structures. It is very likely that such structure have, for example, similar biological activity. The asymmetrical index of Tversky (TI) is a modified version of TS where α and β are weight coefficients which allow to consider asymmetry. Another measure is the Euclidean distance (ED), where a_i and b_i are the *i*-values of fingerprints elements for two structures A and B. The greater the ED value is, the smaller is the similarity between structures. ED is used as a measure mostly in methods which are based on descriptors. Another chemical similarity measures, different from the ones described above, are also known [14].

$$\mathrm{TS}=\frac{c}{a+b+c}\qquad\qquad \mathrm{TI}=\frac{c}{\alpha\cdot a+\beta\cdot b+c}$$

$$\mathrm{ST}=\frac{V_{AB}}{V_{AA}+V_{BB}-V_{AB}}\qquad\qquad \mathrm{CT}=\frac{\sum V_{AB}}{\sum V_{AA}+\sum V_{BB}-\sum V_{AB}}$$

$$\mathrm{ED}=\sqrt{\sum\left(a_i-b_i\right)^2}$$

$$\mathrm{T_3S}=\frac{\sum_{i=1}^{N}X_{iA}X_{iB}}{\sum_{i=1}^{N}X_{iA}{}^2+\sum_{i=1}^{N}X_{iB}{}^2-\sum_{i=1}^{N}X_{iA}X_{iB}}$$

Fig. 4. Definitions of selected chemical similarity measures which are used in text

3D chemical similarity. In order to measure the 3D chemical similarity, it is necessary to know the position of atoms in 3D coordinates for different conformations of compared molecules. Among all methods of measuring 3D similarity, it is especially worth to mention the following:

- use of descriptors related to distances and angles in 3D (such as bonds angles, distances between atoms),
- evaluation of molecular field similarity (electrostatic potential field, geometrical shape field, electron density and other), for example CoMFA (*Comparative Molecular Field Analysis*) or CoMSIA (*Comparative Molecular Similarity Index Analysis*),
- molecular moments comparison, for example CoMMA (*Comparative Molecular Moments Analysis*) which include descriptors between molecular moments in regard to mass center, charge center, dipol center,
- use of descriptors which are based on molecule shapes, for example van der Waals volume, Taft spherical parameter, STERIMOL parameters (which allow a quantitative description of substituent groups), WHIM (*Weighted Holistic Invariant Molecular*) descriptors.

Conformers' 3D similarity measures. Special measures were made to assess the quantitative 3D similarity [15]. One of them is Tanimoto similarity in the modified version for molecular fields (T_3S in Fig. 4): X_{iA} and X_{iB} are values of some attribute of compared molecules A and B in the *i*-th element of the field. PubChem database uses two measures (Fig. 4) [16]. Shape Tanimoto (ST) is a 3D shape similarity measure: V_{AA}, V_{BB} are volumes of conformer's fragments for molecules A and B which are absent in their common superposition, V_{AB} is the common volume of A i B in their superposition. Color Tanimoto (CT) is a compatibility measure for six structural features: hydrogen bonds donors and acceptors, cations, anions, hydrophobicity, rings. CT is a sum over features: V_{AB} – volume of fragments which are in accordance for some feature, V_{AA} and V_{BB} – volume of fragments on which given features are different. ST and CT are determined in a two-stage process: (i) when the 3D superposition for two conformers such that their common part (V_{AB}) is maximal is created and ST is determined, (ii) where CT is determined in every point of this superposition. PubChem algorithm supposes that two conformers are similar when ST ≥ 0.8 and CT ≥ 0.5.

6. Structure-properties strategies

In addition to methods which assess chemical similarity based on analysis and comparison of their structures, methods where similarity is based on the properties were also proposed. According to this framework, two molecules can be similar even when their structures are different, and similarity, assessed by classical methods, is low. This is of particular importance in the creation of sets with high chemical diversity, but with similar properties, such as biological activity. An example of this method is LASSO (*Ligand Activity in Surface Similarity Order*) method [17]: two ligands have similar activity when their surfaces have similar properties (compounds with similar surfaces bonds with the same proteins). For every molecule in the database the LASSO descriptor is evaluated – this is a vector in which every element is a number of points on compound surface with specified properties. There are 23 distinguished surface points, such as hydrophobic places, places with π effects, places with hydrogen bonds donors etc. Two molecules are similar when their LASSO descriptors are identical or highly similar. Fast selection of compounds which may have a specified biological activity can be provided using neural networks (authors say that the scan through 1 mln compounds takes less than 1 minute). LASSO method is implemented, for example, in the chemical search engine ChemSpider.com.

