Graph mining: procedure, application to drug discovery and recent advances

Ichigaku Takigawa and Hiroshi Mamitsuka

Bioinformatics Center, Institute for Chemical Research Kyoto University, Gokasho, Uji 6110011, Japan

Combinatorial chemistry has generated chemical libraries and databases with a huge number of chemical compounds, which include prospective drugs. Chemical structures of compounds can be molecular graphs, to which a variety of graph-based techniques in computer science, specifically graph mining, can be applied. The most basic way for analyzing molecular graphs is using structural fragments, so-called subgraphs in graph theory. The mainstream technique in graph mining is frequent subgraph mining, by which we can retrieve essential subgraphs in given molecular graphs. In this article we explain the idea and procedure of mining frequent subgraphs from given molecular graphs, raising some real applications, and we describe the recent advances of graph mining.

Recent experimental development in chemistry has accelerated the efficiency of chemical compound synthesis, by which a huge number of compounds can be accumulated and stored into chemical libraries. Table 1 shows a list of currently available public or commercial databases on chemical compounds. Public databases, such as PubChem from NCBI and ChEBI from EMBL, have been rapidly expanded due to governmental support, which has been motivated by the success of open-access databases, such as GenBank (gene sequences) and PDB (protein structures) [1]. Commercial databases were originally derived from compound catalogs available from chemical market suppliers. Compared with public databases, commercial libraries provide richer information, particularly biological or medical data, such as links to disorder or disease information. In either case of public or commercial databases, the number of entries now exceeds a manually tractable size. For example, there are more than 30 million chemical compounds in PubChem. This situation needs computational analysis, so-called virtual screening [2–4]. Virtual screening can be divided into two types: structure-based screening, which includes docking simulation [5], and ligand-based (or similarity-based) screening [6,7], which identifies molecular structures closely related with some particular biological or medical property. The focus of this article is on ligand-based screening.

Corresponding author:. Takigawa, I. (takigawa@cris.hokudai.ac.jp)

Simple approaches to ligand-based virtual screening are criteria or rules, which define chemical structures related with a focused property, such as 'drug-likeness', for which a typical criterion is 'rule of five' [8]. Rather than such simple criteria, major approaches of virtual screening are quantitative structure–activity relationship (QSAR) or absorption, distribution, metabolism, excretion and toxicology (ADMET) prediction [9–12]. Both have a long history and are based on multivariate analysis, specifically multiple regression analysis. Recently QSAR and ADMET actively incorporate the latest techniques in data mining and machine learning, which both have a discipline similar to multivariate analysis [13–20]. That is, rules or hypotheses are obtained from given data and applied to new data for prediction. In particular, regarding chemical structures of compounds as graphs, this article focuses on 'graph mining' in which machine learning and/or data mining techniques are applied to graphs for analysis [21]. Figure 1 shows an overview of various graph mining-based approaches for knowledge discovery, which can be classified into roughly three categories. The first, traditional group has two steps: (i) chemical compounds are first represented by descriptors which show a variety of graph features, such as the size of compounds, the number of subgraphs with particular topology and physicochemical properties among others; and (ii) any statistical analysis, such as clustering analysis and principal component analysis (PCA) among others, are then applied. The method in the second step can be replaced with regular machine learning methods, such as

 1359-6446/06/$ - see front matter © 2012 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.drudis.2012.07.016

TABLE 1

List of chemical databases

Database	Provider	URL	Approx. size
Public databases			
PubChem	NCBI, NIH, US	http://pubchem.ncbi.nlm.nih.gov/	32M
ChEBI	EMBL-EBI, Europe	http://www.ebi.ac.uk/chebi/	27K
KEGG LIGAND	KEGG, Kyoto University, Japan	http://www.genome.jp/kegg/ligand.html	16K
KEGG DRUG	KEGG, Kyoto University, Japan	http://www.genome.jp/kegg/drug/	9K
ChemBank	Chemical Biology Program, Broad Institute, US	http://chembank.broadinstitute.org/	800K
ChemDB	University of California, Irvine, US	http://cdb.ics.uci.edu	5M
ChemMine	University of California, Riverside, US	http://bioweb.ucr.edu/ChemMineV2/	6.2M
DrugBank	University of Alberta, Canada	http://drugbank.ca	6K
CSLS	NCI, NIH, US	http://cactus.nci.nih.gov/cgi-bin/lookup/search	46M
NLM ChemID-plus	NLM, NIH, US	http://chem.sis.nlm.nih.gov/chemidplus	350K
eMolecules	eMolecules Inc.	http://www.emolecules.com	5M
ZINC	University of California, San Francisco, US	http://zinc.docking.org	21M
Commercial databases			
CMC	Accelrys	http://accelrys.com/products/databases/bioactivity/	9K
MDDR	Accelrys	http://accelrys.com/products/databases/bioactivity/	197K
ACD	Accelrys	http://accelrys.com/products/databases/sourcing/	3.2M
WDI	Daylight Chemical Information Systems	http://www.daylight.com/products/wdi.html	80K
WOMBAT	Sunset Molecular Discovery	http://www.sunsetmolecular.com	322K

