



Genetic algorithm as a variable selection procedure for the simulation of ^{13}C nuclear magnetic resonance spectra of flavonoid derivatives using multiple linear regression

Raof Ghavami*, Amir Najafi, Mohammad Sajadi, Farhad Djannaty

Department of Chemistry, Faculty of Science, University of Kurdistan, P.O. Box 416, Sanandaj, Iran

ARTICLE INFO

Article history:

Received 3 December 2007

Received in revised form 15 March 2008

Accepted 16 March 2008

Available online 25 March 2008

Keywords:

QSPR

^{13}C NMR chemical shift

Nuclear magnetic resonance

MLR

GA-MLR

Flavones

ABSTRACT

In order to accurately simulate ^{13}C NMR spectra of hydroxy, polyhydroxy and methoxy substituted flavonoid a quantitative structure–property relationship (QSPR) model, relating atom-based calculated descriptors to ^{13}C NMR chemical shifts (ppm, TMS = 0), is developed. A dataset consisting of 50 flavonoid derivatives was employed for the present analysis. A set of 417 topological, geometrical, and electronic descriptors representing various structural characteristics was calculated and separate multilinear QSPR models were developed between each carbon atom of flavonoid and the calculated descriptors. Genetic algorithm (GA) and multiple linear regression analysis (MLRA) were used to select the descriptors and to generate the correlation models. Analysis of the results revealed a correlation coefficient and root mean square error (RMSE) of 0.994 and 2.53 ppm, respectively, for the prediction set.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Flavonoids are a large group of polyphenolic compounds of low molecular weight possessing a basic flavan nucleus with two aromatic rings (A and B rings) interconnected by a three-carbon-atom heterocyclic ring (C ring). The most widespread flavonoids contain a double bond between C-2 and C-3 ($\Delta^{2,3}$) and a keto function at C-4 of ring C, which is attached to ring B at C-2 (flavone) or at C-3 (isoflavone). As a result of a number of further modifications on all three rings, particularly on ring C, flavonoids represent one of the largest and the most diverse class of plant secondary metabolites. These compounds are naturally present in vegetables, fruits, chocolate, beverages, seeds, nuts and red wine as well as herbal preparations [1–4]. Flavonoids are often hydroxylated in positions 3, 5, 7, 3', 4' and/or 5'. Frequently, one or more of these hydroxyl groups are methylated, acetylated, prenylated or sulfated [5]. Up to now, more than 4000 different naturally occurring flavonoid compounds are known and new ones are still to be discovered [6–9]. A number of positive health effects of these products such as antioxidant, anticoagulant, antiestrogenic,

antiinflammatory (digestive tract), antimicrobial, or spasmolytic are discussed for these compounds [3,10–12].

Many hydroxylated and methoxylated flavones have been found from natural sources, and whenever they were isolated, they have been identified by instrumental analysis. Because the substitution of hydroxyl and methoxyl groups caused the changes of the ^1H and ^{13}C chemical shifts, the complete ^1H and ^{13}C NMR spectral assignments and structural elucidation of hydroxylated and methoxylated flavones were possible, and these led us to identify these compounds without further experiments [13–17]. If it is possible to predict the carbon chemical shift from the constitution of a molecule quickly and accurately, an automated ranking of the structure generator results becomes possible. Consequently the prediction of ^{13}C chemical shifts plays an important role in structure elucidation of flavonoids [18]. 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 [19–24] algorithm or multiple linear regression [25,26].

Over the past several decades, the quantitative structure–activity/property relationships (QSAR/QSPR) have become an

* Corresponding author. Tel.: +98 871 6624204; fax: +98 871 6660075.

E-mail address: rghavami@uok.ac.ir (R. Ghavami).

important branch of modern chemistry for description and prediction of properties of complex molecular systems in different environments. The success of the QSAR/QSPR approach is critically dependent on the accurate definition and appropriate use of molecular descriptors. Molecular descriptors are numerical values used to describe different characteristics of a certain structure in order to yield information about the property/activity being studied. It must be underlined that a necessary requirement for the application of the QSAR/QSPR approach is the knowledge of the exact chemical constitution and the three-dimensional molecular structure of the chemical compounds being studied.

In most cases, it is more convenient that a linear relationship between activity/property and descriptors is considered. Multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) are the most widely used linear modeling methods in QSPR [27,28]. However, in MLR analysis, the number of compounds in samples is initially required to be at least five times the number of descriptors and the descriptors should be orthogonal [29,30]. One of the main problems encountered in elaborating on large data sets is the detection of the irrelevant variables (i.e., variables containing information) and the elimination of the noise. Therefore, given the correlation of all the calculated descriptors, and the impossibility of performing MLR, a variable selection procedure is needed to select the predictive variables. Variable selection methods range from simple methods such as stepwise technique processes with forward inclusion or backward elimination [31] to more sophisticated methods such as simulated annealing [32], evolutionary programming [33] and genetic algorithms (GAs) [34]. GA, is a powerful tool which is recently used to optimize many decision making problems [35,36], which were initially proposed by Leardi et al. as strategy for variable subset selection in multivariate calibration analysis [37].

In the current work, we will use the strong global search ability of genetic algorithms to optimize QSPR models for predicting ^{13}C chemical shifts values. QSPR analyzes of 50 flavonoid derivatives were performed using different theoretical molecular descriptors by a GA-based MLR. Finally, the prediction capabilities of both approaches are tested unambiguously by application of the models to subsets of compounds excluding the calibration set.

2. Materials and methods

2.1. Software

A Pentium IV personal computer (CPU at 2.6 MB) under Windows XP operating system, was used. Molecular modeling and geometry optimization were employed by Hyperchem (version 7.0, HyperCube, Inc.). Dragon software [38] 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, MathWorks, Inc.) is utilized and other calculations were also performed in the MATLAB environment.

2.2. Experimental database set

Experimentally measured carbon-13 chemical shifts (CS_{exp}) of 50 hydroxy and methoxy substituted flavonoid derivatives, each containing 15 carbon atoms and consequently a total of 750 chemically non-equivalent carbon centers were obtained from the literature [39,40]. This data set was selected in this study because the chemical shifts values were measured under the similar experimental conditions. The range of CS_{exp} values was from 93.7 to 183.2 ppm relative to tetramethylsilane (TMS, 0 ppm) or dimethylsulfoxide (DMSO, 39.50 ppm) as internal reference in deuteride dimethylsulfoxide solvent (CD_3SOCD_3 or $\text{DMSO}-d^6$). The

molecular structure, nomenclature and experimental ^{13}C NMR chemical shift values of all 750 atoms in the 50 flavonoid derivatives examined are listed in Tables 1 and 2, respectively.

2.3. Calculation of structural descriptors

The molecular structures of all the flavonoid derivatives were built with Hyperchem (Version 7, HyperCube, Inc.) software. AM1 semi-empirical calculation was used to optimize the 3D geometry of the molecules. Different quantum chemical descriptors including, heat of formation, dipole moment, HOMO and LUMO energies and their combinations and local charges were calculated by the software. The Polak-Ribier algorithm with root mean squares gradient of 0.1 kcal/mol was selected for optimization. In order to prevent the structures locating at local minima, geometry optimization was run many times with different starting points for each flavonoid derivative. Dragon software was employed to calculate molecular descriptors. Dragon can compute up to 1612 descriptors, which may have very different complexity but can be classified according to their 'dimensionality' in: zero-dimensional (0D) or constitutional, 1D (e.g., empirical descriptors and molecular properties), 2D (such as: 2D autocorrelations, topological indices, BCUT descriptors, Galvez topological charges indices, molecular walk counts) 3D (aromaticity indices, Randic molecular profiles, charge-, geometrical-, RDF-, 3D-MORSE-, GETAWAY-, and WHIM-descriptors) molecular descriptors. A brief definition and description of some of these molecular descriptors used in study are given in Table 3. Thus the initial set of molecular descriptors used as input for the modeling consists of 453 descriptors.

2.4. Selection of structural descriptors

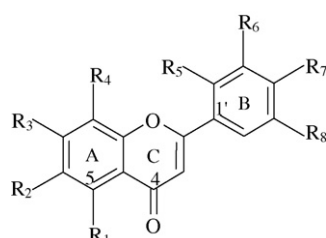
The calculated molecular descriptors were collected in a data matrix (**D**) whose number of rows and columns were the number of molecules and descriptors, respectively. At the beginning, in order to minimize the information overlap in descriptors and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors (NRD) [41] was used in our study. That is, when two descriptors are correlated by a linear correlation coefficient value greater than 0.95, both descriptors are correlated with the dependent variable, the better correlation is used for the actual analysis, leaving out the descriptors showing a lower correlation. This objective-based feature selection left reduced and predictive descriptors for the studied compounds.

