Chemometrics and Intelligent Laboratory Systems 149 (2015) 177–204

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Chemometrics tools in QSAR/QSPR studies: A historical perspective

Saeed Yousefinejad, Bahram Hemmateenejad *

Chemistry Department, Shiraz University, Shiraz, Iran

ARTICLE INFO

Article history:
Received 29 May 2015
Received in revised form 22 June 2015
Accepted 25 June 2015
Available online 3 July 2015

Keywords:
QSAR
QSPR
Variable selection
Model development
Validation

ABSTRACT

One of the most extended subfields of chemometrics, at least by considering the number of publications and interested researchers, is QSAR/QSPR. During the time, various improved and/or alternative methods have been developed in different principal steps of QSAR/QSPR including (i) variable selection, (ii) model construction and (iii) validation evaluation. In the current manuscript, it is tried to represent a short overview on critical and bold developments of chemometrics tools utilized in QSAR/QSPR studies.

1. Introduction

Quantitative structure–activity relationship (QSAR) is a quantitative method that deals with finding a model to relate chemical structural features of compounds (descriptors) to their definite biological activity. Modeling the biological activity was extended to other physical/chemical properties which is called quantitative structure–property relationship (QSPR). Different properties or behaviors of chemical molecules have been investigated in the field of QSPR. Some examples are quantitative structure–reactivity relationships (QSRRs) [1], quantitative structure–chromatography relationships (QSCRs) [2,3], quantitative structure–toxicity relationships (QSTRs) [4], quantitative structure–electrochemistry relationships (QSERs) [5–7], and quantitative structure–biodegradability relationships (QSBRs) [8,9].

Searching on the Web of Science core collection using keywords such as "quantitative structure–activity", "quantitative structure–property", "quantitative structure–electrochemistry", "quantitative structure–toxicity", "quantitative structure–function", "quantitative structure–retention", "quantitative structure–chromatography", "quantitative structure–biodegradability" or "structure–activity correlation" resulted in a sum of 11,000 records. The contribution of each kind of the recorded documents, including research articles, proceeding papers, review articles, meting abstracts and book chapters, is shown in Fig. 1. It is not strange that research articles occupy 82% of the records. However, it is interesting that the noted keywords led to the finding of about 670 review articles.

Fig. 2 shows the evolution of the number of records using the above keywords as a function of published year from 1966 to 2015. The first article used the keyword "structure–activity correlation" which appeared in 1966 and then this keyword and the term "quantitative structure–activity" were used repeatedly by adding other keywords. From Fig. 2 one can observe three sudden increases in publications' growth. The first and second sudden changes are observed at the beginning of the 1990 and 2000 decades. The third sharp change happened after 2006. However, the growth of publications has been stopped after 2010.

In a QSAR/QSPR study, collecting or designing a subset of chemical or biological compounds, generation of potent descriptors capable to reflect the structure of compounds, selection of reverent descriptors to include in model, construction of a regression model and checking the validity and stability of the suggested model are five essential steps. In the current article, our short overview is focused on the last three steps i.e. (i) variable selection, (ii) model construction and (iii) validation evaluation in which chemometrics tools have been involved more significantly. However it should be emphasized that without numerous efforts in the development and generation of various molecular descriptors [10], the growth and effectiveness of QSAR/QSPR studies were impossible.

2. Historical roots of QSAR tree

By searching in literature, the root of huge QSAR tree could be found in the thesis of Cros, entitled "Action de l'alcoholamyliquesurl'organisme" (1863, Faculty of Medicine, University of Strasbourg, France) which noticed the relationship between toxicity of primary aliphatic alcohols and their water solubility. Without doubt, focusing on the concept of

* Corresponding author. Tel.: +98 713 6460724; fax: +98 713 6460788.
E-mail address: hemmatb@shirazu.ac.ir (B. Hemmateenejad).

http://dx.doi.org/10.1016/j.chemolab.2015.06.016
0169-7439/

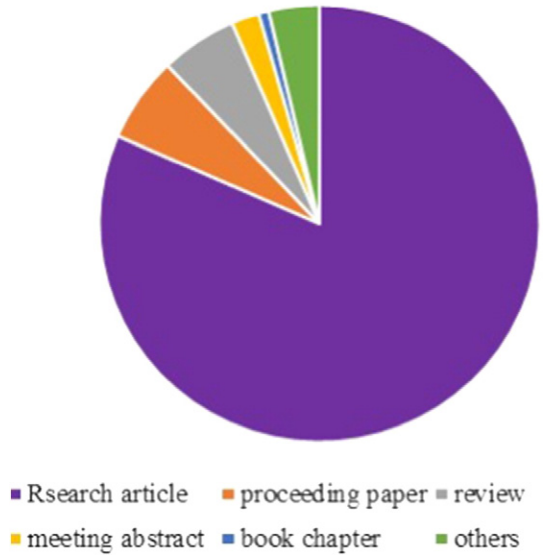

Fig. 1. Contribution of different kinds of documents recorded inWeb of Science core collection on QSAR/QSPR subject (prepared on May 27, 2015). Others include note, editorial material, letter, correction and book review.

