Chemometrics and Intelligent Laboratory Systems 171 (2017) 65–73

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

ChemBCPP: A freely available web server for calculating commonly used physicochemical properties

Jie Dong [a,1], Ning-Ning Wang [a,1], Ke-Yi Liu [b], Min-Feng Zhu [a], Yong-Huan Yun [d], Wen-Bin Zeng [a,*], Alex F. Chen [a,c], Dong-Sheng Cao [a,c,**]

[a] Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, PR China
[b] West China School of Medicine, Sichuan University, Chengdu, People's Republic of China
[c] Center for Vascular Disease and Translational Medicine, The Third Xiangya Hospital of Central South University, Changsha 410013, PR China
[d] Institute of Environment and Plant Protection, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, PR China

ARTICLE INFO

Keywords:
QSPR
Cheminformatics
Physiochemical property prediction
Machine learning
Web server

ABSTRACT

The behavior of a chemical in human or environment mostly depends on its several key physicochemical properties, such as aqueous solubility, octanol-water partition coefficient (logP), boiling point (BP), density, flash point (FP), viscosity, surface tension (ST), vapor pressure (VP) and melting point (MP). Commonly, these properties are important for the environmental sciences and drug discovery, such as the absorption, distribution, metabolism, excretion, and toxicity (ADMET) for medicinal compounds and the common risk assessment for problematic chemicals. At present, the quantitative structure-property relationship (QSPR) model was widely applied to save time and money investment in the early stage of chemical research. Although some satisfactory models were already obtained, most of them are not available for the public researchers and thus cannot be directly applied to practical research projects. Herein, in this study, we developed a user-friendly web server named ChemBCPP that can be used to predict aforementioned 8 important physicochemical properties and calculate several other commonly used properties just by uploading a molecular structure or file. In addition, for a new chemical entity, users can not only get its predicted value but also obtain a leverage value (*h* value) which can be used to evaluate the reliability of predictive result. We believe that ChemBCPP could be widely applied in environmental science, chemical synthesis and drug ADMET fields with the demand for high quality of chemical properties. ChemBCPP could be freely available via http://chembcpp.scbdd.com.

1. Introduction

In recent years, there are about 32,000 chemicals with potential for human exposure which have been identified by the U.S. Environmental Protection Agency (EPA) [1–4]. However, only a small fraction of them have been fully assessed and characterized because of various reasons [5,6]. In general, the behavior of a chemical in human or environment mostly depends on its several key physicochemical properties, such as aqueous solubility, octanol-water partition coefficient (logP), boiling point (BP), density, flash point (FP), viscosity, surface tension (ST), vapor pressure (VP) and melting point (MP) [7–10]. These basic properties play important roles in two main fields: firstly, in the body, they can determine and control the absorption, distribution, metabolism, excretion, and toxicity (ADMET) of chemicals; secondly, in the environment, they are extensively used for new or problematic chemicals evaluation and common risk assessment [11]. Therefore, it is essential to obtain these parameters as early as possible in the stage of chemical research.

Considering the practical situation, how to get these physicochemical properties accurately and quickly has become an urgent problem for chemists and medicinal chemists. At present, there are two main approaches to measure the basic properties for a compound: experimental measures and quantitative structure-property relationship (QSPR) model prediction. For experimental measures, they are usually expensive and time-consuming, what's worse, it is difficult to handle hazardous or reactive chemicals and some pre-manufacturing chemicals which are unavailable for testing. Considering the evaluation efficiency and the

* Corresponding author.
** Corresponding author. Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, PR China.
E-mail addresses: wbzeng@hotmail.com (W.-B. Zeng), oriental-cds@163.com (D.-S. Cao).
[1] The first two authors contribute to the paper equally.

https://doi.org/10.1016/j.chemolab.2017.10.006
Received 24 July 2017; Received in revised form 28 August 2017; Accepted 10 October 2017
Available online 12 October 2017

animal protection, *in silico* QSPR modeling was proposed as an alternative approach to quickly estimate these physicochemical properties for unknown compounds. QSPR models can identify the relationship between the molecular structure and the physicochemical property of interest, and then could give a reasonable prediction for a compound according to its chemical structure [12–16]. At present, there are a lot of QSPR studies based on relatively large datasets for each above-mentioned property and they also obtained some satisfactory predictive models [13,17–19]. However, although the QSPR models based on machine learning offer an effective and powerful way to evaluate the chemical basic properties, most of them are not available for the public researchers and thus cannot be directly applied to practical research projects [20–22]. Therefore, the accessibility of the approved models is a main barrier for these chemists and medicinal chemists who usually don't have any programming skills and are not very familiar with QSPR modeling. Now, there are also some software programs that are already available for physicochemical property evaluation. They almost invariably use QSPR methodology for their predictions, and all that users have to do is to input a chemical structure, for example using SMILES string or a 'mol' file. Other than some freely available software programs such as VCCLAB and Molinspiratio [23,24], most software have to be purchased. For instance, ACD/PhysChem Suite ADMET Predictor and so on [25–27]. In addition, some software programs only cover a subset of these physicochemical properties which are not so sufficient for chemicals evaluation and common risk assessment [28,29]. Therefore, it is an urgent need to develop a freely-available web server for physicochemical properties prediction in a more friendly way.