Using GA-based MLR feature selection procedures, the dependent variables, i.e., the ^{13}C NMR chemical shifts were used to find subsets of molecular descriptors that provide a good relationship to the carbon-13 chemical shifts. The genetic algorithm used was the same as that previously used [42–44]. In GAs, the initial step is to generate a random population (array), consisting of a predefined number of individuals (rows) and variables (columns). Each individual represents a subset of the original variables within the larger superset of data under analysis. The next step in the GA is analogous to the process of Darwinian evolution whereby, through the processes of crossover, mutation and survival of the fittest, individuals are selected for the next generation until a particular stopping criterion has been reached. The GA uses an algorithm known as a fitness function to assess the robustness of the model proposed by each individual. This usually takes the form of a minimization function; therefore the fittest individuals are those with the lowest fitness value. The population size was varied between 50 and 250 for different GA runs.

Typically, the evolutionary stage of a simple GA proceeds as follows: (1) extract a proportion of the fittest individuals from the current (parent) population, (2) recombine the selected offspring

Table 1

Molecular structures and nomenclature of the hydroxy- and methoxyl-flavone derivatives (validation set in bold) used in this study



Derivative	Nomenclature	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈
1	Flavone	H	H	H	H	H	H	H	H
2	5-Hydroxyflavone	OH	H	H	H	H	H	H	H
3	6-Hydroxyflavone	H	OH	H	H	H	H	H	H
4	7-Hydroxyflavone	H	H	OH	H	H	H	H	H
5	5,7-Dihydroxyflavone/chrysin	OH	H	OH	H	H	H	H	H
6	7,8-Dihydroxyflavone	H	H	OH	OH	H	H	H	H
7	6,7-Dihydroxyflavone	H	OH	OH	H	H	H	H	H
8	5,6,7-Trihydroxyflavone/baicalein	OH	OH	OH	H	H	H	H	H
9	2'-Hydroxyflavone	H	H	H	H	OH	H	H	H
10	7,2'-Dihydroxyflavone	H	H	OH	H	OH	H	H	H
11	5,7,2'-Trihydroxyflavone	OH	H	OH	H	OH	H	H	H
12	4'-Hydroxyflavone	H	H	H	H	H	H	OH	H
13	5,4'-Dihydroxyflavone	OH	H	H	H	H	H	OH	H
14	6,4'-Dihydroxyflavone	H	OH	H	H	H	H	OH	H
15	7,4'-Dihydroxyflavone	H	H	OH	H	H	H	OH	H
16	5,7,4'-Trihydroxyflavone	OH	H	OH	H	H	H	OH	H
17	7,8,4'-Trihydroxyflavone	H	H	OH	OH	H	H	OH	H
18	2',3'-Dihydroxyflavone	H	H	H	H	OH	OH	H	H
19	2',4'-Dihydroxyflavone	H	H	H	H	OH	H	OH	H
20	3',4'-Dihydroxyflavone	H	H	H	H	H	OH	OH	H
21	5,3',4'-Trihydroxyflavone	OH	H	H	H	H	OH	OH	H
22	6,3',4'-Trihydroxyflavone	H	OH	H	H	H	OH	OH	H
23	7,3',4'-Trihydroxyflavone	H	H	OH	H	H	OH	OH	H
24	5,7,3',4'-Tetrahydroxyflavone/luteolin	OH	H	OH	H	H	OH	OH	H
25	7,8,3',4'-Tetrahydroxyflavone	H	H	OH	OH	H	OH	OH	H
26	3',4',5'-Trihydroxyflavone	H	H	H	H	H	OH	OH	OH
27	7,3',4',5'-Tetrahydroxyflavone	H	H	OH	H	H	OH	OH	OH
28	8,3',4',5'-Tetrahydroxyflavone	H	H	H	OH	H	OH	OH	OH
29	5,7,3',4',5'-Pentahydroxyflavone	OH	H	OH	H	H	OH	OH	OH
30	7,8,3',4',5'-Pentahydroxyflavone	H	H	OH	OH	H	OH	OH	OH
31	4'-Methoxyflavone	H	H	H	H	H	H	OCH ₃	H
32	5-Hydroxy-4'-methoxyflavone	OH	H	H	H	H	H	OCH ₃	H
33	6-Hydroxy-4'-methoxyflavone	H	OH	H	H	H	H	OCH ₃	H
34	7-Hydroxy-4'-methoxyflavone/pratol	H	H	OH	H	H	H	OCH ₃	H
35	5,7-Dihydroxy-4'-methoxyflavone/acacetin	OH	H	OH	H	H	H	OCH ₃	H
36	7,8-Dihydroxy-4'-methoxyflavone	H	H	OH	OH	H	H	OCH ₃	H
37	3',4'-Dimethoxyflavone	H	H	H	H	H	OCH ₃	OCH ₃	H
38	5-Hydroxy-3',4'-dimethoxyflavone	OH	H	H	H	H	OCH ₃	OCH ₃	H
39	6-Hydroxy-3',4'-dimethoxyflavone	H	OH	H	H	H	OCH ₃	OCH ₃	H
40	7-Hydroxy-3',4'-dimethoxyflavone	H	H	OH	H	H	OCH ₃	OCH ₃	H
41	5,7-Dihydroxy-3',4'-dimethoxyflavone	OH	H	OH	H	H	OCH ₃	OCH ₃	H
42	6,7-Dihydroxy-3',4'-dimethoxyflavone	H	OH	OH	H	H	OCH ₃	OCH ₃	H
43	7,8-Dihydroxy-3',4'-dimethoxyflavone	H	H	OH	OH	H	OCH ₃	OCH ₃	H
44	3',4',5'-Trimethoxyflavone	H	H	H	H	H	OCH ₃	OCH ₃	OCH ₃
45	5-Hydroxy-3',4',5'-trimethoxyflavone	OH	H	H	H	H	OCH ₃	OCH ₃	OCH ₃
46	6-Hydroxy-3',4',5'-trimethoxyflavone	H	OH	H	H	H	OCH ₃	OCH ₃	OCH ₃
47	7-Hydroxy-3',4',5'-trimethoxyflavone	H	H	OH	H	H	OCH ₃	OCH ₃	OCH ₃
48	5,7-Dihydroxy-3',4',5'-trimethoxyflavone	OH	H	OH	H	H	OCH ₃	OCH ₃	OCH ₃
49	6,7-Dihydroxy-3',4',5'-trimethoxyflavone/prosogerin E	H	OH	OH	H	H	OCH ₃	OCH ₃	OCH ₃
50	7,8-Dihydroxy-3',4',5'-trimethoxyflavone	H	H	OH	OH	H	OCH ₃	OCH ₃	OCH ₃

(crossover), (3) mutate the mated population, (4) assess the newly evolved offspring for fitness, (5) reinsert a proportion of the offspring into the population, replacing the worst parents, and (6) repeat the process until a stopping criterion is reached. The stopping criterion can be defined in many different ways. For example, it could be simply defined as a maximum number of generations, a maximum target outcome value for the fitness, or as a pre-specified number of generations for which the fitness value for the fittest individual has remained constant.

3. Results and discussions

In the current QSPR study we initially calculate a multitude of structural descriptors as mathematical representation of chemical structure. For each of the 50 compounds, sharing 15 similar carbon atoms in their structural backbone, a total of 453 calculated structural features including constitutional descriptors, topological descriptors, geometrical descriptors, quantum chemical descriptors, physicochemical descriptors and liquid properties

Table 2The ^{13}C chemical shift of hydroxylated (1–30) and methoxylated (31–50) flavone derivatives in ppm in $\text{DMSO}-d_6$

