Application of Graph Theory: Relationship of Topological Indices with the Partition Coefficient (logP) of the Monocarboxylic Acids

Fatemeh Shafiei
Department of Chemistry, Arak Branch, Islamic Azad University, Arak, Iran

Received November 2015; Accepted November 2015

ABSTRACT
It is well known that the chemical behavior of a compound is dependent upon the structure of its molecules. Quantitative structure – activity relationship (QSAR) studies and quantitative structure – property relationship (QSPR) studies are active areas of chemical research that focus on the nature of this dependency. Topological indices are the numerical value associated with chemical constitution purporting for correlation of chemical structure with various physical properties, chemical reactivity or biological activity. Graph theory is a delightful playground for the exploration of proof techniques in Discrete Mathematics and its results have applications in many areas of sciences. One of the useful indices for examination of structure- property relationship is Randic’ index. In this study is represented the relationship between the Randic’, Balaban and Szeged indices and Harary numbers to the octanol-water partition coefficient (logP) of monocarboxylic acids (C2- C20) are established, and then, some useful topological indices for examination of the structure- property relationship are presented.

Keywords: Topological indices; logP; Monocarboxylic acid; QSPR

INTRODUCTION
Carboxylic acids are characterized by the carboxyl group which combines the carbonyl group of aldehydes and ketones with the hydroxyl group of alcohols and phenols. Since the carbonyl and hydroxyl groups are directly bonded to each other each affects the properties of the other. The result is that the hydroxyl group of a carboxylic acid is considerably, but not completely, different from its alcohol or phenol sibling; the same can be said when one compares the carbonyl of a carboxylic acid with that of an aldehyde or ketone.

The fact that the properties of a molecule are tightly connected to its characteristics is one of the fundamental concepts in chemistry. In this connection, graph theory has been successfully applied and some thermodynamic properties [1-5].

Chemical graph theory is a branch of mathematical chemistry. It is concerned with handling chemical graphs that represent chemical systems. Hence, chemical graph theory deals with analyses of all consequences of connectivity in a chemical system. In other words, chemical graphs theory is concerned with all aspects of the application of graph theory to chemistry area.

A graph is a topological concept rather...
than a geometrical concept of fixed
gometry. Therefore, Euclidean metric
lengths, angles and three-dimensional
spatial configurations have no signiﬁcance
in topological concept.
Chemists employ various types of
designations and formulas when they want
to communicate information about
chemical compounds and their structures.
In spite of this fact, most of the names and
formulas have no direct, immediate or
explicit mathematical meaning.
It has been found to be a useful tool in
QSR (Quantitative Structure-Activity
Relationship) and QSPR (Quantitative
Structure-Property Relationship)[6-11].
Numerous studies have been made relating
to the above mentioned ﬁelds by using
what are called topological indices (TI). In
1975, Randic proposed a topological index
that has become one of the most widely
used in both QSR and QSPR studies
[12].
Quantitative structure-activity
relationship (QSR) are mathematical
models designed for the correlation of
various types of biological activity,
chemical reactivity, equilibrium, physical
and physicochemical properties with
electronic, steric, hydrophobic and other
factors of a molecular structure of a given
series of compounds such as substitution
constants, topological indices (TI) as well
as with solvent and other physicochemical
parameters.
In this paper, a variant of QSR
studies, the so-called quantitative
structure-property relationships (QSPR),
using topological indices as molecular
descriptors [10], is used. The incredibly
great number of works devoted to this has
led to the appearance of hundreds of new
indices, which are useful to describe with
more or less accuracy speciﬁc properties of
given compounds.
In the last few years, also the necessity
of describing the three-dimensional
character of molecular structures has
contributed to the development of three-
dimensional indices [13]. The classical
topological approach [5] relates the
chemical structure constitution (the two-
dimensional model of a molecule, which is
represented by a structural formulae) with
a non-dimensional numerical entity, the so-
called topological indices.
In this correspondence, each structure
has a single associated descriptor, but not
vice versa; one index may correspond to
more than a graph. Here arises the problem
of the degeneracy; so it is desirable that the
working indices present low degeneracy.
To translate chemical structures into a
single number, the graph theory visualizes
chemical structures as mathematical object
sets consisting of vertices or points, which
symbolize atoms, and vertices or lines,
linking a pair of edges, which represent
covalent bonds or shared electron pairs of
covalently linked atoms situated at a
topological distance equal to unit.
The studies of quantitative structure-
activity relationship (QSR) and the
quantitative structure - property
relationships (QSPR) involve various
chemometric methods in which the
chemical behavior of a compound is
correlated with its structure represented by
the topological indices [1]. For example,
QSPR methods were applied for the
prediction of an octanol-water partition
coefficient [2].
In this study, the interesting results of
structure-property relationship between the
Randic(1X ), Balaban (J) and Szeged (Sz)
indices and Harary numbers (H) to the
octanol-water partition coefficient (logP)
of monocarboxylic acids (C2H4O2- C20H40O2)
are presented.
At ﬁrst, a brief review on the classical
topological indices are calculated. After
wards, the molecular descriptors that
include the necessary structural information for properly description of system are employed to derive a numerical correlation with thermodynamic properties. Finally, some useful topological indices for examination of the structure- property relationship are presented.

