Journal of Molecular Structure 1030 (2012) 57–66

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Structure

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

Structure-based predictions of ^{13}C-NMR chemical shifts for a series of 2-functionalized 5-(methylsulfonyl)-1-phenyl-1*H*-indoles derivatives using GA-based MLR method

Raouf Ghavami [a,*], Faridoon Sadeghi [b], Zolikha Rasouli [a], Farhad Djannati [c]

[a] Department of Chemistry, Faculty of Science, University of Kurdistan, P.O. Box 416, Sanandaj, Iran
[b] Department of Chemistry, Islamic Azad University, Saveh Branch, P.O. Box 39187-366, Saveh, Iran
[c] Department of Mathematics, Faculty of Science, Soran University, Soran, Iraq

HIGHLIGHTS

- A central issue of this study was to assess the capabilities of GA-MLR algorithm in simulation of ^{13}C-NMR spectra of some indole derivatives using different GIPF descriptors.
- The accuracy of all developed models were confirmed using different types of internal and external procedures and various statistical tests.
- The domain of applicability for each model which indicates the area of reliable predictions was defined.
- The successful results lead to the conclusion that ^{13}C-NMR chemical shifts property can be successfully modeled.

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form 18 June 2012
Accepted 19 June 2012
Available online 26 June 2012

Keywords:
^{13}C-NMR
QSPR
Nuclear magnetic resonance
Chemical shift calculation
GA-MLR
Applicability domain

ABSTRACT

Experimental values for the ^{13}C NMR chemical shifts (ppm, TMS = 0) at 300 K ranging from 96.28 ppm (C4′ of indole derivative 17) to 159.93 ppm (C4′ of indole derivative 23) relative to deuteride chloroform ($CDCl_3$, 77.0 ppm) or dimethylsulfoxide (DMSO, 39.50 ppm) as internal reference in $CDCl_3$ or DMSO-d^6 solutions have been collected from literature for thirty 2-functionalized 5-(methylsulfonyl)-1-phenyl-1*H*-indole derivatives containing different substituted groups. An effective quantitative structure–property relationship (QSPR) models were built using hybrid method combining genetic algorithm (GA) based on stepwise selection multiple linear regression (SWS-MLR) as feature-selection tools and correlation models between each carbon atom of indole derivative and calculated descriptors. Each compound was depicted by molecular structural descriptors that encode constitutional, topological, geometrical, electrostatic, and quantum chemical features. The accuracy of all developed models were confirmed using different types of internal and external procedures and various statistical tests. Furthermore, the domain of applicability for each model which indicates the area of reliable predictions was defined.

1. Introduction

Indole and its derivatives have been a topic of substantial research interest and continue to be one of the most active areas of heterocyclic chemistry for more than 100 years, particularly due to their natural occurrence and pharmacological activities [1]. A large number of indoles myriad derivatives are at the fore as pharmacologically active lead compounds for drug development. Indole derivatives also occur widely distributed in many natural products such as those from plants [2], fungi [3], and marine organisms [4]. The indole nucleus have attracted the attention of organic chemists, medicinal chemists, biologists and pharmacists.

The first preparation of indole dates from 1866 and the Fischer indole synthesis was introduced as the most versatile method for preparing indoles in 1883 [5]. At the moment, there are approximately 1500 indole alkaloids described [6,7]. In 2003, Cruz-López and coworkers prepared a series of 1-substituted 5-(methylsulfonyl)-1*H*-indole-2-carboxylic acid esters, -2-nitriles-, and -2-carboxamides by replacing the benzoyl group of indomethacin with a substituted phenyl group [8] and after that, they synthesized some derivatives of this group with different substituents [9]. They found that substitution on the indole nitrogen atom with the *p*-methylthiophenyl group on *N*-butyl and *N*-pentyl-2-carboxamides provided active and selective COX-2 enzyme inhibitors. Although

* Corresponding author. Tel./fax: +98 871 6624133.
E-mail addresses: rghavami2000@yahoo.com, r.ghavami@uok.ac.ir (R. Ghavami).

0022-2860/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.molstruc.2012.06.042

the synthesis and reactivity of indole derivatives have been a topic of research interest for well over a century, the ^{1}H and ^{13}C unambiguous assignments of the indole nucleus are scanty. Chemical shifts in NMR spectra have for a long time been one of the most important experimental techniques for the characterization of molecules [10]. One method of spectral simulation techniques for the identification of chemical compounds and for the validation of their spectral assignments involves developing mathematical models that relate the ^{13}C chemical shift of an atom to its structural environment. Recently, many linear and non-linear chemometric methods predicting ^{13}C NMR chemical shifts of organic compounds have been developed by means of artificial neural network [11–16] algorithm or multiple linear regression [17–19].

At the present time, a number of modeling approaches are being used in various areas of chemistry, biology, and pharmacy to predict and understand physicochemical, biological, environmental and pharmacological properties of organic compounds [20–23]. The ideal way to find compounds showing the desired chemical properties require tools for predicting the behavior with a reasonable degree of reliability [24]. Foremost among these are various implementations of empirical quantitative structure–property relationships (QSPRs) which provides a promising method for the estimation of the compounds' behavior based on the molecular-level structure features of chemical compounds (geometric and electronic), derived solely from the molecular structure to fit experimental data [25–27]. The main steps of QSPR method include: data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development and finally model performance evaluation [28]. The objective of QSPR analyses is thus to obtain the best possible regressions in terms of statistical quality, chemical meaning, and to use resulting model equations for characterizing and estimating properties of chemical compounds [29].