Derivative	Position of the carbon														
	2	3	4	5	6	7	8	9	10	1'	2'	3'	4'	5'	6'
1	162.4	106.8	177.0	124.7	125.4	134.1	118.4	155.6	123.3	131.0	126.2	129.0	131.7	129.0	126.2
2	164.1	105.6	183.2	159.8	107.5	135.9	111.0	155.9	110.1	130.5	126.6	129.1	132.3	129.1	126.6
3	162.1	105.9	176.9	107.5	154.9	123.0	119.7	149.3	124.2	131.3	126.1	129.0	131.5	129.0	126.1
4	161.8	102.5	176.3	126.5	115.0	162.7	106.5	157.4	116.1	131.2	126.1	129.0	131.4	129.0	126.1
5	163.0	105.1	181.7	161.4	99.0	164.4	94.0	157.4	103.9	130.7	126.3	129.0	131.8	129.0	126.3
6	161.8	106.0	176.9	115.2	114.1	150.6	131.4	146.7	117.0	133.1	126.3	129.0	131.5	129.0	126.3
7	161.3	103.1	176.2	107.2	144.6	152.3	105.9	150.8	116.0	131.5	125.9	129.0	131.2	129.0	125.9
8	162.9	104.4	182.1	147.0	129.3	153.6	94.0	149.8	104.3	130.9	126.2	129.0	131.7	129.0	126.2
9	160.7	111.0	177.2	124.7	125.2	134.1	118.4	155.8	123.1	117.7	156.6	117.0	132.5	119.4	128.5
10	160.1	110.8	176.6	126.5	114.8	162.6	102.4	157.6	116.0	118.0	156.4	117.0	132.3	119.4	128.5
11	161.3	109.1	181.9	161.4	98.8	164.3	93.9	157.5	109.1	117.3	156.7	117.1	132.8	119.5	128.5
12	163.1	104.9	176.9	125.3	124.8	133.9	118.3	155.4	123.4	121.7	128.4	116.0	161.0	116.0	128.4
13	163.5	103.2	182.9	161.4	107.2	135.5	110.7	155.7	109.8	120.9	128.7	115.9	159.8	115.9	128.7
14	162.6	103.8	176.7	107.5	154.6	122.6	119.5	149.1	124.1	121.7	128.1	115.8	160.7	115.8	128.1
15	162.3	104.4	176.2	126.4	114.7	162.4	102.4	157.3	116.0	121.7	128.0	115.8	160.6	115.8	128.0
16	164.0	102.7	181.6	161.4	98.7	163.6	93.9	157.2	103.6	121.1	128.4	115.8	161.1	115.8	128.4
17	162.2	103.9	176.7	114.9	113.7	150.2	132.9	146.4	116.8	121.9	128.2	115.7	160.6	115.7	128.2
18	161.2	111.0	177.2	124.7	125.2	134.1	118.4	155.9	123.2	118.3	145.4	146.0	117.6	119.2	118.5
19	161.1	109.0	177.2	124.6	125.0	133.8	118.3	155.7	123.2	109.0	158.6	103.3	161.5	108.0	129.8
20	164.2	104.8	176.8	133.9	125.2	135.7	118.7	155.5	123.4	121.9	113.3	145.7	149.4	116.0	124.7
21	164.7	103.3	182.7	159.8	107.1	135.5	110.7	155.7	109.8	121.1	113.5	145.7	150.0	116.0	119.3
22	162.8	103.8	176.6	107.5	154.6	122.6	118.6	149.1	124.1	122.1	113.2	145.6	149.1	115.9	119.5
23	162.5	102.3	176.2	126.4	114.7	162.5	104.4	157.2	116.0	122.1	113.1	145.5	149.0	115.9	118.4
24	163.8	102.8	181.6	161.4	98.7	164.1	93.7	157.2	103.6	121.4	113.2	145.6	149.6	115.9	118.9
25	162.3	103.8	176.6	114.9	113.3	150.1	133.0	146.6	116.9	122.3	113.6	145.6	149.0	115.8	118.6
26	163.6	102.1	176.7	124.3	135.7	137.9	118.0	154.3	124.6	121.1	107.3	145.7	137.9	145.7	107.3
27	163.1	104.8	176.6	126.8	115.1	162.8	102.5	157.6	116.3	121.4	105.7	146.5	137.5	146.5	105.7
28	164.2	103.2	181.6	161.6	99.0	164.2	93.9	157.5	104.0	120.8	106.0	146.5	137.9	146.5	106.0
29	164.0	102.8	181.8	161.4	98.7	162.8	93.7	157.3	103.8	121.6	105.6	146.3	137.8	146.3	105.6
30	162.6	103.9	176.5	114.8	113.6	150.0	133.1	146.7	117.0	121.3	105.6	146.2	137.2	146.2	105.6
31	162.7	105.4	176.9	125.3	124.7	134.1	118.4	155.6	123.2	123.3	128.2	114.6	162.1	114.6	128.2
32	164.2	104.0	183.0	159.9	107.5	135.7	110.9	155.9	109.9	122.6	128.6	114.6	162.6	114.6	128.6
33	162.2	104.4	176.8	105.5	154.7	122.8	119.6	149.2	123.4	124.1	128.0	114.5	161.9	114.5	128.0
34	162.0	102.5	176.2	126.4	114.8	162.6	105.1	157.4	116.1	123.4	127.9	114.5	162.9	114.5	127.9
35	163.1	103.4	181.7	161.4	98.8	164.1	93.9	157.2	103.6	122.7	128.2	114.4	162.2	114.4	128.2
36	164.8	104.6	176.8	115.0	113.8	150.4	133.0	146.5	116.9	123.2	128.1	114.4	161.8	114.4	128.1
37	162.4	105.7	176.9	125.3	124.7	134.0	118.4	155.6	123.3	123.3	109.5	149.0	152.0	111.7	119.9
38	164.1	104.2	183.0	155.7	107.4	135.6	110.8	159.8	109.9	122.6	109.4	148.9	152.3	111.6	120.3
39	162.2	104.7	176.8	107.4	154.7	122.7	119.6	149.2	123.5	124.1	109.2	148.9	151.7	111.6	119.7
40	162.0	102.6	176.3	126.4	114.8	162.6	105.4	157.4	116.1	123.4	109.3	149.0	151.7	111.7	119.4
41	163.2	103.8	181.8	161.4	98.8	164.2	94.0	157.1	103.8	122.8	109.4	148.9	152.1	111.6	120.0
42	162.6	104.9	176.9	119.7	144.3	157.6	107.4	146.2	122.9	128.0	114.2	148.8	151.8	116.8	120.0
43	161.8	104.9	176.8	115.1	113.8	150.5	132.9	146.6	116.9	123.8	109.6	148.9	151.7	111.7	119.9
44	162.3	103.9	177.0	125.4	124.6	134.1	118.6	155.5	123.2	126.3	106.7	153.1	140.4	153.1	106.7
45	163.9	104.4	183.2	155.8	107.6	135.0	110.9	159.8	110.0	125.7	105.5	153.3	141.1	153.3	105.5
46	162.0	104.0	177.0	107.4	154.8	122.9	119.9	149.3	124.2	126.7	105.8	153.2	140.4	153.2	105.8
47	162.6	102.7	176.4	126.4	107.3	161.8	106.5	157.4	114.8	126.2	103.5	153.2	140.4	153.2	103.5
48	162.9	104.9	181.8	161.3	98.9	164.3	94.2	157.3	103.7	125.3	104.0	153.1	140.6	153.1	104.0
49	161.2	103.3	176.2	107.5	144.5	152.2	105.8	150.7	116.0	126.9	103.7	153.1	140.1	153.1	103.7
50	161.6	104.3	176.9	116.8	114.0	150.7	132.8	146.7	115.2	126.8	106.1	153.1	140.3	153.1	106.1

were generated (Table 3). There are 50 types of flavonoid derivatives whose basic structures contain 15 centers of carbons namely from 2 to 10 and 1' to 6' (see Table 1).

At first, multiple linear regression analysis with stepwise selection and elimination of variables was employed to model the ^{13}C chemical shifts with different set of descriptors. However, this procedure did not produce good results. Taking advantage of GA-based MLR feature selection procedures, the best subsets of molecular descriptors that provide a good relationship with the chemical shift for each carbon center were derived.

3.1. GA-MLR modes

For each position of carbon atom center which is labeled 2 to 10 and 1' to 6' in Tables 1 and 3, a subset of most informative molecular descriptors was selected by GA to build the relationship between molecular structure and chemical shift by MLR analysis.

The resulted QSPR models obtained for each carbon atom are given in Table 4. These models contain 54 different molecular descriptors in which the symbols and definitions selected and used in this study are shown in Table 5. Some statistical parameters such as squared correlation coefficients (R^2), root-mean-square error (RMS), relative error of prediction (REP) and Fisher statistic ratio (F) are included in Table 4 for the best fitted equations. As can be seen, for all type of carbon centers the QSPR models presented in Table 4 indicate that the MLR models have good statistical qualities with low prediction error.

3.2. Model prediction-validation

To demonstrate that the resulted models have good prediction abilities for ^{13}C chemical shifts property, the prediction abilities for the external samples have to be further tested with three different methods. In order to assess the predictive ability and to check the

Table 3

Brief description of the molecular descriptors used in this study

Descriptor type	Molecular descriptors
Constitutional	Molecular weight, chemical composition (wt.% of C, H, O, N, S, Cl, F in molecular mass), atom count (C, H, N, S, Cl, F, O), no. of bonds, no. of multiple bonds, no. of aromatic bonds, no. of functional groups (amine, aldehyde, amide, carbonyl, carboxylate, cyano, ether, hydroxyl, methyl, methylene, nitro, nitroso, sulfide, sulfone, sulfoxide and thio), no. of rings, no. of circuits, no. of H-bond donors, no. of H-bond acceptors, chemical composition, etc.
Topological indices	Kier and Hall connectivity indices (χ^0 – χ^2) and valence connectivity indices (χ^{0v} – χ^{2v}), molecular size index, molecular connectivity indices, information contents, total walk count, path/walk-Randic shape indices, Zagreb indices, Schultz indices, Balaban <i>J</i> index, Wiener indices, topological charge indices, topological shape indices (κ^0 – κ^2), etc.
Molecular walk counts	Molecular walk counts of order 1–10, self-re-turning of order 1–10, etc.
Burden eigenvalues	Positive and negative Burden eigenvalues weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, etc.
Two-dimensional autocorrelation	Broto-Moreau autocorrelation of a topological structure, Moran autocorrelation, Geary autocorrelation, H-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, leverage autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, R-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, etc.
Quantum chemical	Highest occupied molecular orbital energy (EHOMO), lowest unoccupied molecular orbital energy (ELUMO), partial charges, etc.
WHIM	Unweighted size, shape, symmetry and accessibility directional indices; size, shape, symmetry and accessibility directional indices weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume; total size, shape symmetry and accessibility indices, etc.
Chemical descriptors	Log <i>P</i> , hydration energy, polarizability, molar refractivity, molecular volume, molecular surface area, polar surface area, parachor, density, molecular mass, surface tension, pK_a , pK_a^0 , etc.
Three-dimensional and geometrical	3-D MorSE signal weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, 3D-Wiener index, solvent accessible surface, molar volume, average geometric distance degree, maximal electrotopological negative/positive variation, etc.

statistical significance of the developed model, Leave-one-out cross-validation (LOO-CV) and external validation (EV) procedures are used.

