Chemical Methodologies

journal homepage: http://chemmethod.com/

Original Research article

Approach to Chemometrics Models by Artificial Neural Network for Structure: First Applications for Estimation Retention Time of Doping Agent

Mehrdad Shahpar[a] *, Sharmin Esmaeilpoor[b]

[a] Director of Ilam Petrochemical Company

[b] Department of Chemistry, Payame Noor University, P.O. BOX 19395-4697, Tehran, Iran

ARTICLE INFORMATION

Received: 2 August 2017
Received in revised: 29 August 2017
Accepted: 10 September2017
Available online: 20 November 2017

DOI:
10.22631/chemm.2017.96397.1008

KEYWORDS

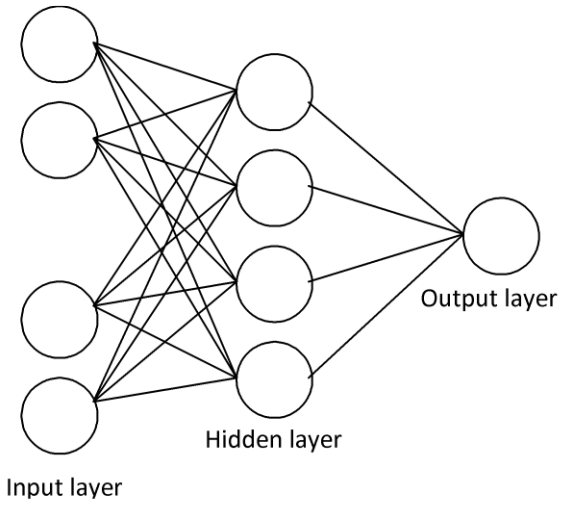
Doping agents
Ultra-high-pressure liquid chromatogram
QSRR
Genetic algorithms

ABSTRACT

A quantitative structure–retention relationship (QSRR), was developed by using the genetic algorithm-partial least square (GA-PLS), Kernel partial least square (GA-KPLS) and Levenberg-Marquardt artificial neural network (L-M ANN) approach for the prediction of the retention time (RT) of the doping agents in urine. The values of the retention time were obtained by using ultra-high-pressure liquid chromatography–quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS). A suitable set of the molecular descriptors was calculated and the important descriptors were selected by the aid of the GA-PLS and GA-KPLS. By comparing the results, GA-KPLS descriptors are selected for L-M ANN. Finally a model with a low prediction error and a good correlation coefficient was obtained by L-M ANN. This model was used to predict the RT values of some of doping agents which were not used in the modeling procedure. This is the first research on the QSRR of doping agents against the RT using the GA-PLS, GA-KPLS and L-M ANN model.

Corresponding author:email:shahpar2012@gmail.con, Director of Ilam Petrochemical Company, Tel:+98 9143100801

Graphical Abstract



Introduction

A win at all costs ethos that undermines the integrity of sport has entered the arena and a new game is at stake, the dangerous and sometimes deadly game of doping. Doping in sport is not a new phenomenon; athletes have taken performance-enhancing agents since the beginning of time. Doping not only contravenes the spirit of fair competition, it can be seriously detrimental to health. Anabolic steroids affect the cardiovascular and mental health and are associated with an increased risk of neoplasms [1, 2]. Dietary supplements containing ephedra alkaloids have been linked to serious health risks including hypertension, tachycardia, stroke, seizures, and death. Deaths under the influence of drugs and combinations thereof are not uncommon in sport. The peptide hormones or so-called "sports-designer drugs" are thought to be the most dangerous; although, the combination of amphetamines, anabolic steroids or anti hypertensives combined with intense exertion in athletes are just as hazardous [3].

The banned substances and techniques fall into the following categories: androgens, blood doping, peptide hormones, stimulants, diuretics, narcotics, and cannabinoids from the World Anti-Doping Agency (WADA) prohibited list. Also, substances in the banned list may be restricted according to the route, sport and governing body regulations. For instance, steroid inhalers and beta-agonist inhalers are mostly permitted with prior written notification but are banned orally. Bambuterol, fenoterol, and reproterol are banned completely, regardless of route, as is the vetinerary beta-agonist clenbuterol. Similarly, steroids are permitted with notification by intraarticular

administration, but they are banned intramuscularly or intravenously. Beta-blockers are banned in control sports only, such as archery, shooting, bobsleigh, snooker, darts, and synchronised swimming. Alcohol (ethanol) is banned in sports such as motor-racing and shooting where performance of skilled tasks might be a detriment for both competitors and spectators [4].

The method has allowed a reduction of analysis time up to 5-fold compared to the accredited methods (STS 288), meeting the minimal required performance limit (MRPL) concentration of the WADA [5].

Generally, the confirmatory analysis is conducted for one specific analyte found positive during the screening step. In certain cases, the determination of the major metabolite or of a concomitant drug intake is simultaneously achieved. Commonly, qualitative results are required, as trace of the drugs of abuse detected in a urine sample is considered as the final result. However, an estimation of the concentration found in urine was required for threshold compounds (e.g., cathine, ephedrine, and methylephedrine), which were considered as the doping agents only above a given cut-off value. Criteria must be established at the confirmatory level for the complete identification of a prohibited substance by high-pressure liquid chromatography (HPLC) coupled to MS [6]. First, all materials should be submitted to the entire analytical process with a strict sample injection order. The first sample to be analyzed is a negative blank urine, followed by the suspect sample, a second negative blank urine, a quality control (QC) and finally a reference collection sample (administration study sample) or a reference material [7]. The retention time (RT) tolerance window must be within the range of ±2% between the suspect analyte and the QC of the same batch. Finally, for MS/MS experiments there should be three diagnostic ions that may include the precursor ion, which must have intensity equal to or greater than 5% of that of the most intense diagnostic ion of the MS/MS spectrum. These should be considered with a S/N ratio >3 and the relative intensity of any of the ions shall not differ by more than 10% (absolute) or 25% (relative) from that of the positive control urine [8].