"molecular structure" in the years between 1860 and 1880 has led to the development of QSAR researches [11]. Among the works on the molecular structure especially "three-dimensional" concept, the attempts of Butlerov (1861–65), Wislicenus (1869–73) and Van't Hoff (1874–75), are notable [11].

Maybe the first primary QSAR study could be related to Brown and Fraser [12], who proposed the existence of a correlation between molecular constitution and biological activity of different alkaloids. In 1884, Mills explained more on the hypothesis of structure–property correlation in his article entitled "On melting-point and boiling-point as related to chemical composition" [13]. After that, at the end of the 19th to beginning of the 20th centuries, other attempts were made to clarify this hypothesis (e.g. [14,15]). Somebody believes that the works of Hammet on substituent effects in organic reactions, 1935–1938 [16–19], had an outstanding role in developing QSAR models. Some years later, when the first theoretical QSAR/QSPR studies were proposed in the middle of the 20th century [20,21] based on the descriptors obtained from graph theory, it was not expected that QSAR can become an inseparable part of molecular and drug design.

After the development of some other categories of structural descriptors, the world came ready to see the revolution of QSAR/QSPR. To reach this step, attempts of Pauling and also Coulson on the chemical bond concept [22,23], Sanderson on the atomic electronegativity [24], and the researches on electronic distribution and quantum-chemical descriptors were very important and determinant [25–28]. After that, some specific QSAR/QSPR approaches were proposed step by step which is prominent from the historical point of view. After the suggestion of Linear Free Energy Relationships (LFER) by Hammet [19], one of the most principal approaches was linear solvation energy relationship (LSER) which had been proposed in 1952–3 by Taft [29,30].

Someone considers the start of the modern period of QSAR (or official birth date of QSAR) in 1962, when Hansch et al. correlated the biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients [31]. Their work was the first multi-parametric QSAR model. In continuation of this official birthday, other work of Hansch and coworkers [32,33] attracted a lot of interests to this field and caused the QSAR explosion [34].

A historical bold point in structure–activity studies was the publication of Free and Wilson in 1964 [35] which had an effective role in the development of QSAR/QSPR. The basic idea in their proposed approach was the modeling of a biological activity (or chemical property) by looking at the presence/absence of substituent functional group on a common structural skeleton.

In the 1980s, proposing and utilizing different categories of topological and geometrical descriptors, entered the 3D geometry of molecules in QSAR/QSPR [11] and derived enhancement in prediction of models as well as their description ability. Another shiny and really important stage for QSAR in the 1980s was the development of molecular descriptors based on molecular interaction fields (MIFs) which led to the well-known field of 3D QSAR. The concern of 3D QSAR was the finding of the interaction energies between a compound and specified chemical probes at certain spatial points of 3D space [36,37]. Different interaction probes, such as hydrogen atom, water molecule and methyl group [36] have been proposed to detect and calculate the interaction energies of a molecule in a grid space. Historically, the first approach of 3D QSAR category was the GRID method proposed by Goodford in 1985 [38] and then was developed by Cramer et al. in 1988 by the name of *comparative molecular field analysis* (CoMFA) [39]. These methods are done in a lattice model by aligning molecules to compare them and exploring their MIF information in 3D space [39]. Other important 3D QSAR methods based on MIF was introduced later such as

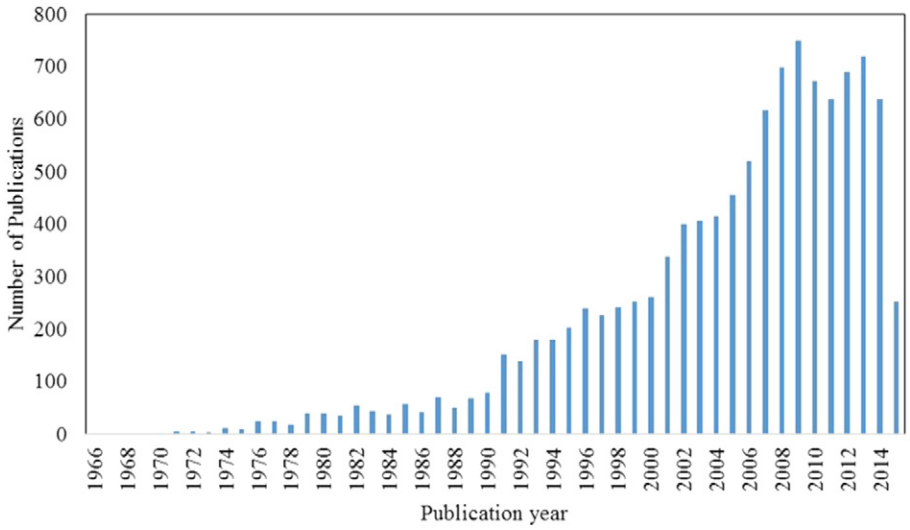

Fig. 2. Growth pattern of publications on QSAR/QSPR subject (prepared on May 27, 2015).