3.2.1. Leave-one-out cross-validation

In LOO-CV procedure, $n - 1$ samples from a total data set were chosen to construct a QSPR model between the descriptors and the

^{13}C chemical shift values, and the property of the left out sample (removed) was estimated by the designed model. This procedure is repeated until every sample in the total data set is used for a prediction. Then, the prediction residual error sum of squares (PRESS) and the sum of the squared deviation from the mean (SSD) are calculated for each regression equation. The squared correlation coefficient for cross-validation (R_{cv}^2) is then calculated by the

Table 4Multivariate linear regression models and statistical parameters of different ^{13}C NMR chemical shift of flavone derivatives

Position of the carbon	Equation	R^2	RMS	REP	<i>F</i>	R_{cv}^2
C2	$CS_{C2} = 191.7(\pm 3.9) - 46.6(\pm 3.0) R6m+ - 11.6(\pm 1.1) Mor25v + 11.8(\pm 1.4) ATS6m + 48.1(\pm 9.0) H7m - 8.3(\pm 1.3) BEHm8 - 7.7(\pm 2.1) HATS6u$	0.8712	0.38	0.24	48.5	0.8071
C3	$CS_{C3} = 206.6(\pm 10.0) - 83.5(\pm 9.9) SPH - 36.8(\pm 7.8) JG11 + 3.9(\pm 1.0) Mor12e + 11.4(\pm 2.7) Mor23m - 40.7(\pm 12.5) G3s$	0.8673	0.78	0.74	57.5	0.8250
C4	$CS_{C4} = 71.3(\pm 6.3) + 349.8(\pm 20.1) qpos - 1.1(\pm 0.1) RDF070m + 0.008(\pm 0.001) TIE + 2.6(\pm 0.5) Mor04m - 9.0(\pm 1.5) Mor19v$	0.9328	0.68	0.38	122.2	0.8949
C5	$CS_{C5} = 2177.0(\pm 263.4) + 2547.0(\pm 127.1) qpos - 858.9(\pm 79.1) BEHm3 + 0.09(\pm 0.01) TIE + 41.28(\pm 7.05) H4u - 152.38(\pm 36.72) R6u+$	0.9275	5.40	4.08	112.5	0.8758
C6	$CS_{C6} = 2029.1(\pm 138.1) - 1924.8(\pm 123.4) qpos + 98.0(\pm 14.4) ASP + 352.4(\pm 30.1) H5m - 484.1(\pm 47.4) BEHv4$	0.9058	5.35	4.46	108.4	0.8796
C7	$CS_{C7} = 105.9(\pm 7.1) + 9.1(\pm 0.6) RDF070m + 71.8(\pm 6.9) Mor08m + 6.3(\pm 0.7) Mor02m + 20.0(\pm 5.2) R7e + 0.03(\pm 0.03) TIE$	0.9298	3.87	2.64	116.7	0.9067
C8	$CS_{C8} = 497.2(\pm 34.6) - 0.07(\pm 0.01) TIE + 69.6(\pm 5.7) Mor25m - 23.5(\pm 2.0) Mor02v + 2.6(\pm 0.4) RDF045u - 126.3(\pm 50.5) G3p$	0.8828	4.22	3.78	66.5	0.8339
C9	$CS_{C9} = 4.3(\pm 0.2) + 1392.9(\pm 171.0) X4A + 90.9(\pm 9.3) R7m - 65.8(\pm 9.8) Mor30m + 26.0(\pm 3.5) Mor30u - 36.8(\pm 10.0) ATS4m$	0.8504	1.65	1.08	50.0	0.8141
C10	$CS_{C10} = 155.8(\pm 5.3) - 141.0(\pm 15.3) ATS4m + 72.2(\pm 6.0) R6e - 0.02(\pm 0.00) TIE + 25.9(\pm 5.7) Mor09m + 79.2(\pm 14.3) Mor29v$	0.9178	2.06	1.78	97.9	0.8865
C1'	$CS_{C1'} = 334.6(\pm 9.7) - 14.0(\pm 0.7) PCR + 0.3(\pm 0.1) RDF060e - 5.0(\pm 0.7) GATS7e + 5.6(\pm 0.8) Mor04m - 3.4(\pm 0.9) Mor10u$	0.9256	1.20	0.97	109.5	0.8973
C2'	$CS_{C2'} = -127.9(\pm 5.2) - 8.8(\pm 0.4) Mor02m + 27.1(\pm 2.0) RDF035m + 1770.7(\pm 257.2) X3A + 32.5(\pm 5.5) GGI5$	0.9532	3.74	3.10	230.0	0.9429
C3'	$CS_{C3'} = -12.5(\pm 1.6) + 113.4(\pm 4.6) MATS7e + 16.1(\pm 1.1) L2s + 43.0(\pm 5.6) Mor11m + 24.8(\pm 3.4) IC5$	0.9610	3.07	2.28	277.3	0.9513
C4'	$CS_{C4'} = -0.3(\pm 0.1) + 163.1(\pm 9.8) ASP + 24.9(\pm 1.9) GATS7e + 38.9(\pm 5.6) MATS7v + 17.6(\pm 3.9) MATS8e - 23.7(\pm 4.9) Mor32m$	0.9218	3.31	2.27	86.0	0.8981
C5'	$CS_{C5'} = 413.8(\pm 21.2) + 66.3(\pm 5.4) R8u - 68.1(\pm 4.2) IC5 + 66.3(\pm 5.36) MATS7e + 24.6(\pm 2.9) Mor17u + 143.0(\pm 18.9) R6m+$	0.9572	3.09	2.45	197.0	0.9412
C6'	$CS_{C6'} = 143.5(\pm 1.0) - 4.9(\pm 0.4) RDF090m - 223.0(\pm 15.9) R7m+ - 13.4(\pm 2.5) CIC5$	0.9607	1.78	1.48	372.0	0.9491