Nowadays, different techniques such as gas chromatography (GC), capillary electrophoresis (CE), and HPLC are used to confirm and quantify the doping agents in urine matrix. GC is the most frequently used technique for the confirmatory step (e.g., cannabis, ephedrine and related substances, and anabolic steroids) [9]. This technique has been known for years and the coupling of GC with MS detectors is reliable with electron ionisation (EI) sources. Indeed, it allows the construction of worldwide spectral reference libraries and, with the development of fast-GC technologies; analysis time could be drastically shortened. However, the major drawback of GC is its

incompatibility with thermolabile substances, the necessity of hydrolysing conjugate molecules, and derivatising polar analytes.

Methods by CE coupled to laser-induced fluorescence (LIF) detector or to MS were also used to quantify or detect some stimulants [10] and furosemide [11] and for separating chiral isomers (such as ephedrine and related compounds) [12]. Finally, HPLC–MS/MS currently constitutes the method of choice for anti-doping analysis. Indeed, it allows the straightforward determination of polar analytes excreted in urine. Therefore, HPLC–MS/MS methods were successfully developed in the anti-doping field to confirm or quantify amphetamine and derivatives, diuretics, ephedrines, or corticosteroids and anabolic agents [13, 14].

Fast analyses are emerging for anti-doping purposes, since the number of samples to be screened is continuously increasing. Moreover, the time delivery response to give results is required to be 24 h or less after sample reception during the major sporting events.

The use of fast HPLC techniques, such as UHPLC, is of particular interest for screening and confirmatory analysis. UHPLC is a recognized approach to reduce the analysis time and improve or maintain the chromatographic performance by using the columns packed with small particles (i.e., sub-2 μ m diameters). This technique is especially recommended because of its high resolution and excellent retention time repeatability [15]. Benefits of the UHPLC approach have been experimentally highlighted using fast duty cycle mass analysers such as triple quadrupole or time-of-flight (TOF) mass spectrometers in the anti-doping field [16].

The hyphenation of the QTOF mass spectrometer with UHPLC is a very attractive tool for performing the confirmatory analysis. Indeed, the QTOF mass spectrometer can acquire MS/MS spectra with high reproducibility and give accurate mass measurements, allowing the determination of the analyte elemental composition. Moreover, it ensures high selectivity in complex biological matrices and is also proven to be a satisfactory tool for quantitative analysis [17].

Prediction of physic-chemical properties of materials based on their molecular structure has been one of the wishes of scientists and engineers for a long time. One of the best methods, applied for this purpose, is quantitative structure-property relationships (QSRR). QSRR analysis is now a well established and highly respected technique to correlate chromategraphic retention time of a compound with its molecular structure, through a variety of descriptors. The basic strategy of QSRR analysis is to find optimum quantitative relationships, which can then be used to predict the retention from the molecular structures [18, 19]. Once a reliable relation has been obtained, it is

possible to use it to predict that retention for other structures not yet measured or even not yet prepared. QSRR on the retention time have been reported for different types of the organic compounds [20-22].

The application of this technique usually requires variable selection for building well-fitted models. Nowadays, the genetic algorithm method (GA) is well known as an interesting and more widely used variable selection method. GA is a stochastic method that solves the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions such as crossover and mutation [23, 24].

In this work, for the first time, we constructed a QSRR model of the retention time of doping agents and their theoretically derived descriptors. After the variables were selected, the linear multivariate regressions (e.g. the partial least squares (PLS)) as well as the non-linear regressions (e.g. the kernel PLS (KPLS), Levenberg-Marquardt artificial neural network (L-M ANN)) were utilized to construct the linear and non-linear QSRR models. The sets of variables, which provide the best-fitted models for PLS and KPLS methods, were selected with the help of the genetic algorithm. The present study is a first research on QSRR of the doping agents, using GA-PLS, GA-KPLS, and L-M ANN.

Materials and Methods

Equipment

A Pentium IV personal computer (CPU at 3.06 GHz) with the Windows XP operating system was used. The geometry optimization was performed with HyperChem (Version 7.0 Hypercube, Inc). For the calculation of the molecular descriptors, the Dragon 2.1 software was used. The GA-PLS, GA-KPLS, L-M ANN, cross validation, and the other calculations were performed in the MATLAB (Version 7.0, Math works, Inc).

Data set and descriptor generation

The data set, used in this study, is the retention time (RT) of doping agents in urine (a total number of 103 molecules), which obtained by ultra-high-pressure liquid chromatography–quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS) were taken from the literature [25] is shown in Table 1. The prohibited list covers nine pharmaceutical classes of substances (e.g., stimulants, diuretics, anti-estrogens), three forbidden doping methods (e.g., enhancement of oxygen transfer, chemical and physical manipulation and gene doping), and two groups of analytes prohibited in specific activities (e.g., alcohol, β-blockers). In this study, agent doping consist β -Blocker, Stimulant,

Diuretic, Aromatase, inhibitor, Narcotic, Antiestrogen, α-Reductase inhibitor, Uricosuric, and Oxygen transfer enhancer.

Table 1. The data set, structure, class and the corresponding observed retention time values