**DATA SOURCES**

(C\textsubscript{2}H\textsubscript{4}O\textsubscript{2} - C\textsubscript{20}H\textsubscript{40}O\textsubscript{2}) are taken from chemicalize Web Book\[20\].

**TOPOLOGICAL INDICES**

All the used topological indices were calculated using all hydrogen suppressed graph by deleting all the carbon hydrogen as well as heteroatomic hydrogen bonds from the structure of the monocarboxylic acids. The calculations of these indices are well documented in the literature and therefore, their detailed calculations are not given here. However, blow we have given the final expression for the calculation of these indices.

**The connectivity index (1X)**

The connectivity index 1\textit{X} = 1\textit{X} \text{ (G)} of \text{G} is defined by Randic' \[12\] as:

\[ 1_X = 1_X (G) = \sum_{\text{alleges}} \left[ d_{ij} d_{ij} \right]^{-0.5} \quad (1) \]

where \textit{d}_{ij} and \textit{d}_{ij} are the valencies of the vertices and \textit{j} that define the edge \textit{ij}.

**Szeged index (Sz)**

The Szeged index, \textit{Sz} = \text{Sz} \text{(G)}, is calculated \[16,17\] according to the following expression:

\[ \text{Sz} = \text{Sz} (G) = \sum_{\text{alleges}} n_u \cdot n_v \quad (2) \]

Edges where \textit{n}_u is the number of vertices lying closer to one end of the edge \textit{e}=uv; the meaning of \textit{n}_v is analogous. Edges equidistance from both the ends of an edges, \textit{e}=uv are not taken into account.

**Balaban index (J)**

The Balaban index, \textit{J}=J(G) of \text{G}, was introduced by Balaban \[1-5-18\] as the average- distance sum connectivity. It is defined as:

\[ J = \frac{M}{\mu + 1} \sum_{\text{alleges}} \left[ d_{ij} d_{ij} \right]^{0.5} \quad (3) \]

where \textit{M} is the number of edges in \text{G}; \textit{\mu} is the cyclomatic number of \text{G}; and \textit{d} (\textit{i}) is the distance sum where \textit{i}=1,2,..,\textit{N}. The cyclomatic number \textit{M}=\text{M(}G\text{)} of a polycyclic graph \text{G} is equal to the minimum number of edges that must be removed from \text{G} to transform it to the related acyclic graph. For trees, \textit{M}=0; for monocycles \textit{M}=1.

**Harary number (H)**

The Harary number (\textit{H}) was introduced in 1991 by Harary \[14, 15\]. This index is defined from the inverse of the squared elements of the distance matrix according to the expression:

\[ H = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( D_{ij} \right)^2 \quad (4) \]

where \textit{D}_{ij} is the matrix whose elements are the squares of the reciprocal distances.

**Graphs**

For drawing the graphs of our results, we used the Microsoft Office Excel – 2003 program.

**DISCUSSION AND DESIGNING THE QSPR MODELS**

A graph – theoretical approach to QSPR is based on the use of topological indices for encoding the structural informations. The topological indices term indicates a characterization of a molecule (or a corresponding molecular graph by a single number). The need to represent molecular structure by a single number arises from the fact that most molecular properties are
recorded as single numbers. Therefore, QSPR modeling reduces to a correlation between the two sets of numbers via algebraic expression. (One set of numbers represents the properties, and the other set represents the structures of molecules under study.)