Table 5

Description of some of the molecular descriptors used in the present study

Symbol	Definition	Class
BEHm3	Highest eigenvalue n. 3 of Burden matrix/weighted by atomic masses	BCUT
BEHm8	Highest eigenvalue n. 8 of Burden matrix/weighted by atomic masses	BCUT
BEHv4	Highest eigenvalue n. 4 of Burden matrix/weighted by atomic van der Waals volumes	BCUT
Qpos	Maximum positive charge	Charge
GGI5	Topological charge index of order 5	Galvez charge
JGI1	Mean topological charge index of order 1	Galvez charge
TIE	E-state topological parameter	Geometrical
SPH	Spherosity	Geometrical
ASP	Asphericity	Geometrical
H4u	H autocorrelation of lag 4/unweighted	GETAWAY
HATS6u	Leverage-weighted autocorrelation of lag 6/unweighted	GETAWAY
H5m	H autocorrelation of lag 5/weighted by atomic masses	GETAWAY
H7m	H autocorrelation of lag 7/weighted by atomic masses	GETAWAY
R8u	R autocorrelation of lag 8/unweighted	GETAWAY
R6u+	R maximal autocorrelation of lag 6/unweighted	GETAWAY
R7m	R autocorrelation of lag 7/weighted by atomic masses	GETAWAY
R6m+	R maximal autocorrelation of lag 6/weighted by atomic masses	GETAWAY
R7m+	R maximal autocorrelation of lag 7/weighted by atomic masses	GETAWAY
R6e	R autocorrelation of lag 6/weighted by atomic Sanderson electronegativities	GETAWAY
R7e	R autocorrelation of lag 7/weighted by atomic Sanderson electronegativities	GETAWAY
RDF045u	Radial Distribution Function – 4.5/unweighted	RDF
RDF035m	Radial Distribution Function – 3.5/weighted by atomic masses	RDF
RDF070m	Radial Distribution Function – 7.0/weighted by atomic masses	RDF
RDF090m	Radial Distribution Function – 9.0/weighted by atomic masses	RDF
RDF060e	Radial Distribution Function – 6.0/weighted by atomic Sanderson electronegativities	RDF
Mor10u	3D-MorSE – signal 10/unweighted	3D-MorSE
Mor17u	3D-MorSE – signal 17/unweighted	3D-MorSE
Mor30u	3D-MorSE – signal 30/unweighted	3D-MorSE
Mor02m	3D-MorSE – signal 02/weighted by atomic masses	3D-MorSE
Mor04m	3D-MorSE – signal 04/weighted by atomic masses	3D-MorSE
Mor08m	3D-MorSE – signal 08/weighted by atomic masses	3D-MorSE
Mor09m	3D-MorSE – signal 09/weighted by atomic masses	3D-MorSE
Mor11m	3D-MorSE – signal 11/weighted by atomic masses	3D-MorSE
Mor23m	3D-MorSE – signal 23/weighted by atomic masses	3D-MorSE
Mor25m	3D-MorSE – signal 25/weighted by atomic masses	3D-MorSE
Mor30m	3D-MorSE – signal 30/weighted by atomic masses	3D-MorSE
Mor32m	3D-MorSE – signal 32/weighted by atomic masses	3D-MorSE
Mor02v	3D-MorSE – signal 02/weighted by atomic van der Waals volumes	3D-MorSE
Mor19v	3D-MorSE – signal 19/weighted by atomic van der Waals volumes	3D-MorSE
Mor25v	3D-MorSE – signal 25/weighted by atomic van der Waals volumes	3D-MorSE
Mor29v	3D-MorSE – signal 30/weighted by atomic van der Waals volumes	3D-MorSE
Mor12e	3D-MorSE – signal 12/weighted by atomic Sanderson electronegativities	3D-MorSE
X3A	Average connectivity index chi-3	Topological
X4A	Average connectivity index chi-4	Topological
IC5	Information content index (neighborhood symmetry of 5-order)	Topological
CIC5	Complementary information content (neighborhood symmetry of 5-order)	Topological
PCR	Ratio of multiple path counts to path counts	Topological
ATS4m	Broto-Moreau autocorrelation of a topological structure – lag 4/weighted by atomic masses	2D autocorrelations
ATS6m	Broto-Moreau autocorrelation of a topological structure – lag 6/weighted by atomic masses	2D autocorrelations
MATS7v	Moran autocorrelation – lag 7/weighted by atomic van der Waals volumes	2D autocorrelations
MATS7e	Moran autocorrelation – lag 7/weighted by atomic Sanderson electronegativities	2D autocorrelations
MATS8e	Moran autocorrelation – lag 8/weighted by atomic Sanderson electronegativities	2D autocorrelations
GATS7e	Geary autocorrelation – lag 7/weighted by atomic Sanderson electronegativities	2D autocorrelations
G3p	3rd component symmetry directional WHIM index/weighted by atomic polarizabilities	WHIM
L2s	2nd component size directional WHIM index/weighted by atomic electrotopological states	WHIM
G3s	3rd component symmetry directional WHIM index/weighted by atomic electrotopological states	WHIM

following equation:

$$\text{PRESS} = \sum_{i=1}^n \frac{(y_{i,\text{obs}} - y_{i,\text{pred}})^2}{n-1} \quad (1)$$

$$\text{SSD} = \sum_{i=1}^n \frac{(y_{i,\text{obs}} - y_{i,\text{avg}})^2}{n-1} \quad (2)$$

$$R_{\text{cv}}^2 = 1 - \frac{\sum_{i=1}^n (y_{i,\text{obs}} - y_{i,\text{pred}})^2}{\sum_{i=1}^n (y_{i,\text{obs}} - y_{i,\text{avg}})^2} = 1 - \frac{\text{PRESS}}{\text{SSD}} \quad (3)$$

where n is the number of compounds included in the model, $y_{i,\text{obs}}$, and $y_{i,\text{pred}}$ are the experimental, and predicted chemical shifts of the left-out compound i , respectively and $y_{i,\text{avg}}$ is the average

experimental chemical shift of left-in compounds different from i . The R_{cv}^2 values can be considered as a measure of the predictive power of a model: whereas R^2 can always be increased artificially by adding more parameters, R_{cv}^2 decreases if a model is over-parameterized [45], and is therefore a more meaningful summary statistic for predictive models. The correlation coefficients for each subset are presented in last column of Table 4. The cross-validation results show that the models have PRESS/SSD ratios less than 0.2 or R_{cv}^2 values greater than 0.807.

3.2.2. Odd–even external validation

To further check the prediction ability of the resulting models, external validation (EV) was also employed. In the EV procedure applied here, the method proposed by Hawkins and

Table 6

Statistical parameters of the overfitting and predictive ability of the models

Position of the carbon	Odd samples				Even samples			
	RMSE _{RS}	R _{RS} ²	RMSE _{HO}	R _{HO} ²	RMSE _{RS}	R _{RS} ²	RMSE _{HO}	R _{HO} ²
C2	0.25	0.8632	0.28	0.8254	0.20	0.9005	0.25	0.8530
C3	0.59	0.9036	0.88	0.7942	0.76	0.8759	0.94	0.8121
C4	0.19	0.9832	0.45	0.9332	0.42	0.9200	0.77	0.7450
C5	3.27	0.9532	4.97	0.8965	4.07	0.9285	5.50	0.8709
C6	4.91	0.8862	5.88	0.8621	3.48	0.9426	4.36	0.9145
C7	2.53	0.9350	3.47	0.8843	1.94	0.9624	4.05	0.8630
C8	3.38	0.9030	6.25	0.6787	2.97	0.9299	4.85	0.8398
C9	0.98	0.8766	1.32	0.8434	1.02	0.8646	1.28	0.8017
C10	1.02	0.9734	2.40	0.8558	1.99	0.8958	2.50	0.8419
C1'	0.95	0.9397	1.31	0.9002	0.84	0.9326	1.11	0.9004
C2'	3.06	0.9335	3.47	0.9150	1.84	0.9781	2.38	0.9655
C3'	2.46	0.9567	3.00	0.9373	1.65	0.9785	2.64	0.9561
C4'	1.97	0.9450	2.51	0.9121	2.31	0.9120	2.73	0.8825
C5'	2.47	0.9565	3.09	0.9359	2.10	0.9688	2.77	0.9487
C6'	0.99	0.9825	1.41	0.9768	1.70	0.9480	2.06	0.9429

Table 7

Equations based on only the 36 carbon atoms in calibration sets for each position of the carbon

Position of the carbon	Equations calibration sets
C2	CS _{C2} = 190.6(±4.8) – 45.3(±3.7) R6m+ – 11.4(±1.3) Mor25v + 11.5(±1.9) ATS6m + 47.1(±11.6) H7m – 7.9(±1.6) BEHm8 – 7.4(±2.9) HATS6u
C3	CS _{C3} = 204.9(±10.1) – 79.9(±9.8) SPH – 47.0(±8.3) JG1 + 4.6(±1.0) Mor12e + 11.0(±2.4) Mor23m – 42.0(±15.7) G3s
C4	CS _{C4} = 70.0(±7.7) + 354.3(±24.6) qpos – 1.1(±0.1) RDF070m + 0.01(±0.00) TIE + 2.6(±0.6) Mor04m – 10.4(±1.8) Mor19v
C5	CS _{C5} = 2214.6(±335.4) + 2521.2(±172.0) qpos – 867.1(±101) BEHm3 + 0.09(±0.01) TIE + 39.7(±9.8) H4u – 154.6(±53.2) R6u+
C6	CS _{C6} = 2129.4(±182) – 1924.4(±159.2) qpos + 102.3(±17.7) ASP + 365.2(±37.7) H5m – 520.7(±62.9) BEHv4
C7	CS _{C7} = 102.4(±10.2) + 8.7(±0.7) RDF070m + 69.4(±9.5) Mor08m + 6.3(±1.1) Mor02m + 19.2(±6.7) R7e + 0.03(±0.01) TIE
C8	CS _{C8} = 483.4(±48.7) – 0.07(±0.01) TIE + 67.4(±7.0) Mor25m – 22.4(±2.8) Mor02v + 2.7(±0.6) RDF045u – 147.4(±62.4) G3p
C9	CS _{C9} = 32.5(±2.2) + 1147.1(±175.2) X4A + 91.4(±9.5) R7m – 69.2(±10.2) Mor30m + 23.7(±3.4) Mor30u – 36.1(±10.0) ATS4m
C10	CS _{C10} = 153.6(±7.3) – 136.1(±19.4) ATS4m + 69.8(±7.5) R6e – 0.02(±0.01) TIE + 25.5(±7.2) Mor09m + 83.8(±18.5) Mor29v
C1'	CS _{C1'} = 332.8(±12.0) – 13.9(±0.8) PCR + 0.3(±0.1) RDF060e – 5.1(±0.9) GATS7e + 5.3(±1.0) Mor04m – 2.9(±1.1) Mor10u
C2'	CS _{C2'} = –145.0(±6.3) – 8.7(±0.6) Mor02m + 28.0(±2.4) RDF035m + 1852.7(±314.2) X3A + 30.7(±7.0) GGI5
C3'	CS _{C3'} = –15.9(±1.9) + 111.9(±5.7) MATS7e + 16.1(±1.3) L2s + 42.1(±6.9) Mor11m + 25.4(±4.2) IC5
C4'	CS _{C4'} = –8.0(±0.8) + 170.2(±11.0) ASP + 25.5(±2.4) GATS7e + 42.0(±6.9) MATS7v + 17.9(±4.3) MATS8e – 30.0(±11.6) Mor32m
C5'	CS _{C5'} = 390.8(±25.5) + 73.1(±6.3) R8u – 63.9(±5.1) IC5 + 56.9(±6.4) MATS7e + 21.9(±3.8) Mor17u + 120.9(±22.7) R6m+
C6'	CS _{C6'} = 143.5(±1.4) – 4.9(±0.6) RDF090m – 223.6(±21.6) R7m+ – 13.6(±3.1) CIC5