No	Name	Class	Structure	RT
	Calibration Set			
1	Methylecgonine	Stimulant	$C_{10}H_{18}NO_3$	0.8
2	Benzylpiperazine	Stimulant	$C_{11}H_{17}N_2$	1.15
3	Oxilofrine	Stimulant	$C_{10}H_{16}NO_2$	1.41
4	Pholedrine	Stimulant	$C_{10}H_{16}NO$	1.58
5	Amiloride	Diuretic	$C_6H_9ClN_7O$	1.64
6	Sotalol	β-Blocker	$C_{12}H_{21}N_2O_3S$	1.67
7	Cathine	Stimulant	$C_9H_{14}NO$	1.79
8	Acetazolamide	Diuretic	$C_4H_5N_4O_3S_2$	1.9
9	Ephedrine	Stimulant	$C_{10}H_{16}NO$	1.92
10	Methylephedrine	Stimulant	$C_{11}H_{18}NO$	1.99
11	Aminogluthetimide	Aromatase inhibitor	$C_{13}H_{17}N_2O_2$	2.03
12	Chlorothiazide	Diuretic	$C_7H_5ClN_3O_4S_2$	2.05
13	Nikethamide	Stimulant	$C_{10}H_{15}N_2O$	2.06
14	Nadolol	β -Blocker	$C_{17}H_{28}NO_4$	2.09
15	Etafedrine	Stimulant	$C_{12}H_{20}NO$	2.12
16	Phendimetrazine	Stimulant	$C_{12}H_{18}NO$	2.16
17	Phenpromethamine	Stimulant	$C_{10}H_{16}N$	2.17
18	Indapamide	Diuretic	$C_{16}H_{15}ClN_3O_3S$	2.18
19	MDMA	Stimulant	$C_{11}H_{16}NO_2$	2.22
20	Amfepramone	Stimulant	$C_{13}H_{20}NO$	2.24
21	Phentermine	Stimulant	$C_{10}H_{16}N$	2.25
22	Dimethamphetamine	Stimulant	$C_{11}H_{18}N$	2.26
23	Fenproporex	Stimulant	$C_{12}H_{17}N_2$	2.29
24	Ritalinic acid	Stimulant	$C_{13}H_{18}NO_2$	2.32
25	Norfentanyl	Narcotic	$C_{14}H_{21}N_2O$	2.38
26	Methoxyphenamine	Stimulant	$C_{11}H_{18}NO$	2.42
27	para-Methylamphetamine	Stimulant	$C_{10}H_{16}N$	2.49
28	Isometheptene	Stimulant	$C_9H_{20}N$	2.56

29	Metoprolol	β-Blocker	$C_{15}H_{26}NO_3$	2.57
30	Celiprolol	β-Blocker	$C_{20}H_{34}N_3O_4$	2.72
31	Esmolol	β-Blocker	$C_{16}H_{26}NO_4$	2.73
32	Pethidine	Narcotic	$C_{15}H_{22}NO_2$	2.81
33	Mefenorex	Stimulant	$C_{12}H_{19}ClN$	2.83
34	Chlorthalidone	Diuretic	$C_{14}H_{10}ClN_2O_4S$	2.86
35	Furfenorex	Stimulant	$C_{15}H_{20}NO$	2.89
36	Dichlorphenamide	Diuretic	$C_6H_5Cl_2N_2O_4S_2$	2.9
37	Bupropion	Stimulant	$C_{13}H_{19}ClNO$	2.93
38	Crotetamide	Stimulant	$C_{12}H_{23}N_2O_2$	2.98
39	Etamivan	Stimulant	$C_{12}H_{18}NO_3$	3.01
40	Fenfluramine	Stimulant	$C_{12}H_{17}F_3N$	3.07
41	Prolintane	Stimulant	$C_{15}H_{24}N$	3.08
42	Torasemide	Diuretic	$C_{16}H_{21}N_4O_3S$	3.18
43	Modafinil	Stimulant	$C_{15}H_{16}NO_2S$	3.27
44	Buprenorphine	Narcotic	$C_{29}H_{42}NO_4$	3.38
45	Pentazocine	Narcotic	$C_{19}H_{28}NO$	3.41
46	Probenecide	Uricosuric	$C_{13}H_{18}NO_4S$	3.55
47	Hydrochlorothiazide	Diuretic	$C_7H_7ClN_3O_4S_2$	3.59
48	Methadone	Narcotic	$C_{21}H_{28}NO$	3.83
49	Salmeterol	β-Agonist	$C_{25}H_{38}NO_4$	3.87
50	Sibutramine	Stimulant	$C_{17}H_{27}ClN$	3.97
51	Bendroflumethiazide	Diuretic	$C_{15}H_{13}F_3N_3O_4S_2$	4.08
52	Furosemide	Diuretic	$C_{12}H_{10}ClN_2O_5S$	4.24
53	Mesocarb	Stimulant	$C_{18}H_{19}N_4O_2$	4.29
54	Bumetanide	Diuretic	$C_{17}H_{21}N_2O_5S$	4.37
55	Xipamide	Diuretic	$C_{15}H_{14}ClN_2O_4S$	4.48
56	Spironolactone	Diuretic	$C_{24}H_{33}O_4S$	4.57
57	Canrenone	Diuretic	$C_{23}H_{28}O_3$	4.62
58	Ethacrynic acid	Diuretic	$C_{13}H_{13}Cl_2O_4$	4.63
59	Clomiphen	Antiestrogen	$C_{26}H_{29}ClNO$	4.66
60	Amfetaminil	Stimulant	$C_{17}H_{19}N_2$	5.24
	Prediction Set			
61	Heptaminol	Stimulant	$C_8H_{20}NO$	1.54
62	Phenylpropanolamine	Stimulant	$C_9H_{14}NO$	1.79