A novel method for computing the new descriptors to construct QSPR is presented. First, molecular topological indices are calculated. The values of topological indices of Randic’ (1X), Harary (H), Balaban (J) and Szeged (Sz) for the 19 compounds of training set are also described in Table 1.

After words, the molecular descriptors, which include the structural information necessary to property describe the system, are employed to derive numerical correlation with property. The octanol-water partition coefficient (logP) of monocarboxylic acids Tables 2 are taken from book and web book [19,20] for the compounds of training set.

Figs. 1-8 were shown, respectively, the plots of logP for the compounds of training set against the 1X, log1 X, J, log J, H, logH, Sz and logSz.

There are several ways to design the QSPR models. Here we outline one possible strategy which contains five steps:

Step 1. To get a reliable source of experimental data for a given set of molecules. This initial set of molecules is sometimes called the training set. The data in this set must be reliable and accurate. The quality of the selected data is important because it will affect all the following steps.

Step 2. The topological index is selected and computed. This is also an important step because selecting the appropriate topological index (or indices) can facilitate finding the most accurate model.

Step 3. The quality of the QSPR models can be conveniently measured by the correlation coefficient (R²). A good QSPR model must have R² > 0.99. Therefore, step 3 is a central step in the design of the structure – property models.

Step 4. Predictions are made for the values of the molecular property for species that are not part of the training set via the obtained initial QSPR model. The unknown molecules are structurally related to the initial set of compounds.

Step 5. The predictions are tested with unknown molecules by experimental determination of the predicted properties. This step is rather involved because it requires acquiring or preparing the test molecules.

We will apply the procedure from the preceding section, to give an instructive example of the design of the QSPR model for predicting the octanol-water partition coefficient of monocarboxylic acids. As the initial set we will consider monocarboxylic acids with up to 5 carbon atoms (4 molecules). The octanol-water partition coefficient of these monocarboxylic acids are taken from [19,20], and the molecular topological indices such as the Randic’ index (1X), Harary number (H), Balaban index (J) and Szeged index (Sz) of the above mentioned are calculated (see table 1).

Models 5 to 12 exhibit the applicability of topological and logP for QSPR study of mentioned monocarboxylic acids.

The coefficient of determination (R²) for applying model (5) to the experimental data (Table 2) is 0.9965 and for applying model (7) is 0.9949. Thus, according to this result, the topological 1X and H can be used for predicting the octanol-water partition coefficient (logP) of monocarboxylic acids.

Table 3. shows the variance and highest (positive and negative) differences between experimental values of logP and the values obtained using 1X and H for the 4 compounds of training set (Table 1).
Comparison of the predicted logP with the experimental values, revealed that models 5 and 7 are shown that the topological $^1X$ is quite suitable for predicting the logP of these carboxylic acids.

**CONCLUSION**

Graph theory has provided the chemists with a variety of very useful tools that can be used to predict many interesting physical and chemical properties of considered materials. Regarding this aspect we have presented a structure–property relationship based on topological indices. The instructive example was directed to the design of the structure–property model for predicting the octanol-water partition coefficient of monocarboxylic acids (C$_2$H$_4$O$_2$ - C$_{20}$H$_{40}$O$_2$). Four selected topological indices were tested. The correlation of the Randic index ($^1X$) with logP, appeared to have better results than the other indices. Meanwhile, the discussion of relationship areas become straightforward.

$$
\log P = 0.9063(^1X) - 1.6449 \\
\log P = 10.637(\log^1X) - 3.8932 \\
\log P = 0.1465(H) - 0.3766 \\
\log P = 7.5182(\log H) - 6.3435 \\
\log P = 14.413(J) - 36.138 \\
\log P = 86.424(\log J) - 34.379 \\
\log P = 0.0044(Sz) + 1.6272 \\
\log P = 3.6531(\log Sz) - 4.7304
$$

![Table 1. The values of randic ($^1X$), Balaban (J) and Szeged indices (Sz) and Harary numbers (H) of monocarboxylic acids (C$_2$H$_4$O$_2$ - C$_{20}$H$_{40}$O$_2$)](attachment:table1.png)
Table 2. The values of the octanol-water partition coefficient (logP) monocarboxylic acids (C_{2}H_{4}O_{2} - C_{20}H_{40}O_{2})

