Minerals Engineering 21 (2008) 1038–1043

Contents lists available at ScienceDirect

Minerals Engineering

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

Quantitative structure–activity relationship (QSAR) approach for the selection of chelating mineral collectors

R. Natarajan, I. Nirdosh *

Department of Chemical Engineering, Lakehead University, Thunder Bay, Ontario, Canada P7B 5E1

ARTICLE INFO

Article history:
Received 27 December 2007
Accepted 19 March 2008
Available online 6 May 2008

Keywords:
Topological indices
Molecular similarity
Dissimilarity selection
Mineral collectors
Froth flotation

ABSTRACT

Quantitative structure–activity/property relationship (QSAR/QSPR) studies are mathematical quantification of relations between structure and activity or property. These are extensively used in pharmaceutical and agricultural chemistry for screening potential compounds for specific biological activity and also in environmental toxicology. The linear or nonlinear regression models that establish a relation between the structure and the activity/property are then used to predict the activity/property/toxicity of any number of structurally related compounds including those that are yet to be synthesized. Computable molecular descriptors are preferred to experimental properties in QSAR analyses because they require molecular structure as the only input and can be inexpensively calculated for a chemical in less than a millisecond. The use of QSAR approach for modeling the efficiencies of chelating collectors was tested and it was learnt that the performances of chelating mineral collectors are amenable to QSAR modeling. The molecular similarity approach used for the selection of structural analogues was used for the selection of *N*-arylhydroxamic acids. In the case study, 10 *N*-arylhydroxamic acids were selected from a virtual database of 3800 compounds and thus, a large structural space was explored without spending much resource. The 10 compounds were then synthesized and tested as collectors for sphalerite. Molecular similarity–dissimilarity clustering was found to be an effective scientific tool in the pursuit of finding new mineral collectors.

1. Introduction

Pharmacological activities, chemical reactivity and physicochemical properties of organic compounds depend on molecular structure. The quantitative-relation that brings out the dependence of a property or activity (P) on entire molecular structure or its substructural fragment is called quantitative structure–activity relationship (QSAR)

$$P = a_1 \times S_1 + a_2 \times S_2 + \cdots \tag{1}$$

where a indicates a coefficient and S indicates a structural property or an empirical property of the whole molecule or any of its substructural units. Depending on the nature of P in Eq. (1), the relation is also called as quantitative structure–property relationship (QSPR) or quantitative structure–toxicology relationship (QSTR). QSAR approach is extensively used in pharmaceutical chemistry and predictive toxicology. In pharmaceutical industry, application of QSAR may be considered as the first step in a multi-step computer-assisted screening of therapeutically active compounds. Development of QSAR models with good predictive abilities rely on the effective transformation of structural details of molecules into numerical quantities. Topological indices (TIs) are extensively used to encode structural details into numbers, i.e. they are the molecular metrics. There are over 300 TIs and their definitions, and calculations can be obtained from the monographs by Devillers and Balaban (1999) or Todeschini and Consonni (2000). In addition to TIs, quantum mechanical parameters such as energy of highest occupied molecular orbital (E_{HOMO}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), and partial atomic charges are also used in forming QSAR models (Karelson et al., 1996). Computable molecular descriptors are preferred to experimental physicochemical properties because their computation requires no information other than molecular structure as input.

Though the dependence of efficiency of chelating collectors on their molecular structures had been studied (Somasundaran et al., 1993; Nagaraj 1988; Marabini et al., 1988; Marabini, 1993; Das et al., 1995; Fuerstenau et al., 1964; Ackerman et al., 1984, 1987, 1999; Nirdosh et al., 1994; Urbina, 2003; Natarajan and Nirdosh, 2001a, 2006), no quantitative treatment had been reported. There is implicit dependence of efficiency of a chelating agent (organic compound) as flotation collector on the molecular architecture, and QSAR is the approach that quantifies this intuitiveness. Natarajan et al. (1999) extended the QSAR approach to model the efficiency of chelating collectors. They successfully

---

* Corresponding author. Tel.: +1 807 343 8343.
E-mail addresses: rnataraj@lakeheadu.ca (R. Natarajan), inirdosh@lakeheadu.ca (I. Nirdosh).

0892-6875/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mineng.2008.03.009

showed that the separation efficiencies of chelating collectors are amenable to QSAR modeling and topological indices (Natarajan and Nirdosh, 2001b), electronic parameters, physicochemical properties such as octanol–water coefficient ($\log P$), soil–water partition coefficient ($\log K_{oc}$), and geometrical parameters that describe the molecular surface area could be used to form the QSAR models (Natarajan et al., 2002a; Natarajan and Nirdosh, 2003).

A large number of TIs can be calculated for a given set of compounds and many topological indices are mutually related. Principal Component Analysis (PCA) is one of the ways to eliminate the problem of over-fitting by using correlated descriptors and it also extracts minimum number of orthogonal parameters that explain the maximum variance among the calculated descriptors. Principal components (PC) thus extracted can be used as independent parameters (descriptors) to form the regression equations. Natarajan et al. (2002b) used this approach to correlate the separation efficiencies of cupferron derivatives. A brief summary of the QSAR work on different collectors is given below:

Multiple linear regression (MLR) using topological indices, as well as principal components were carried out on three sets of data:

(1) A set of 22 cupferron derivatives used to float a uranium ore (Nirdosh et al., 1994).
(2) A set of 14 mercaptobenzothiozole tested to float a lead ore (Marabini et al., 1988).
(3) A set of 9 *o*-aminothiophenol tested to float a zinc ore (Marabini et al., 1988).