63	Carteolol	β-Blocker	$C_{16}H_{25}N_2O_3$	2.01
64	Metamphetamine	Stimulant	$C_{10}H_{16}N$	2.17
65	Triamterene	Diuretic	$C_{12}H_{12}N_7$	2.23
66	Strychnine	Stimulant	$C_{21}H_{23}N_2O_2$	2.25
67	Pemoline	Stimulant	$C_9H_9N_2O_2$	2.31
68	Ethylamphetamine	Stimulant	$C_{11}H_{18}N$	2.34
69	Acebutolol	β-Blocker	$C_{18}H_{29}N_2O_4$	2.45
70	Methylphenidate	Stimulant	$C_{14}H_{20}NO_2$	2.62
71	Cocaine	Stimulant	$C_{17}H_{22}NO_4$	2.78
72	Norbuprenorphine	Narcotic	$C_{25}H_{36}NO_4$	2.85
73	Pipradol	Stimulant	$C_{18}H_{22}NO$	2.91
74	Fencamfamine	Stimulant	$C_{15}H_{22}N$	3.01
75	Adrafinil	Stimulant	$C_{15}H_{14}NO_3S$	3.17
76	Clobenzorex	Stimulant	$C_{16}H_{19}ClN$	3.34
77	Anastrozole	Aromatase inhibitor	$C_{17}H_{20}N_5$	3.74
78	Piretanide	Diuretic	$C_{17}H_{19}N_2O_5S$	4.14
79	RSR13	Oxygen transfer enhancer	$C_{20}H_{24}NO_4$	4.45
80	Dextromoramide	Narcotic	$C_{25}H_{33}N_2O_2$	4.62
	Validation Set			
81	Etilefrine	Stimulant	$C_{10}H_{16}NO_2$	1.41
82	Atenolol	β-Blocker	$C_{14}H_{23}N_2O_3$	1.61
83	Amiphenazole	Stimulant	$C_9H_{10}N_3S$	1.79
84	Pseudoephedrine	Stimulant	$C_{10}H_{16}NO$	1.92
85	Caffeine	Stimulant	$C_8H_{11}N_4O_2$	2.03
86	Amphetamine	Stimulant	$C_9H_{14}N$	2.08
87	MDA	Stimulant	$C_{10}H_{14}NO_2$	2.12
88	Metolazone	Diuretic	$C_{16}H_{15}ClN_3O_3S$	2.18
89	Pentetrazole	Stimulant	$C_6H_{11}N_4$	2.24
90	Benzoylecgonine	Stimulant	$C_{16}H_{20}NO_4$	2.36
91	Fenetylline	Stimulant	$C_{18}H_{24}N_5O_2$	2.56
92	Carphedon	Stimulant	$C_{12}H_{15}N_2O_2$	2.66
93	Chlorphentermine	Stimulant	$C_{10}H_{15}ClN$	2.75
94	Propylhexedrine	Stimulant	$C_{10}H_{22}N$	2.83
95	Norfenfluramine	Stimulant	$C_{10}H_{13}F_3N$	2.86
96	Clopamide	Diuretic	$C_{14}H_{20}ClN_3O_3S$	2.93

97	Metipranolol	β-Blocker	$C_{17}H_{28}NO_4$	3.08
98	Fentanyl	Narcotic	$C_{22}H_{29}N_2O$	3.24
99	Cropropamide	Stimulant	$C_{13}H_{25}N_2O_2$	3.42
100	Fenbutrazate	Stimulant	$C_{23}H_{30}NO_3$	3.97
101	Finasteride	α-Reductase inhibitor	$C_{23}H_{37}N_2O_2$	4.28
102	Exemestane	Aromatase inhibitor	$C_{20}H_{25}O_2$	4.56
103	Hydroxybromantan	Stimulant	$C_{16}H_{21}BrNO$	5.1

The chemical structure of the 103 studied molecules were drawn with the Hyperchem software and saved with the HIN extension. To optimize the geometry of the studied molecules, the AM1 geometrical optimization was applied. The DRAGON software was used to calculate the descriptors in this research and a total of 1497 molecular descriptors, belonging to 18 different types of the theoretical descriptors, were calculated for each molecule.

Experimental

Stock standard solutions of the 103 substances were prepared at a concentration of 1mg/mL in methanol and kept at −20 °C in glass tubes fitted with PTFE caps. Quality controls (QCs) solutions (103) were prepared by spiking 10 μ L of the diluted standard solutions in an aliquot of 500 μ L of urine to obtain a final concentration at the MRPL level for each analyte.

Separations were carried out on an Acquity UPLC system (Waters, Milford, MA, USA) with Waters Acquity UPLC columns (BEH C_{18} 100 mm×2.1 mm, 1.7 μ m) at 30 °C and 400 μ L × min^{-1}. UHPLC conditions were maintained identical for the screening and the pre-confirmatory methods. Analyses were performed using a Micromass-Q-Tof Premier mass spectrometer (Waters) equipped with an ESI source. MS operating conditions were set as follows: the desolvation gas flow was 800 L/h at a temperature of 300 °C, the capillary voltages were defined as 3.0 kV in positive mode and 2.4 kV in negative mode, and the cone voltage was constant at 40 V in both modes.

UHPLC allows an increase in resolution, throughput and sensitivity using sub-2_m particles. Therefore, a fast gradient of 3minwith 1.5 min of equilibration time was generated on a short column (50 mm). A selective QTOF-MS and MS/MS detection was performed for each analyte to meet the WADA's identification criteria. With the QTOF mass analyzer, it was possible to obtain a QTOF-MS full scan acquisition in a first channel and a QTOF–MS/MS spectrum in a second channel in the same analytical run. The acquisition of simultaneous MS and MS/MS methods at two collision energies allows the determination of precursor and product ions with high mass accuracy. A

dedicated MS/MS method was developed for each analyte by setting the cone voltage and the collision energy at the analyte expected RT to obtain at least three diagnostic ions, including the protonated molecule.

Data pretreatment

The calculated descriptors were first analyzed to check the existence of the constant or near-constant variables, which were removed, in case they existed at all. Furthermore, in order to decrease the redundancy existing in the descriptor data matrix, the correlation of the descriptors with each other and with the property (RT) of the molecules was examined and the collinear descriptors (i.e. $r > 0.9$) were detected. Among the collinear descriptors, the one with the highest correlation with the property was retained and the others were removed from the data matrix. Then, the remaining descriptors were collected in an $n \times m$ data matrix (D), where n=103 and m=906 are the number of the compounds and the descriptors, respectively. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program.