Abbreviations: ACD: Available Chemicals Directory; CSLS: Chemical Structure Lookup Service; CMC: Comprehensive Medicinal Chemistry; MDDR: MDL Drug Data Report. World Drug Index.

decision trees, artificial neural networks and support vector machines (SVM), among others, for classification, and self-organizing maps and *k*-means, among others, for clustering. By contrast, in the past 10 years, an active topic in machine learning and data mining is to uncover statistical rules behind input graphs. Two representative machine learning approaches on graphs are (i) 'graph kernels' [20,22,23], which show similarities between two graphs or chemical compounds, and (ii) 'frequent subgraph

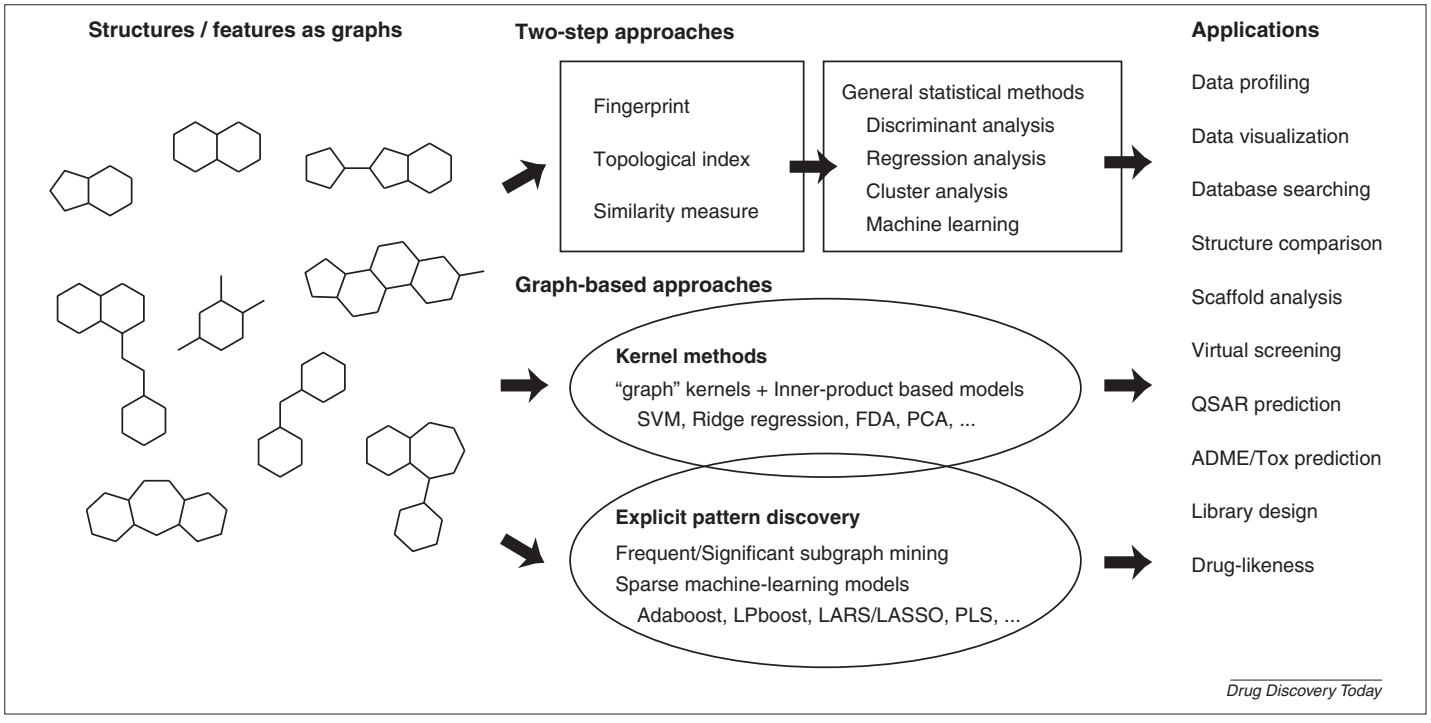

FIGURE 1

Three types of graph mining approaches. *Abbrevations*: ADME/Tox: absorption, distribution, metabolism, excretion and toxicology; LARS: least square regression; LASSO: least absolute shrinkage and selection operator; PCA: principal component analysis; PLS: partial least squares; QSAR: quantitative structure–activity relationship; SVM: support vector machines.