Table 8

Statistical parameters of the QSPR models obtained using different molecular descriptors

Position of the carbon	Calibration set					Prediction set		
	RMS	REP	F	R ²	R _{cv} ²	RMS	REP	R ²
C2	0.41	0.25	29.68	0.8600	0.7221	0.32	0.20	0.9309
C3	0.64	0.61	57.90	0.9062	0.8600	1.10	1.05	0.7871
C4	0.72	0.40	71.21	0.9231	0.8663	0.59	0.33	0.9526
C5	6.22	4.73	64.21	0.9146	0.8427	2.32	1.74	0.9805
C6	5.94	4.86	74.60	0.9058	0.8696	3.78	3.32	0.8377
C7	4.23	2.92	54.69	0.9194	0.8849	3.06	2.02	0.9641
C8	4.62	4.12	38.14	0.8640	0.7874	3.39	3.06	0.9237
C9	1.39	0.91	47.54	0.8866	0.8437	2.48	1.60	0.7107
C10	2.26	1.95	60.77	0.9104	0.8688	1.59	1.38	0.9608
C1'	1.30	1.05	72.45	0.9241	0.8778	0.94	0.76	0.9347
C2'	3.30	2.72	150.80	0.9509	0.9340	2.90	2.44	0.9618
C3'	3.23	2.41	178.85	0.9584	0.9426	2.68	1.97	0.9698
C4'	3.30	2.27	97.04	0.9267	0.8940	3.62	2.47	0.9215
C5'	2.99	2.37	146.69	0.9608	0.9407	3.77	3.05	0.9382
C6'	0.41	0.25	29.68	0.8600	0.9360	0.32	0.20	0.9309

Table 9Statistical parameters obtained using GA-MLR model for prediction ¹³C chemical shifts of flavone derivatives

	R ²	RMS	REP	R _{cv} ²	N ^a
Total	0.9821	3.09	2.32	0.9750	750
Calibration set	0.9790	3.35	2.51	0.9675	540
Prediction set	0.9882	2.53	1.91	–	210

^a N denotes of the number of carbon atom centers.

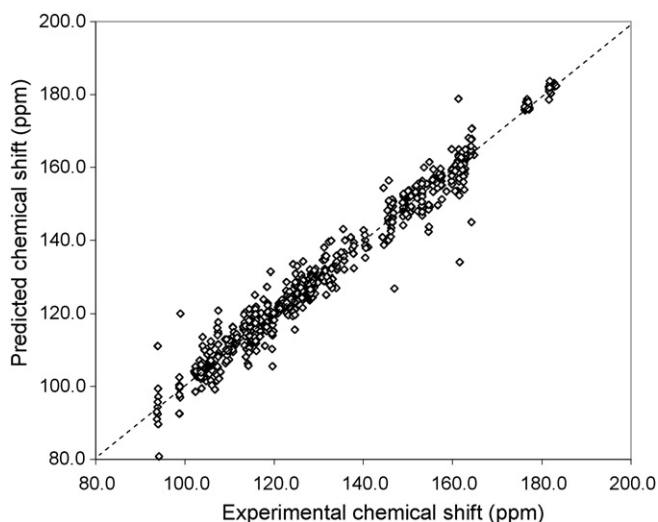


Fig. 1. Plot of the predicted chemical shifts by GA-MLR (LOO-CV) for the total data set of 50 flavone derivatives used in this study against the experimental values. The dotted line is the ideal fit to the straight line.

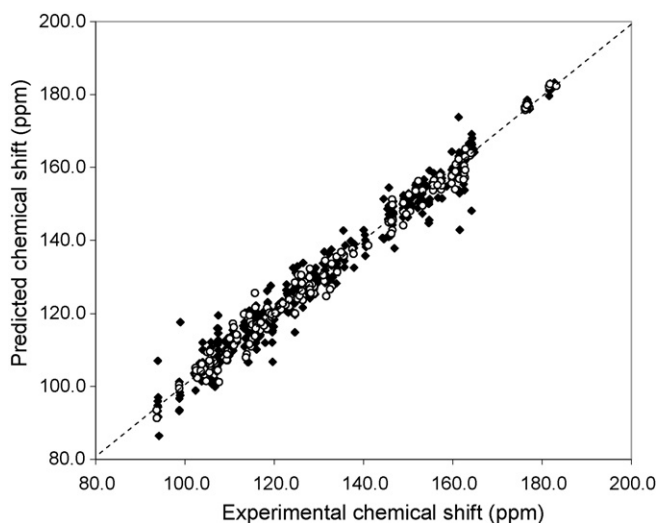


Fig. 2. Plot of the predicted chemical shifts by GA-MLR for calibration set (■) and validation set (□) against the experimental values. The dash line is the ideal fit to the straight line.

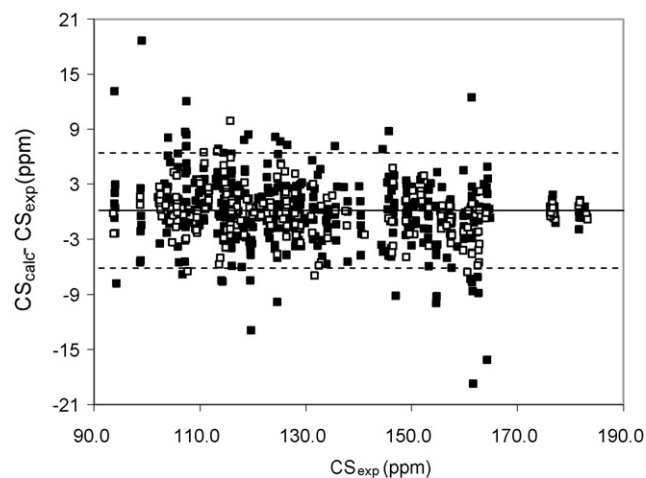


Fig. 3. Calculated errors vs. experimental values of flavone derivatives carbon-13 chemical shifts (ppm) for the calibration set (■) and prediction set (□) using GA-MLR model.

co-workers [45,46] was used. To do this, the data were sorted in the ascending order of chemical shift values and then divided into two sets namely odd-number and even-number chemical shifts. This way of splitting ensures that the distribution of chemical shift values of the two subsets were very similar. The QSPR models were fitted to the odd-number and even-number samples separately and the resulted fitness were assessed by applying QSPR models to both samples. To compare the estimation abilities of the models, two statistical parameters namely root mean squares error (RMSE) and R^2 , were calculated. The same data set (i.e. 'calibration set') that was already used to fit the models was employed to determine resubstitution

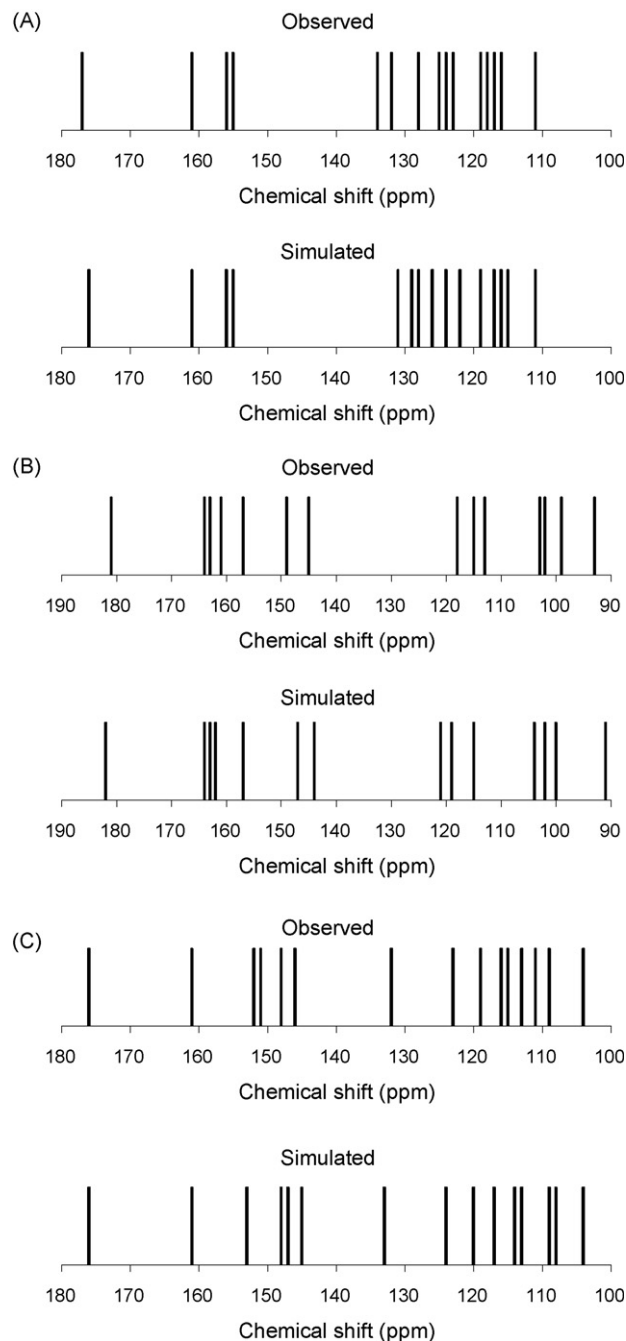


Fig. 4. Simulated and observed spectra of three molecules in the prediction set (A) 2'-hydroxyflavone, (B) 5,7,3',4'-tetrahydroxyflavone, and (C) 7,8-dihydroxy-3',4'-dimethoxyflavone.