Genetic algorithm for descriptor selection

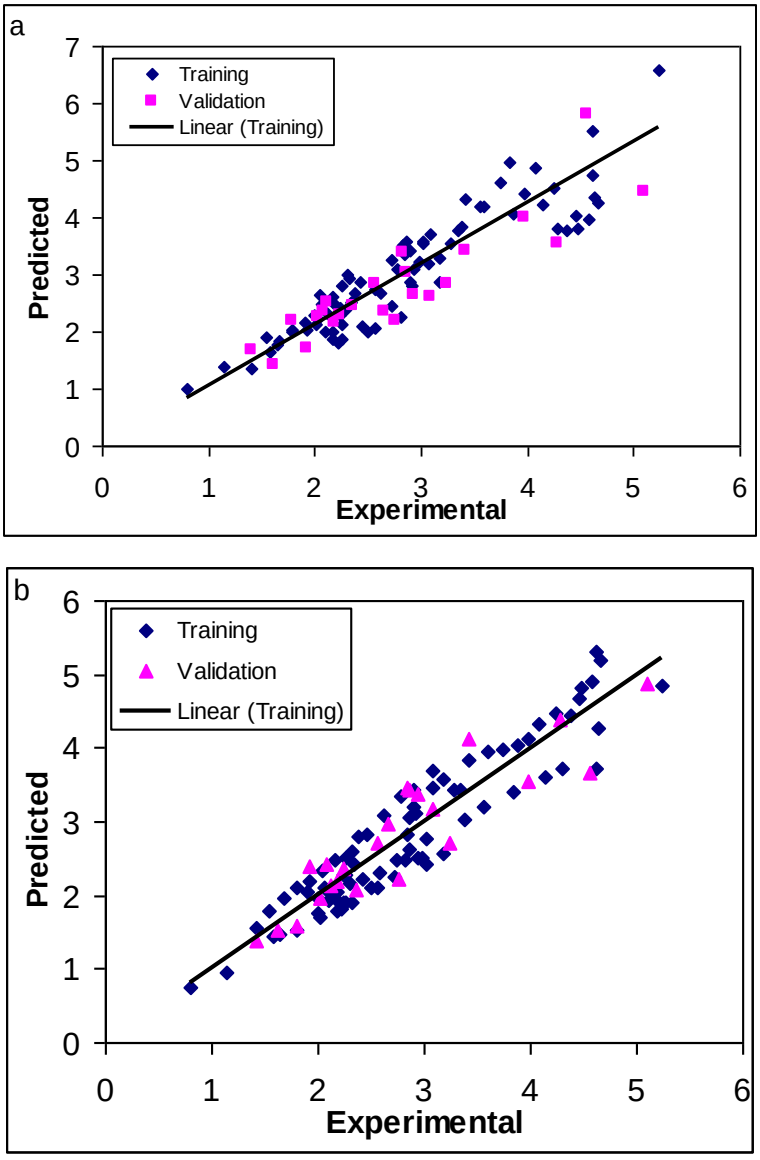
Genetic algorithm is a problem solving method that uses generic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected based on a fitness criterion to survive and pass information on to the next generation [26]. GA uses a binary bit string representation as the coding technique for a given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0. A string is composed of several genes that represent a specific characteristic to be studied. In the present case, a string is composed of 561 genes representing the presence or absence of a descriptor. By encoding various descriptors with bit strings, called chromosomes, the initial population was created randomly. The population size was varied between 50 and 300 for different GA runs. For a typical run, the evolution of the generation was stopped, when 90% of the generations had taken the same fitness [27, 28]. In this paper, size of the population is 30 chromosomes, the probability of initial variable selection is 5:V (V is the number of independent variables), crossover is multi Point, the probability of crossover is 0.5, mutation is multi Point, the probability of mutation is 0.01 and the number of evolution generations is 1000. For each set of data, 3000 runs were performed.

Nonlinear model

Artificial neural network

A three-layer back propagation artificial neural network ANN (Figure 2) with a sigmoid transfer function was used to investigate the feature sets.

Figure 2. Plots of predicted retention time against the experimental values by (a) GA-PLS model and (b) GA-KPLS model



The descriptors from the training set were used for the model generation whereas the descriptors from the validation set were used to stop the overtraining of the network. In addition, the descriptors from the validation set were used to verify the predictively of the model. Before training the networks, the input and output, values were normalized with auto-scaling of all data [29, 30]. The initial weights were selected randomly between -0.3 and 0.3. For the purpose of comparison of results, the same number of hidden layer nodes was used for the ANN models from

all other feature sets of each database. The goal of training the network is to minimize the output errors by changing the weights between the layers.

$$\Delta W_{ij,n} = F_n + \alpha \Delta W_{ij,n-1} \quad (1)$$

In this, ΔW_{ij} is the change in the weight factor for each network node, α is the momentum factor, and F is a weight update function, which indicates how weights are changed during the learning process. The weights of hidden layer were optimized using the Levenberg-Marquardt algorithm, a second derivative optimization method [31].

Levenberg-Marquardt Algorithm

In Levenberg-Marquardt algorithm, the update function, F_n, is calculated using equations.

$$F_0 = -g_0 \quad (2)$$

$$g = J^T e \quad (3)$$

$$F_n = -[J^T \times J + \mu I]^{-1} \times J^T \times e \quad (4)$$

Where g is gradient and J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights, and e is a vector of network errors. The parameter μ is multiplied by some factor (λ) whenever a step would result in an increased e and when a step reduces e, μ is divided by λ [32, 33].