<table>
<thead>
<tr>
<th>compounds</th>
<th>log P</th>
<th>compounds</th>
<th>log P</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_{2}H_{4}O_{2}</td>
<td>-0.22</td>
<td>C_{12}H_{24}O_{2}</td>
<td>4.48</td>
</tr>
<tr>
<td>C_{3}H_{6}O_{2}</td>
<td>0.48</td>
<td>C_{13}H_{30}O_{2}</td>
<td>4.92</td>
</tr>
<tr>
<td>C_{4}H_{8}O_{2}</td>
<td>0.92</td>
<td>C_{14}H_{32}O_{2}</td>
<td>5.37</td>
</tr>
<tr>
<td>C_{5}H_{10}O_{2}</td>
<td>1.37</td>
<td>C_{15}H_{36}O_{2}</td>
<td>5.81</td>
</tr>
<tr>
<td>C_{6}H_{12}O_{2}</td>
<td>1.81</td>
<td>C_{16}H_{38}O_{2}</td>
<td>6.26</td>
</tr>
<tr>
<td>C_{7}H_{14}O_{2}</td>
<td>2.26</td>
<td>C_{17}H_{40}O_{2}</td>
<td>6.70</td>
</tr>
<tr>
<td>C_{8}H_{16}O_{2}</td>
<td>2.70</td>
<td>C_{18}H_{42}O_{2}</td>
<td>7.15</td>
</tr>
<tr>
<td>C_{9}H_{18}O_{2}</td>
<td>3.14</td>
<td>C_{19}H_{44}O_{2}</td>
<td>7.59</td>
</tr>
<tr>
<td>C_{10}H_{20}O_{2}</td>
<td>3.59</td>
<td>C_{20}H_{40}O_{2}</td>
<td>8.30</td>
</tr>
<tr>
<td>C_{11}H_{22}O_{2}</td>
<td>4.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Comparison between predicted (three model) and experimental values of enthalpies of combustion and their logarithmic of respect monocarboxylic acids

<table>
<thead>
<tr>
<th>compounds</th>
<th>log P</th>
<th>log P predicted</th>
<th>Deleted residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model (5)</td>
<td>Model (7)</td>
</tr>
<tr>
<td>C_{6}H_{12}O_{2}</td>
<td>1.810</td>
<td>1.772</td>
<td>0.038</td>
</tr>
<tr>
<td>C_{9}H_{18}O_{2}</td>
<td>3.140</td>
<td>3.131</td>
<td>0.009</td>
</tr>
<tr>
<td>C_{12}H_{24}O_{2}</td>
<td>4.480</td>
<td>4.491</td>
<td>-0.011</td>
</tr>
<tr>
<td>C_{17}H_{34}O_{2}</td>
<td>6.700</td>
<td>6.756</td>
<td>-0.056</td>
</tr>
</tbody>
</table>

Fig. 1. Plots of the Randic' index (1X) versus logP for (C_{2} - C_{20}) monocarboxylic acids (without C_{6}, C_{9}, C_{12}, C_{17}).

Fig. 2. Plots of the logX versus logP for (C_{2} - C_{20}) monocarboxylic acids (without C_{6}, C_{9}, C_{12}, C_{17}).
Fig. 3. Plots of the Harary index (J) versus logP for (C_2- C_20) monocarboxylic acids (without C_6, C_9, C_12, C_17).

Fig. 4. Plots of the log H versus log P for (C_2- C_20) monocarboxylic acids (without C_6, C_9, C_12, C_17).

Fig. 5. Plots of the Balaban index (J) versus log P monocarboxylic acids (without C_6, C_9, C_12, C_17).

Fig. 6. Plots of the log J versus for (C_2- C_20) for (C_2- C_20) monocarboxylic acids (without C_6, C_9, C_12, C_17).

Fig. 7. Plots of the Szeged index (Sz) versus log P monocarboxylic acids (without C_6, C_9, C_12, C_17).

Fig. 8. Plots of the logSz versus logP for (C_2- C_20) for (C_2- C_20) monocarboxylic acids (without C_6, C_9, C_12, C_17).
REFERENCE


[20]. WWW. Chemicalize.org