parameters, i.e. $RMSE_{RS}$ and R_{RS}^2 and to determine holdout parameters, i.e. $RMSE_{HO}$ and R_{HO}^2 for the other data set which was not involved in the fitting. The resubstitution statistical parameters of the samples bases their predictions on the regression fitted to those samples, and holdout statistical parameters bases their predictions on the regression fitted to the other samples. Table 6 summarizes these statistical parameters archived by this approach. As can be seen, in the odd- and even-number samples, the resubstitution and holdout

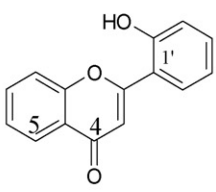
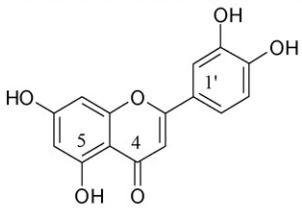
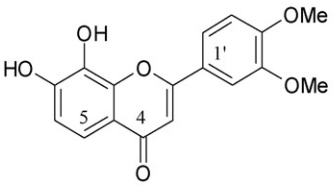
RMSE are very similar, indicating that same sample and other sample predictions are equally precise.

3.2.3. Calibration set and prediction set

In addition to the traditional LOO-CV and odd–even external validation, further attempts were made to examine the quality of the resulted models. In this case, before each training run, all data sets were split randomly into two separate sections: the calibration and external prediction sets. Out of 50 compounds, 36 compounds

Table 10

Experimental and calculated values of chemical shifts for the molecules included in the prediction set

Prediction set	Carbon position	Experimental (ppm)	GA-MLR predicted (ppm)	$\Delta(\text{Exp.}-\text{Cal.})$
 (A)	2	160.7	160.8	0.1
	3	111.0	111.3	0.3
	4	177.2	176.3	-0.9
	5	124.7	128.3	3.6
	6	125.2	124.1	-1.1
	7	134.1	131.5	-2.6
	8	118.4	115.7	-2.7
	9	155.8	156.5	0.7
	10	123.1	122.2	-0.9
	1'	117.7	116.7	-1.0
	2'	156.6	155.1	-1.5
	3'	117.0	117.5	0.5
	4'	132.5	126.7	-5.8
	5'	119.4	119.1	-0.3
	6'	128.5	128.0	-0.5
 (B)	2	163.8	163.6	-0.2
	3	102.8	102.5	-0.3
	4	181.6	182.2	0.6
	5	161.4	162.2	0.8
	6	98.7	100.4	1.7
	7	164.1	163.9	-0.2
	8	93.7	91.3	-2.4
	9	157.2	157.8	0.6
	10	103.6	104.3	0.7
	1'	121.4	121.4	0.0
	2'	113.2	119.8	6.6
	3'	145.6	144.9	-0.7
	4'	149.6	147.2	-2.4
	5'	115.9	115.9	0.0
	6'	118.9	120.2	1.3
 (C)	2	161.8	161.9	0.1
	3	104.9	104.5	-0.4
	4	176.8	176.5	-0.3
	5	115.1	113.1	-2.0
	6	113.8	108.8	-5.0
	7	150.5	148.0	-2.5
	8	132.9	133.5	0.6
	9	146.6	145.5	-1.1
	10	116.9	117.0	0.1
	1'	123.8	124.2	0.4
	2'	109.6	109.6	0.0
	3'	148.9	147.9	-1.0
	4'	151.7	153.9	2.2
	5'	111.7	113.9	2.2
	6'	119.9	120.3	0.4

(A) 2'-hydroxyflavone, (B) 5,7,3',4'-tetrahydroxyflavone, and (C) 7,8-dihydroxy-3',4'-dimethoxyflavone.

were used for calibration set and 14 compounds were used as external validation. The compounds that constituted the calibration and validation sets are clearly presented in Table 2. The validation examples are marked as bold font. The calibration set was used to obtain the best fit equation of MLR and the validation set was used to monitor overfitting the MLR models. The resulted GA-MLR models for 36 compounds were the same as those obtained for the entire set of 50 flavonoid derivatives in each subset subject to using descriptors of the 50 compounds models to provide for prediction set of 14 compounds. Each of related calibration set equations is indexed in Table 7, and those statistical parameters for the best-fitted models are also presented in Table 8.

3.2.4. Overall prediction ability

Each individual flavonoid compound considered here has 15 carbon atom centers, that is, altogether we have 750 carbon atoms. In order to obtain an overall estimate about the prediction ability of the resulted QSPR models, we plotted all cross-validated predicted values of ^{13}C chemical shifts of different carbon atoms against the experimental values in a single graph. The corresponding statistical parameters are calculated and results are presented in Table 9. The statistical results indicate that the overall model has the capability of predicting and has low level space. These results are presented in Fig. 1. Based on the models introduced in the previous stages, all 50 compounds were divided into two parts, i.e., calibration and prediction sets, which contained 36 (including 540 carbon centers) and 14 (including 210 carbon centers) compounds, respectively. The plot of predicted chemical shift against the corresponding experimental values for all carbon centers of the prediction set molecules is represented in Fig. 2 and the corresponding statistical quantities are listed in Table 9. The achieving respective values of 0.9882, 2.523, and 1.905 ppm for parameters R^2 , RMS and REP demonstrate the high prediction power of the proposed models.

The distribution of calculation errors vs. experimental values of chemical shifts are plotted in Fig. 3. As is observed, some carbon atom centers represent significant deviation from the regression line, where the largest overestimations of chemical shifts 18.7 ppm are observed for the calibration carbon atoms C6 of 8,3',4',5'-tetrahydroxyflavone, and the largest underestimations –18.7 ppm for the calibration carbon atoms C4 of 8,3',4',5'-tetrahydroxyflavone, C2 of 8,3',4',5'-tetrahydroxyflavone (–16.1 ppm), as well as for the prediction carbon atom centers C3' of 7,8,4'-trihydroxyflavone (9.9 ppm). The two dotted lines above and below the zero line indicate a double standard deviation of regression ($\pm 2 \times 3.14$), and roughly 94.13% of the residuals are lied between these two lines. Since no distinct pattern found, it is turned out that no relationship exists between the residual values and the calculated carbon-13 chemical shift values.

The accuracy of the models that have been proposed can be seen from Fig. 4, which compares the simulated and observed spectra of three molecules, i.e., 2'-hydroxyflavone, 5,7,3',4'-tetrahydroxyflavone, and 7,8-dihydroxy-3',4'-dimethoxyflavone in the prediction set. An outstanding visual similarity exists between the simulated and observed spectra of the three molecules in the prediction set. The overall spectral errors between the two spectra for these compounds were –0.807, 0.407 and –0.420 ppm, respectively. The differences between the experimental and calculated values of the chemical shifts of these molecules are also given in Table 10 for the carbon centers.

4. Conclusions

Although numerous ^{13}C NMR spectra are being generated each day, their interpretation has become a hindrance to progress in the

identification process. One way to deal with this difficulty is developing computer-assisted models. In this work, a novel QSPR tool, GA-MLR that combines a genetic algorithm with multiple linear regression is performed to relate the structural parameters of 50 hydroxy, polyhydroxy and methoxy substituted flavonoid derivatives data set to their ^{13}C NMR spectra. MLR with GA produced more predictive, informative and significantly improved QSPR models. The use of physicochemical, topological, and geometric descriptors was revealed to be quite a successful strategy. The effectiveness of the evolutionary programming algorithm is demonstrated by the selection of the best set of molecular descriptors. The validation and predictive ability of the models were examined by both the leave-one-out cross-validation and external validation. Both methods indicated that the resulting multiparametric QSPR models possess high prediction ability and low overfitting. The high correlation coefficient of 0.9982 and low root mean squares error of 2.53 ppm and relative prediction error of 1.91 ppm for the prediction set reveals an excellent prediction ability of the generated model for ^{13}C chemical shifts of flavonoid derivatives.