Nowadays, it is more convenient that a linear relationship between property and descriptors be considered. Multiple linear regression (MLR), principal component regression (PCR) and partial least squares regression (PLSR) are the most widely used linear modeling methods in QSPR [30–32]. Variable selection methods range from simple methods such as stepwise technique processes with forward inclusion or backward elimination [33] to more sophisticated methods such as simulated annealing [34], evolutionary programming [35] and genetic algorithms (GAs) [36]. GA, is a powerful tool which is recently used to optimize many decision making problems [37,38], which were initially proposed by Leardi et al. as strategy for variable subset selection in multivariate calibration analysis [39].

In our previous work, we have reported the application of GA-MLR in predicting ^{13}C NMR chemical shifts of flavonoid derivatives [18]. A central issue of this study was to assess the potential capabilities of GA-MLR algorithm in accurate simulation of ^{13}C NMR spectra of 2-functionalized 5-(methylsulfonyl)-1-phenyl-1*H*-indoles derivatives by using different theoretical molecular descriptors. The validation and predictive ability of the proposed models were examined using several strategies: leave-one-out cross-validation, Y-randomization, and external validation through an odd–even number and division of the entire dataset into training and test sets. Also, the domain of applicability for each model which indicates the area of reliable predictions was defined.

2. Data sets and descriptors preparation

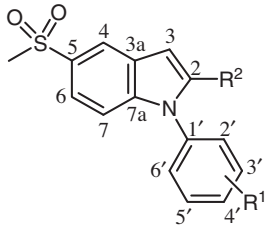
2.1. Computer hardware and software

All calculations were run on an 2.5 GHz Intel® Core™2 Quad Q 8300 CPU with 2 GB of RAM using all four available cores under Windows XP operating system. Molecular modeling and geometry optimization were employed by Hyperchem (version 7.0, HyperCube, Inc.). Dragon software [40] was employed for calculation of theoretical molecular descriptors. SPSS software (version 13.0, SPSS, Inc.) was used for MLR analysis. Genetic algorithm of MATLAB (version 7.0, Math Works, Inc.) is utilized and other calculations were also performed in the MATLAB environment.

2.2. Chemical database sets

Experimentally measured carbon-13 chemical shifts (CS_{exp}) of thirty 5-(methylsulfonyl)-1-phenyl-1*H*-indoles, are belonging to the following four families: (i) 5-(methylsulfonyl)-1-phenyl-1*H*-indoles (Nos. 1–6); (ii) alkyl 5-(methylsulfonyl)-1-phenyl-1*H*-indole-2-carboxylates (Nos. 7–12); (iii) 5-(methylsulfonyl)-1-phenyl-1*H*-indole-2-nitriles (Nos. 13–19); and (iv) 5-(methylsulfonyl)-1-phenyl-1*H*-indole-2-carboxamides (Nos. 20–30), where each containing 14 carbon atoms and consequently a total of 420 chemically non-equivalent carbon centers were obtained from the literature [9]. The chemical structures, numbering and experimental carbon-13 NMR chemical shift values of all 420 atoms in the thirty indole derivatives examined are scheduled in Table 1 and Table S1

Table 1
Molecular structures of thirty 2-functionalized 5-(methylsulfonyl)-1-phenyl-1*H*-indole derivatives used in this study.



Derivative	R^1	R^2
1[c]	H	H
2[p]	*p*-Me	H
3[c]	*p*-Et	H
4[c]	*p*-Pr[n]	H
5[c]	*p*-Cl	H
6[p]	*p*-SMe	H
7[c]	*p*-Me	CO_2Me
8[p]	*p*-Cl	CO_2Me
9[c]	3,5-Cl	CO_2Me
10[c]	*p*-SMe	CO_2Me
11[c]	*p*-SO_2Me	CO_2Me
12[c]	*p*-SMe	CO_2Bu^n
13[p]	*p*-Me	CN
14[p]	*p*-Et	CN
15[c]	*p*-Cl	CN
16[p]	*p*-Br	CN
17[c]	*p*-I	CN
18[c]	*p*-SMe	CN
19[p]	*p*-SO_2Me	$CONHBu^n$
20[c]	*p*-Me	$CONHBu^n$
21[p]	*p*-Cl	$CONHBu^n$
22[p]	3,5-Cl	$CONHBu^n$
23[c]	*p*-OMe	$CONHBu^n$
24[c]	*p*-SMe	$CONHBu^n$
25[c]	*p*-SO_2Me	$CONHBu^n$
26[c]	*p*-SMe	$CONHPr^n$
27[c]	*p*-SMe	$CONHBu^i$
28[c]	*p*-SMe	$CONHPen^n$
29[p]	*p*-SMe	$CONHPen^c$
30[c]	*p*-SMe	$CONHHex^c$

[c] The compounds used in the prediction set.
[p] The compounds used the validation set.