Results and Discussion

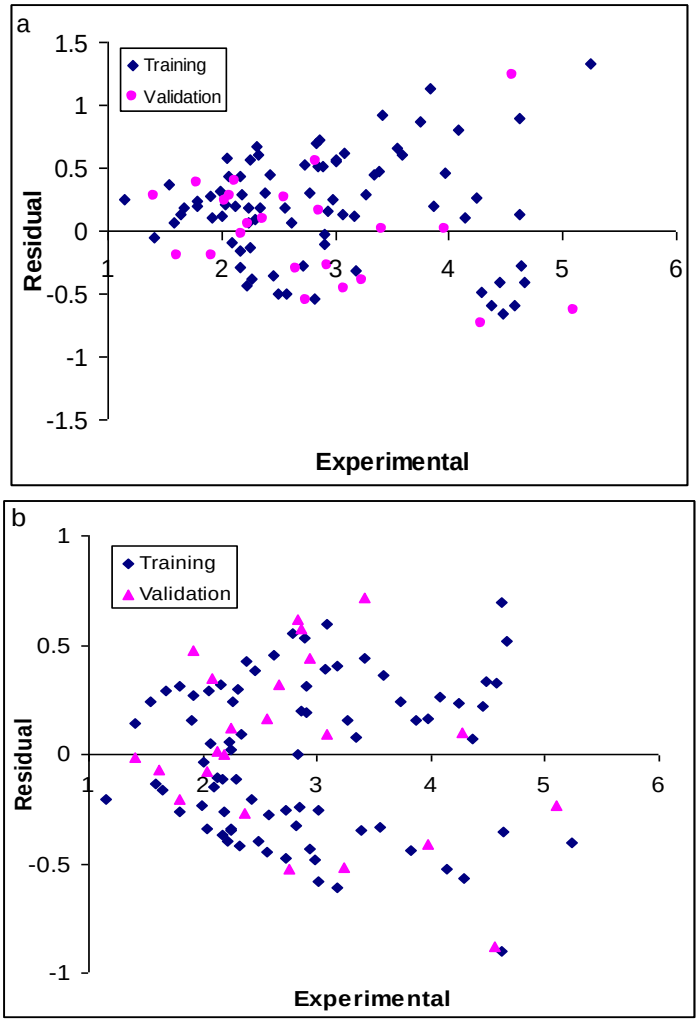
Linear model

Results of the GA-PLS model

The best model is selected on the basis of the highest square correlation coefficient leave-group-out cross validation (R^2), the least root mean squares error (RMSE) and relative error (RE) of prediction. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains thirteen selected descriptors in three latent variables space. These descriptors were obtained constitutional descriptors (number of Hydrogen atoms (nH) and mean atomic Sanderson electronegativity (scaled on Carbon atom) (Me)), topological descriptors (Narumi simple topological index (log) (SNar)), 2D autocorrelations (Moran autocorrelation - lag 1/ weighted by atomic polarizabilities (MATS1p)), GETAWAY descriptors (leverage-weighted autocorrelation of lag 1/ weighted by atomic masses (HATS1m), leverage-weighted autocorrelation of lag 4/ weighted by atomic masses (HATS4m) and H autocorrelation of lag 6/ weighted by atomic Sanderson electronegativities (H6e)), geometrical descriptors (gravitational index G2 (bond-

restricted) (G2)), functional group counts (number of total secondary C(sp^3) (nCs) and number of acceptor atoms for H-bonds (N,O,F) (nHAcc)), atom-centred fragments (CH_3R/ CH_4 (C-001) and H attached to C1(sp^3)/C0(sp^2) (H-047)) and quantum descriptors (highest occupied molecular orbital (HOMO)). The R^2, mean RE and RMSE for training and validation sets were (0.851, 0.803), (14.59, 16.99) and (0.46, 0.90), respectively. The predicted values of RT are plotted versus the experimental values for training and validation sets in Figure 2a. The residuals (predicted RT–experimental RT) obtained by the GA-PLS modeling versus the experimental RT values are demonstrated in Figure 3a. For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The PLS model uses higher number of descriptors that allow the model to extract better structural information from descriptors to result in a lower prediction error.

Figure 3. The residual vs. the experimental RT in (a) GA-PLS and (b) GA-KPLS models



Nonlinear model

Results of the GA-KPLS model

In this paper a radial basis kernel function, k(x, y)= exp(||x-y||^2/c), was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r was set to be 1), m is the dimension of the input space and σ^2 is the variance of the data [34]. It means that the value of c depends on the system under the study. The 9 descriptors in 7 latent variables space chosen by GA-KPLS feature selection methods were contained. These descriptors were obtained constitutional descriptors (number of Carbon atoms (nC)), topological descriptors (spanning tree number (log) (STN) and centralization (CENT)), GETAWAY descriptors (leverage-weighted autocorrelation of lag 1/ weighted by atomic masses (HATS1m), geometrical descriptors (gravitational index G2 (bond-restricted) (G2) and (Qzz COMMA2 value/ weighted by atomic Sanderson electronegativities (QZZe)), functional group counts (number of unsubstituted benzene C(sp2) (nCbH)), molecular properties (Squared Moriguchi octanol-water partition coeff. (logP^2) (MLOGP2)) and quantum descriptors (highest occupied molecular orbital (HOMO)). The R^2, mean RE and RMSE for training and validation sets were (0.873, 0.816), (13.89, 16.26) and (0.43, 0.74), respectively. Figure 2b illustrates the plot of the GA-KPLS predicted versus the experimental values for RT of all the molecules in the data set. The plots of the residuals versus the experimental RT values obtained by the GA-KPLS modeling, is demonstrated in Figure 3b. It can be seen from these results that statistical results for GA-KPLS model are superior to GA-PLS method. Inspection of the results of the table reveals a higher R^2 and lower RMSE and RE for the GA-KPLS method compared with their counterparts for linear model. Also, a lower number of variables have appeared in the former model. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method.

Results of the L-M ANN model

The networks were generated using descriptors appearing in the GA-KPLS model as inputs. For ANN generation, dataset was separated into three groups: calibration, prediction and validation sets. Before training, the input and output values were normalized between 0 and 1. Number of neurons in the hidden layer, learning rate and momentum were optimized. A feed-forward neural network with back-propagation algorithm was constructed to model the retention relationship [35]. This method is an iterative algorithm that allows training of multilayer networks. The algorithm looks for the minimum of the error function. In this way, the training process tries to

diminish the difference between the outputs of the network and the expected values. Of course, there are some other approaches such as Levenberg Marquardt algorithm, gradient descent with variable learning rate back-propagation and resilient back-propagation. These networks are different in weight update functions and can converge faster than steepest decent method [36]. But this paper has not focused on investigating the role of weight update functions or calculation time in artificial neural networks. Our network has nine input layer, four hidden layer, and one output layer. A bias unit with a constant activation of unity is connected to each unit in the hidden and output layers. Once the best topology of the network is obtained and the convergence criterion is reached, a leave-4- out cross-validation procedure is also employed to more validate the performances of the resulted networks. To evaluate the performance of the ANN, RMSE of the calibration was used. The number of neurons in the hidden layer with the minimum value of RMSE was selected as the optimum number. Learning rate and momentum were optimized in a similar way. It was realized that the RMSE for the training and validation sets are minimum when four neurons were selected in the hidden layer. The R^2 and RE for calibration, prediction and validation sets were (0.943, 0.925, 0.901) and (8.11, 10.01, 11.67), respectively. Also, RMSE for calibration, prediction and validation sets were (0.29, 0.41, 0.53), respectively. Inspection of the results reveals a higher R^2 and lowers other values parameter for the validation set compared with their counterparts for other models. Plots of predicted RT versus experimental RT values by L-M ANN for calibration, prediction and validation sets are shown in Figure 4a, 4b, respectively. The residuals of L-M ANN predicted values of RT against the experimental values are plotted in Figure 5a and Figure 5b. As the calculated residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the development of the Neural Network. The relative error and R^2 of validation set for the GA-PLS and GA-KPLS models are (16.99, 0.803) and (16.26, 0.816), respectively which would be compared with the values of (13.19, 0.901, 11.67), respectively, for L-M ANN model.