7. Ontological strategies

As an effect of dynamic development of different fields of science, scientist from whole world create their own nomenclature for describing their new discoveries. As a result, scientist from different parts of the world have problems with comparison and analysis of the research results. Creation of ontologies has been proposed as a solution to this problem. Ontology is set of terms which describe some field of science. Its key task is to ensure the unambiguity in descriptions. In order to do that ontologies use categorization and hierarchization. ChEBI (*Chemical Entities of Biological Interest*) [18] ontology, which is a part of bioactive compounds database ChEMBL, can be show as an example. ChEBI ontology (Fig. 5) is using three subontologies: (i) chemical entities which classify compounds based on their composition and structure (e.g., hydrocarbons, carboxylic acids, amins), (ii) roles which classify compounds based on their biological functions (e.g., antibiotics, coenzymes, hormones), on the usage by people (e.g., pesticides, drugs, fuels), or on their chemical role (e.g., acceptor, donor, solvent, ligand) and (iii) subatomic particles which classifies elements smaller than atoms (e.g., electrons, photons, nucleons).

CHEBI:24431 chemical entity – CHEBI:23367 molecular entity – CHEBI:33579 main group molecular entity – CHEBI:33675 p-block molecular entity – CHEBI:33582 carbon group molecular entity – CHEBI:50860 organic molecular entity – CHEBI:33285 heteroorganic entity – CHEBI:36962 organochalcogen compound – CHEBI:36963 organooxygen compound – CHEBI:37622 carboxamide – CHEBI:29347 monocarboxylic acid amide – CHEBI:22645 arenecarboxamide – CHEBI:22702 benzamides – CHEBI:7496 nelfinavir

Fig. 5. Selected branch from ChEBI ontology tree of nelfinavir –HIV protease inhibitor. Full tree contains c.a. 50 branches: http://www.ebi.ac.uk/chebi/chebiOntology.do?chebiId=CHEBI:7496&treeView=true#vizualisation (17.02.2016)

8. Summary

Some of the strategies described here are more and more often used by chemists (for example strategies based on chemical similarity or fuzzy substructures), another have only specific usage (for example *Superstructure* and *Precursors* strategies which are used in computer assisted organic synthesis [19]).

It seems that one of the most important current applications of different tools for searching in chemical compounds databases is the creation of virtual libraries with compounds structures, which are necessary for virtual screening in the process of designing chemical compounds (such as drugs) with desirable properties. This is an effect of the development of cheminformatical methods, notably algorithms for evaluation of chemical similarity and strategies for searching through large databases. Clusterization, in which structures are grouped in clusters which contain similar structures, is a very useful supporting tool for this type of work. Clusterization can be done based on 2D and 3D similarity (both are available in PubChem).

Ontological strategies can rapidly search for compounds with similar chemical or biological properties. This tool is not yet popular and has been implemented in only a few search engines. PubChem is an exception in this area, as it provides several different ontologies (including ChEBI described earlier).

Future development of structure-properties strategies seems very promising. Even currently, that kind of strategy allows (using ChemSpider.com database) to find molecules with large bioactivity for specified protein and small for another protein group.