In this paper, a freely-available web server, ChemBCPP (Basic Chemical Properties Prediction) was developed to easily estimate these most important physicochemical properties (ST, VP, Viscosity, BP, MP, density, solubility, FP) and basic parameters (logP, molecular refraction (MR), topology surface area (TPSA)) using four QSPR methodologies (random forest, support vector machine, Boosting and Cubist). To reduce model uncertainty, the consensus model was obtained by averaging the outputs from four individual models. All these models were strictly assessed by cross validation and external test protocol. What's more, for each predictive model, the application domain was defined by Williams plot according to the Organization for Economic Cooperation and Development (OECD) principles. More specifically, ChemBCPP allows a user to estimate these properties without requiring any external programs. That is to say, users just need to input a chemical by drawing it in an included chemical sketcher window, or entering a structure text file, or imputing the SMILES, and then various properties will be obtained. It also enables users to perform batch computation by using an *.sdf file with multiple molecules. In short, ChemBCPP provides the medicinal chemists a convenient and effective tool to evaluate essential physicochemical properties for specific molecules.

2. Material and method

2.1. Data collection

For these 8 physicochemical properties, we searched for related data from some representative scientific papers and different public databases (DrugBank database, http://www.drugbank.ca; ChEMBL database, https://www.ebi.ac.uk/chembl/). After combining these data, a heterogeneous and chaotic dataset for each property was obtained. To further improve the quality and reliability of the data, some pretreatment steps were applied: 1) removing compounds that without explicit description and exact value; 2) if there are two or more entries for a molecule, the arithmetic mean value of these values was adopted to reduce the random error when their fluctuations was less than 30%, otherwise, this compound would be deleted; 3) specific treatment for some datasets: for boiling point dataset, chemicals have significantly different (>50 K) experimental values in different sources were omitted; for density dataset, it was restricted to chemicals with boiling points greater than 25 °C and chemical with densities greater than 0.5 and less than 5 g/cm^3; for aqueous solubility dataset, chemicals with water solubility exceeding 1,000,000 mg/L were omitted and the data was limited to data points that are within 10 °C of 25 °C; for flash point dataset, chemicals with flash points greater than 1000 °C were omitted. The final datasets and their detailed information were listed in Table 1. LogP, MR and TPSA were calculated by RDKit packages directly [30].

2.2. Descriptor calculation and selection

To obtain a robust and practical model, the informative descriptors for modeling are of high importance. Before the descriptors were calculated, all the compounds were standardized by MOE (the "wash" function of MOE (Molecular Operating Environment software, version 2015, Chemical Computing Group, Montreal, QC, Canada) to disconnect group metals in simple salts, keep only largest molecular fragments, deprotonate strong acids, protonate strong bases and add explicit hydrogens. And then, we calculated two-dimensional (2-D) molecular descriptors using several molecular calculation tools developed by our group [39–42]. These descriptors included 30 constitutional descriptors, 44 connectivity descriptors, 35 topology descriptors, 7 kappa descriptors, 32 Moran autocorrelation descriptors, 60 MOE-type descriptors, 21 Basak descriptors, 64 Burden descriptors, 6 molecular property descriptors, 25 charge descriptors and 89 E-state descriptors. Two pretreatments were performed to delete some uninformative descriptors before further descriptor selection: 1) delete the descriptors whose variance is 0 or approaches 0; 2) if the correlation coefficient between two descriptors is higher than 0.95, only one was reserved. Finally, a series of descriptors (ST: 108, VP: 132, Viscosity: 95, BP: 125, MP: 126, density: 131, Solubility: 133, FP: 125) were prepared for further feature selection and QSPR model building.

In the practical application of machine learning, there are often too many features and some of them are interdependent or irrelevant with response profiles. Therefore, the feature selection process is needed in order to build a high-quality QSPR model. In this part, we employed the random forest algorithm to take a recursive feature elimination [43]. Firstly, all descriptors were applied to build a classification model and these involved descriptors were sorted according to their importance. Then, the last two descriptors were removed and the rest were used to rebuild a model and a new descriptor order was obtained. This process was repeated until the last two remaining descriptors were used for modeling and finally we get a series of models based on different numbers of descriptors. Among them, we can choose a best feature combination according to the number of descriptors and the error value of the model [44].

2.3. Model training and validation

To build the high-performance QSPR models, we employed four the state-of -the-art machine learning algorithms: random forest (RF), support vector machine (SVM), Boosting and Cubist. Herein, we will briefly introduce the basic theory of four algorithms, and more detailed description about them can refer to the related books and literature

Table 1
The detailed information of the physicochemical properties datasets.

No.	Endpoint	Dataset	Unit[a]	Model type	Reference
1	Surface tension	1416	dyn/cm	Regression	[31–33]
2	Vapor pressure	2510	mmHg	Regression	[32–34]
3	Viscosity	557	cP	Regression	[32,33,35,36]
4	Boiling point	5758	°C	Regression	[32,33,37,38]
6	Melting point	9384	°C	Regression	[32–34]
5	Density	8908	g/cm^3	Regression	[32,33,38]
7	Solubility	5020	mol/L	Regression	[32–34]
8	Flash point	8362	°C	Regression	[32,33,38]

[a] The unit here for each one is the original unit. The units in the prediction result are *-log mol/L* for Solubility and log *mmHg* for Vapor pressure.