Figure 4. Plot of predicted RT obtained by L-M ANN against the experimental values (a) calibration and prediction sets of molecules and (b) for validation set

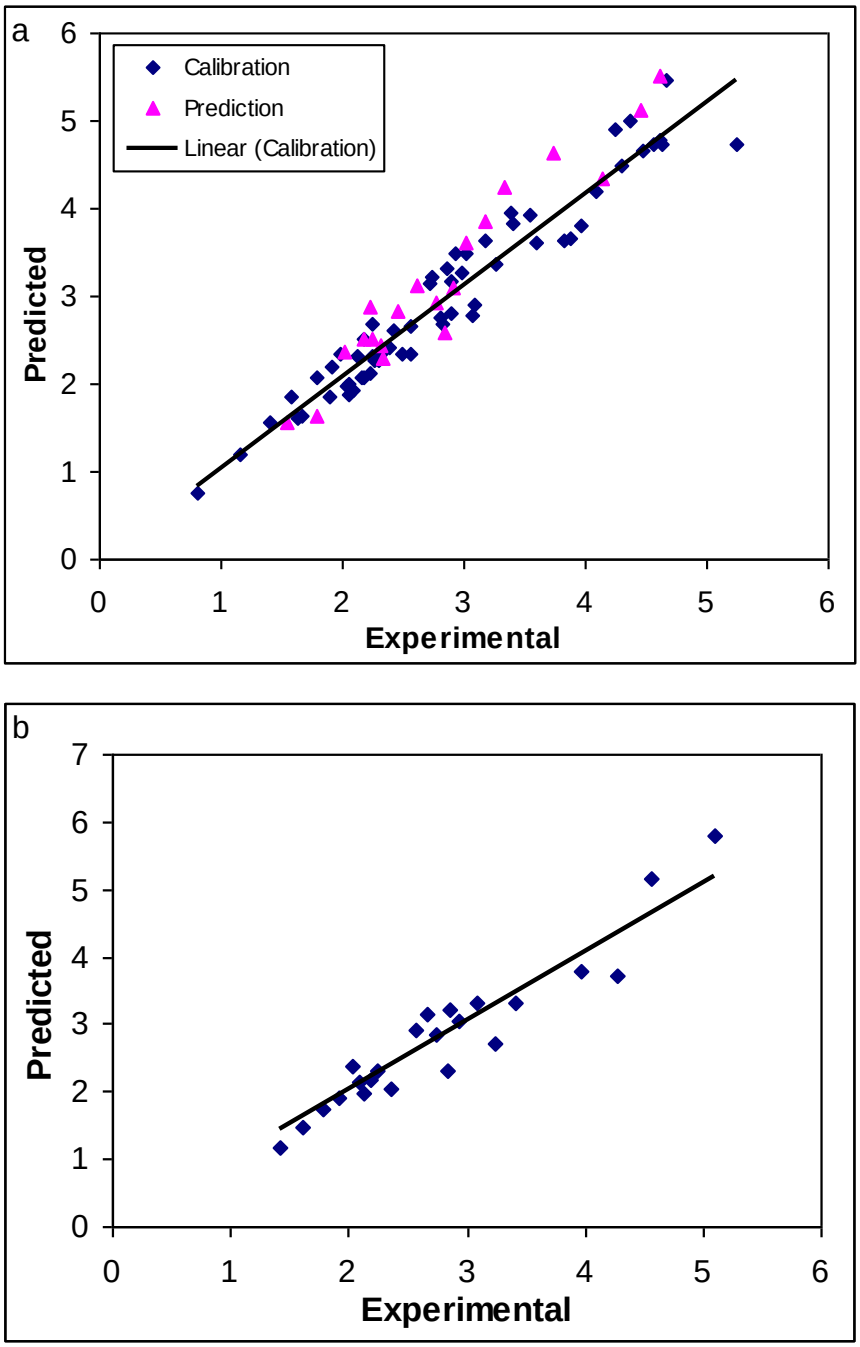
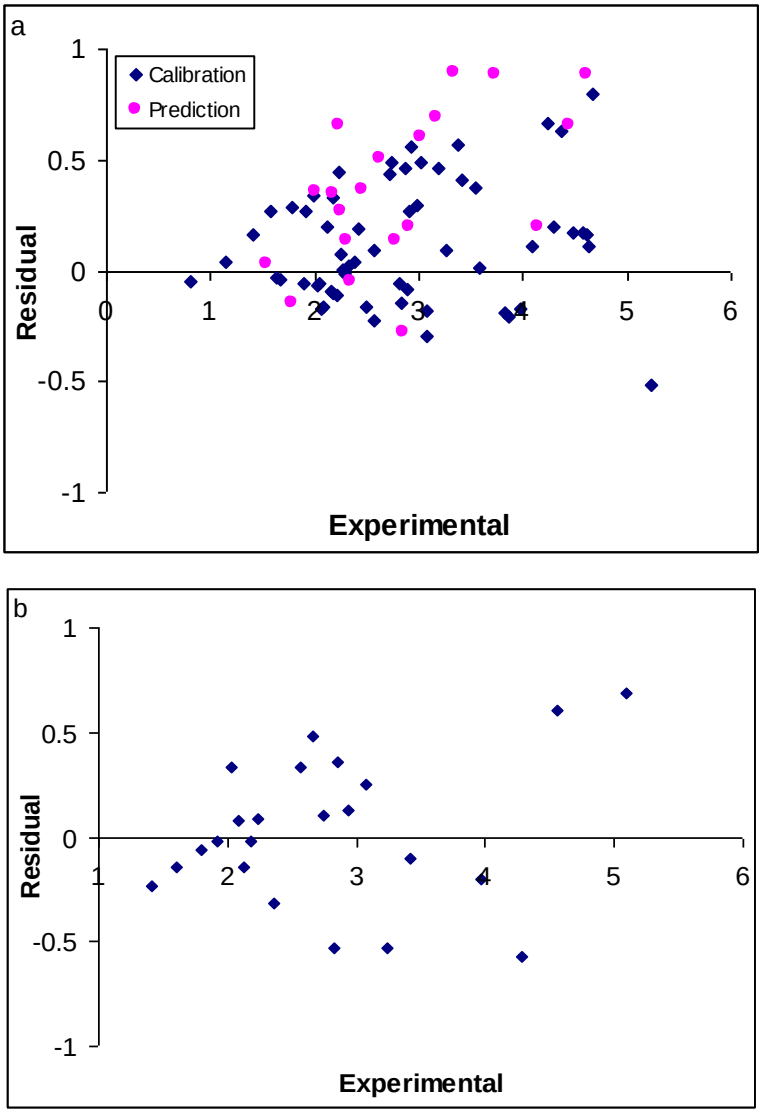