Literature

1. Entrez, the Life Sciences Search Engine; http://www.ncbi.nlm.nih.gov/sites/gquery.
2. SMILES – A Simplified Chemical Language, Daylight Chemical Information Systems, Inc. 2011, http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.
3. The IUPAC International Chemical Identifier, IUPAC, 2012, http://www.iupac.org/home/publications/e-resources/inchi.html; Heller S., McNaught A., Stein S., Tchekhovskoi D., Pletnev I. InChI – the worldwide chemical structure identifier standard, J. Cheminform. 2013, 5, 7, http://jcheminf.springeropen.com/articles/10.1186/1758–2946–5-7.
4. Heller S. R., McNaught A., Pletnev I., Stein S., Tchekhovskoi D.: InChI, the IUPAC International Chemical Identifier, J. Cheminform. 2015, 7, 23, http://jcheminf.springeropen.com/articles/10.1186/s13321–015–0068–4.
5. Williams A.: An InChIkey Collision is Discovered and NOT Based on Stereochemistry, 2011, http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/.
6. Hippe Z. S., Fic G., Nowak G.: Representation of Common Sense in Chemical Syntheses by Means of Molecular Graphs, Found. Comput. Decision Sci. 1994, 19, 21–30; Dębska B., Guzowska-Świder B.: Fuzzy Definition of Molecular Fragments in Chemical Structures, J. Chem. Inf. Comput Sci., 2000, 40, 325–329.
7. SMARTS – A Language for Describing Molecular Patterns, Daylight Chemical Information Systems, Inc. 2011, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
8. specification: ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf.
9. Open Babel v2.3.0 documentation, http://openbabel.org/docs/dev/.
10. MayaChemTools: MACCS, 2015, http://www.mayachemtools.org/docs/modules/html/MACCSKeys.html.
11. Rogers D., Hahn M.: Extended-Connectivity Fingerprints, J. Chem. Inf. Model. 2010, 50(5), 742–754.
12. Todeschini R., Consonni V.: Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.
13. Pearlman R. S., Smith K. M.: Metric Validation and the Receptor-Relevant Subspace Concept, J. Chem. Inf. Comput. Sci. 1999, 39 (1), 28–35.
14. www.daylight.com/dayhtml/doc/theory/theory.finger.html.
15. Todeschini R., Consonni V.: Handbook of Molecular Descriptors, John Wiley & Sons, 2008.
16. Bolton E. E., Kim S., Bryant S. H.: PubChem3D: Similar conformers. J. Cheminform. 2011, 3, 13, http://www.jcheminf.com/content/3/1/13.
17. Reid D., Sadjad B. S., Zsoldos Z., Simon A.: LASSO-ligand activity by surface similarity order: a new tool for ligand based virtual screening, J. Comput. Aided Mol. Des. 2008, 22(6–7),479–87.
18. Hastings J., de Matos P., Dekker A., Ennis M., Harsha B., Kale N.: Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013, Nucleic Acids Res. 2013, 41(Database issue):D456–63, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531142/.
19. Nowak G., Fic G.: Komputerowe projektowanie wieloetapowych syntez organicznych – stan aktualny i kierunki rozwoju, Przem. Chem. 2011, 90(1), 78–83.

(Received – 14.07.2016 r.)

*Grzegorz FIC – PhD., (Eng.), a graduate of Rzeszów University of Technology (chemical technology); Ph.D. in chemistry (Jagiellonian University). Senior lecturer in Rzeszów University of Technology (Faculty of Chemistry, Department of Biotechnology and Bioinformatics). Author of 77 scientific publications, information systems and computer programs for chemical sciences and related. Research interests – CAOS (computer assisted organic synthesis), computer processing of structural information, bioinformatics.

e-mail: gfic@prz.edu.pl, phone: +48 17–865–1838

Mariusz SKOMRA – M.Sc., (Eng.), a graduate of Rzeszów University of Technology (biotechnology). Ph.D. student at Rzeszów University of Technology (Department of Chemistry). Research interests – CAOS (computer assisted organic synthesis), computer processing of structural information.

e-mail: d180@stud.prz.edu.pl, phone: +48 17–865–1838

Barbara DĘBSKA – Ph.D., D.Sc., (Eng.), a graduate of WSI in Rzeszow (School of Engineering); Ph.D., DSc. in technical sciences in chemical technology and chemical informatics. Associate Professor at Rzeszów University of Technology (Department of Biotechnology and Bioinformatics). She was awarded the Knight's Cross of the Order of the Rebirth of Polish. The author of 187 scientific publications. Research interests – chemometrics, intelligent systems for chemical sciences and related, e-learning.

e-mail: bjdebska@prz.edu.pl, phone: +48 17–865–1358