References

- [1] B.J. Halla, M. Chebib, J.R. Hanrahan, G.A.R. Johnstone, 6-Methylflavanone, a more efficacious positive allosteric modulator of γ -aminobutyric acid (GABA) action at human recombinant $\alpha_2\beta_2\gamma_{2L}$ than at $\alpha_1\beta_2\gamma_{2L}$ and $\alpha_1\beta_2$ GABA_A receptors expressed in xenopus oocytes, *Eur. J. Pharmacol.* 512 (2005) 97–104.
- [2] J.V. Formica, W. Regelson, Review of the biology of quercetin and related bioflavonoids, *Food Chem. Toxicol.* 33 (1995) 1061–1080.
- [3] E. Haslam, *Practical Polyphenolics from Structure to Molecular Recognition and Physiological Action*, Cambridge University Press, Cambridge, UK, 1998.
- [4] A. Scalbert, G. Williamson, Dietary intake and bioavailability of polyphenols, *J. Nutr.* 130 (2000) 2073S–2085S.
- [5] E.De. Rijke, P. Out, W.M.A. Niessen, F. Ariese, C. Gooijer, U.A.Th. Brinkman, Analytical separation and detection methods for flavonoids, *J. Chromatogr. A* 1112 (2006) 31–63.
- [6] S.B. Lotito, B. Frei, Consumption of flavonoid-rich foods and increased plasma antioxidant capacity in humans: cause, consequence, or epiphenomenon? *Free Radic. Biol. Med.* 41 (2006) 1727–1746.
- [7] S.M. Lai, R.L. Chen, S.Y. Suen, Adsorption separation for the extracts from Ginkgo Biloba leaves using intermediate polarity resins, *J. Liq. Chromatogr. Rel. Technol.* 26 (2003) 2941–2960.
- [8] B.H. Havsteen, The biochemistry and medical significance of flavonoids, *Pharmacol. Ther.* 96 (2002) 67–202.
- [9] K. Wähälä, S. Rasku, K. Parikka, Deuterated phytoestrogen flavonoids and iso-flavonoids for quantitation, *J. Chromatogr.* 77 (2002) 111–122.
- [10] J.B. Harborne, C.A. Williams, Advances in flavonoid research since 1992, *Phytochemistry* 55 (2000) 481–504.
- [11] B. Klejdus, D. Sterbova, P. Stratil, V. Kuban, Identification and characterization of isoflavones in plant material by HPLC-DADMS tandem, *Chem. Listy* 97 (2003) 530–539.
- [12] C.A. Williams, R.J. Grayer, Anthocyanins and other flavonoids, *Nat. Prod. Rep.* 21 (2004) 539–573.
- [13] P. Joseph-Nathan, J. Mares, M.C. Hernandez, J.N. Schoolery, Proton and carbon-13 nuclear magnetic resonance studies of flavones and deuterated analogs, *J. Magn. Reson.* 16 (1974) 447–453.
- [14] V.M. Chari, H. Wagner, A. Neszmelyi, in: *Proceedings of the 5th Hungarian Bioflavonoid Symposium, Marafured, Hungary*, 1977.
- [15] K.R. Markham, V.M. Chari, in: J.B. Harborne, T.J. Mabry (Eds.), *The Flavonoids: Advances in Research*, Chapman and Hall, 1982.
- [16] A.De. Sousa, J.M.C. Hemmer, J. Gasteiger, Prediction of H-1 NMR chemical shifts using neural networks, *Anal. Chem.* 74 (2002) 80–90.
- [17] *Dictionary of Natural Products on CD-ROM*, Chapman & Hall/CRC Press, London, 2000.
- [18] D.C. Burns, D.A. Ellis, R.E. March, A predictive tool for assessing ^{13}C NMR chemical shifts of flavonoids, *Magn. Reson. Chem.* 45 (2007) 835–845.
- [19] M. Jalali-Heravi, P. Shahbazzikah, B. Zekavat, M.S. Ardejani, Principal component analysis-ranking as a variable selection method for the simulation of ^{13}C nuclear magnetic resonance spectra of xanthenes using artificial neural networks, *QSAR Comb. Sci.* 26 (2007) 764–772.
- [20] D. Svozil, J. Pospichal, V. Kvasnicka, Neural-network prediction of C-13 NMR chemical-shifts of alkanes, *J. Chem. Inf. Comput. Sci.* 35 (1995) 924–928.
- [21] P.C. Jurs, D.L. Cluser, The simulation of ^{13}C nuclear magnetic resonance spectra of dibenzofurans using multiple linear regression analysis and neural networks, *Anal. Chim. Acta* 321 (1996) 127–135.
- [22] O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye, J.P. Doucet, ^{13}C NMR chemical shift prediction of the sp^3 carbon atoms in the α position relative to the double bond in acyclic alkenes, *J. Chem. Inf. Comput. Sci.* 37 (1997) 587–598.

- [23] B.E. Mitchell, D.C. Jurs, Computer assisted simulation of ^{13}C nuclear magnetic resonance spectra of monosaccharides, *J. Chem. Inf. Comput. Sci.* 36 (1996) 58–64.
- [24] J. Meiler, W. Maier, M. Will, R. Meusinger, Using neural networks for ^{13}C NMR chemical shift prediction-comparison with traditional methods, *J. Magn. Reson.* 157 (2002) 242–252.
- [25] G. Liang, H. Mei, Y. Zhou, P. Zhou, Z. Li, Simulation of ^{13}C nuclear magnetic resonance spectra for derivatives of bases and nucleotides, *Chinese J. Anal. Chem.* 34 (2006) 329–333.
- [26] M. Jaiswal, P. Khadikar, QSAR study on ^{13}C NMR chemical shifts on carbinol carbon atoms, *Bioorg. Med. Chem.* 12 (2004) 1793–1798.
- [27] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemomet. Intell. Lab. Syst.* 58 (2001) 109–130.
- [28] K. Tang, T. Li, Combining PLS with GA-GP for QSAR, *Chemomet. Intell. Lab. Syst.* 64 (2002) 55–64.
- [29] Y. Miyashita, Z. Li, S. Sasaki, Chemical pattern recognition and multivariate analysis for QSAR Studies, *TRAC Trends Anal. Chem.* 12 (1993) 50–60.
- [30] M. Randic, Resolution of ambiguities in structure–property studies by use of orthogonal descriptors, *J. Chem. Inf. Comput. Sci.* 31 (1991) 311–320.
- [31] D.C. Montgomery, E.A. Peck, *Introduction to Linear Regression Analysis*, Wiley, New York, 1982.
- [32] Ž.J.H. Kalivas, N. Roberts, J.M. Sutter, Global optimization by simulated annealing with wavelength selection for ultraviolet–visible spectrophotometry, *Anal. Chem.* 61 (1989) 2024–2030.
- [33] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvao, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, *Chemomet. Intell. Lab. Syst.* 57 (2001) 65–73.
- [34] L. Davis, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, N.Y. (USA), 1991.
- [35] J. Devillers, *Genetic Algorithms in Molecular Modelling*, Academic Press Ltd., London, 1996.
- [36] A. Böcker, G. Schneider, A. Teckentrup, Status of HTS data mining approaches, *QSAR Comb. Sci.* 23 (2004) 207–213.
- [37] R. Leardi, R. Boggia, M. Terrile, Genetic algorithm as a strategy for feature selection, *J. Chemom.* 6 (1992) 267–281.
- [38] Dragon Academic version is a product by Milano Chemometrics and QSAR research group, Milano, Italy.
- [39] Y. Park, B.-Ho. Moon, E. Lee, Y. Lee, J.-H. Ahn, Y. Lim, *Magn. Reson. Chem.* 45 (2007) 674–679.
- [40] S. Perruchon, *Synthese und Struktur-Aktivitäts-Beziehungen von Flavonoiden*, Ph.D. Thesis, der Technischen Universität Darmstadt, Germany, 2004, <http://deposit.d-nb.de/cgi-bin/dokserv?idn=97036184x> and <http://elib.tu-darmstadt.de/diss/000409>.
- [41] J. Olivero, T. Garcia, P. Payares, R. Vivas, D. Diaz, E. Daza, P. Geerliger, Molecular structure and gas chromatographic retention behavior of the components of ylang–ylang oil, *J. Pharm. Sci.* 86 (1997) 625–630.
- [42] B. Hemmateenejad, Correlation ranking procedure for factor selection in PC-ANN modeling and application to ADMETox evaluation, *Chemomet. Intell. Lab. Syst.* 75 (2005) 231–245.
- [43] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, Application of ab initio theory for the prediction of acidity constants of some 1-hydroxy-9,10-anthraquinone derivatives using genetic neural network, *J. Mol. Struct. (Theochem)* 635 (2003) 183–190.
- [44] R. Ghavami, A. Najafi, B. Hemmateenejad, Chemometrics-assisted spectrophotometric methods for simultaneous determination and complexation study of Fe(III), Al(III) and V(V) with morin in micellar media, *Spectrochim. Acta Part A: Mol. Biomol. Spec.*, in press.
- [45] D.M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci.* 44 (2004) 1–12.
- [46] D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation, *J. Chem. Inf. Comput. Sci.* 43 (2003) 579–586.