I. 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 be searched by formulating and entering only one search query using the global interface of Entrez.

<table>
<thead>
<tr>
<th>Increase of PubChem database in years 2007–2016 [mln]. Values were obtained by the use of searching query: all[title]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>PubChem Substance</td>
</tr>
<tr>
<td>PubChem Compound</td>
</tr>
<tr>
<td>PubChem BioAssay</td>
</tr>
</tbody>
</table>

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 fulfill) 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.

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 skeletal (for example Flex, Flexplus in ChemIDplus, Same Skeleton in ChemSpider),
   - searching for all isomers (for example All Isomers in ChemSpider).

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...
4. Fuzzy substrcutures

notation (Fig. 3) [7], which belongs to SMILES notation family. Information about fuzziness is entered for forbidden values. Therefore, every fuzzy substructure represents not more than one value. Every attribute has a set of allowed and forbidden values. 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 by SHA-256 algorithm. It consists of three blocks, separated by dashes. The so-called “substructure” is an important element for many 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.

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 substrcutures. The idea of fuzzy substrcutures was introduced in the purpose of providing consistent representation of classes and groups of compounds with similar structures. In fuzzy substrcutures [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 substrcuture represents not one, but many substrcutures. 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.
coefficients which allow to consider asymmetry. Another measure is the Tversky (TI) is a modified version of TS where they have, for example, similar biological activity. The asymmetrical index of two chemical structures. It is very likely that such structure amount of features represented in the Tanimoto similarity (TS): are used to evaluate similarity by that they have the same bit pairs. One needs to mention here that value TS=1 does – 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].

\[
TS = \frac{c}{a + b + c} \quad TI = \frac{c}{\alpha \cdot a + \beta \cdot b + c}
\]

\[
ST = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}} \quad CT = \sum_{i} V_{Ai} + \sum_{j} V_{Bj} - \sum_{i,j} V_{ij}
\]

\[
ED = \sqrt{\sum (a_i - b_i)^2}
\]

\[
TS = \frac{\sum_{i,j} x_{i\alpha} x_{j\beta}}{\sum_{i,j} x_{i\alpha}^2 + \sum_{i,j} x_{j\beta}^2 - \sum_{i,j} x_{i\alpha} x_{j\beta}}
\]

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 (TS in Fig. 4): X_{iα} and X_{iβ} 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_{AB}, V_{AA}, V_{BB}, V_{AB} 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, amines), (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).


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 substrutures), 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 cheminfomatical 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**


(Received – 14.07.2016)