The three data sets and a set of 17 *N*-arylhydroxamic acids used to float a copper zinc ore (Natarajan and Nirdosh, 2001a) were modeled using calculated physicochemical parameters instead of topological indices.

Octanol–water partition coefficient and soil–water partition coefficient were two of the important physicochemical parameters. Octanol–water partition coefficient (usually expressed as its logarithmic value, $\log P$) is extensively used as a measure of hydrophobicity of compounds in QSAR modeling along with electronic and steric parameters. Several software are available for computing $\log P$ from molecular structure using group contribution parameters, and the calculated $\log P$, denoted as $\mathrm{Clog}P$ is used in QSAR modeling. On the other hand soil–water partition coefficient, $K_{oc,}$ gives the extent to which an organic compound partitions itself between soil or sediment and water. Similar to octanol–water partition coefficient soil–water partition coefficient is also used as its logarithmic value and can be calculated using software. The estimated $\log K_{oc}$ therefore expresses the tendency of adsorption independent of soil characteristics. QSAR studies on mineral flotation data were reviewed by Natarajan et al. (2003) and the important findings of the QSAR modeling studies are:

(1) Topological indices or the principal components extracted from them could be used as independent parameters to describe the molecular architecture.
(2) In the case of congeneric collectors, (collectors having the same chelating group), introduction of electronic and quantum mechanical parameters did not improve the predictive ability significantly.
(3) Computed soil–water partition coefficient, $\log K_{oc}$, and octanol–water partition coefficient, $\mathrm{Clog}P$, were found to be good computable property parameters that gave QSAR equations of good predictive ability.
(4) The regression equation formed had good predictive ability and the differences between predicted and experimental separation efficiencies were less than the experimental errors (±5).

Efficiency of a collector depends on the extent to which the collector molecule renders the mineral surface hydrophobic and also on the extent to which it adsorbs on the mineral. $\mathrm{Clog}P$ is a measure of hydrophobicity while $\log K_{oc}$ is a measure of adsorption tendency of an organic compound on soil (i.e. mineral, in flotation) and this might be the reason for good QSAR models obtained using them. Though these two parameters cannot account for the various factors involved in the collector–mineral interaction, they seemed to work well for a set of congeneric compounds.

An earlier study (Natarajan and Nirdosh, 2001a) on *N*-arylhydroxamic acids indicated that flotation of sphalerite may be obtained without activation by copper sulphate. Hence, the research was continued to synthesize more *N*-arylhydroxamic acids including dihydroxamic acids. A total of 31 compounds (8 used in earlier study and 23 new compounds) were tested to float sphalerite from the Cu–Zn ore used in the previous study. Unlike the previous study, *N*-arylhydroxamic acids were used to float sphalerite after floating most of the chalcopyrite with a collectorless prefloat. The results of the study were reported (Natarajan and Nirdosh, 2006) and no QSAR modeling was carried out for this set of data. However, sphalerite was floated without activation as observed in the previous study. It was not sure whether the activation was due to the copper present in the ore. Hence, the study was extended for the flotation of sphalerite from a Pb–Zn ore. Instead of testing the compounds that gave the best results with the Cu–Zn ore, it was decided to use molecular similarity-based selection of compounds in order to explore a large database. QSAR studies on chelating collectors showed that compounds with very close structural similarities had comparable collector efficiencies of flotation of a particular mineral and this gave impetus to test molecular similarity-based clustering, a computational method used in pharmaceutical industry, as a scientific method for selecting compounds for synthesis and testing as sphalerite collectors. A brief outline of molecular similarity clustering and selection is given in the following section.

2. Molecular dissimilarity/similarity clustering

Johnson and Maggiora (1990) showed that molecular similarity clustering was an efficient method in the selection of compounds and prediction of biological and physicochemical properties, and Lajiness (1990) found this method useful for the selection of a small number of compounds from a virtual library (a database of compounds that contain all possible structures) in the process of identification of new therapeutic agents. To create a similarity space for the selection of compounds, an appropriate measure of molecular similarity is needed. Physicochemical properties, topological indices and sub-structures called atom-pairs had been used in constructing structure spaces for selection (Basak et al., 2006). Use of experimental data has a very restricted application due to the limitation in their availability for all the chemicals under consideration. Computable properties are always preferred for the purpose. Clustering procedure using topological indices as a measure of molecular similarity may be explained as below:

Let us consider that t numbers of topological indices are calculated for the given set of m structures. The data set is very large hence, the $(t \times m)$ data matrix is subjected to data reduction by principal component analysis (PCA). The principal components are computed in such a way that the first factor describes the largest part of the data variability, the second factor is orthogonal to the first and describes the largest part of the variability left over by the former factor, and so on. Usually principal components with eigenvalue equal to or greater than one are retained or extracted. Hence, by performing principal component analysis, the majority of the information contained within the original data can be repre-