Figure 5. Plot of residuals obtained by L-M ANN against the experimental RT values (a) training set of molecules and (b) for validation set



Comparison between these values and other statistical parameters reveals the superiority of the L-M ANN model over other models. The key strength of neural networks, unlike regression analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence implicitly. The statistical parameters reveal the high predictive ability of L-M ANN model. The whole of these data clearly displays a significant improvement of the QSRR model to nonlinear statistical treatment. Obviously, there is a close agreement between the experimental and predicted RT and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. As can be seen in this section, the L-M ANN is more

reproducible than GA-PLS and GA-KPLS for modeling the UHPLC-QTOF-MS retention time of doping agents.

Interpretation of descriptors

In the chromatographic retention of compounds in the stationary phase, two important types of interactions contribute to the chromatographic retention of the compounds: the induction and dispersion forces. The dispersion forces are related to steric factors, molecular size, shape, and branching, while the induced forces are related to the dipolar moment, which should stimulate dipole-induced dipole interactions.

Constitutional descriptors are most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. Number of C atoms, the average bond order of a C atom and the minimum atomic state energy for a C atom quantify the bond strength between the C atoms. A molecule locked in a rigid conformation due to strong intramolecular interactions is in fact less free to move and is expected to have a higher boiling point.

The hydrogen bonding is a measure of the tendency of a molecule to form hydrogen bonds. This is related to number of Hydrogen atoms (nH). Hydrogen-bonding may be divided into an electrostatic term and a polarization/ charge transfer term.

The geometrical descriptors are suitable for complex-behaved properties, because they take into account the 3D-arrangement of atoms without ambiguities (as those appearing when using chemical graphs), as well as they do not depend on the molecular size and thus they are applicable to a large number of molecules with great structural variance, which have a characteristic common to all of them.

The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors try to match 3Dmolecular geometry provided by the molecular influence matrix and atom relatedness by molecular topology, with chemical information by using different atomic weights. These descriptors are quickly computed from the atomic positions of the molecule atoms (hydrogens included).

The geometrical descriptors are suitable for complex-behaved properties, because they take into account the 3D-arrangement the atoms without ambiguities (as those appearing when using chemical graphs), as well as they do not depend on the molecular size and thus they are applicable to a large number of molecules with great structural variance, which have a characteristic common to all of them.

Gravitational index (G2) (bond-restricted) is a geometrical descriptor that reflecting the mass distribution in a molecule and defined as Eq. (5):

$$G_2 = \sum_{a=1}^{A} \left(\frac{m_i . m_j}{r_{ij}^2} \right)_a \quad (5)$$

Where m_i and m_j are the atomic masses of the considered atoms; r_{ij} the corresponding interatomic distances; and A the number of all pairs of bonded atoms of the molecule. This index is related to the bulk cohesiveness of the molecules, accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space. This index can be extended to any other atomic property different from atomic mass, such as atomic polarizability, atomic, van der Waals volumeetc.

Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity.

Although these descriptors are often successful in rationalizing RT of doping agents, they cannot account for conformational changes and they do not provide information about electronic influence through bonds or across space. For that reason, quantum chemical descriptors are used in developing QSRR.

Quantum chemical descriptors were defined in terms of the atomic charges and used to describe the both electronic aspects of the whole molecule and of particular regions, such as atoms, bonds, and molecular fragments. They include thermodynamic properties (system energies) and electronic property (HOMO energy). The HOMO as an electron donor represents the ability to donate an electron. The HOMO energy plays a very important role in the nucleophylic behavior and it represents molecular reactivity as a nucleophyle [37].

From the above discussion, it can be seen that the particle size, hydrogen bonding, and electrostatic interactions are the likely three factors controlling the RT of these compounds. All the descriptors involved in the model–which have an explicit physical meaning may account for the structure responsible for the RT of these compounds.

Model validation and statistical parameters

The applied internal (leave-group-out cross validation (LGO-CV)) and external (validation set) validation methods were used for the predictive power of models. In the leave-group-out procedure one compound was removed from the data set, the model was trained with the remaining compounds and used to predict the discarded compound. The process was repeated for each compound in the data set. The predictive power of the models developed on the selected training set is estimated on the predicted values of validation set chemicals. The data set should be divided into three new sub-data sets, one for calibration and prediction (training), and the other one for validation sets. The calibration set was used for model generation. The prediction set was applied deal with overfitting of the network, whereas validation set which its molecules have no role in model building was used to investigate the predictive ability of the models for the external set [38, 39].

In the other hand by means of training set, the best model is found and then, the prediction power of it is checked by validation set, as an external data set. In this work, from all 103 components, 60 components are in calibration set, 20 components are in prediction set and 23 components are in validation set).

The result clearly displays a significant improvement of the QSRR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the validation molecule. In the above analysis, the descriptive power of a given model has been measured by its ability to predict partition of unknown doping agents.

For the constructed models, some general statistical parameters were selected to evaluate the predictive ability of the models for RT values. In this case, the predicted RT of each sample in prediction step was compared with the experimental acidity constant.

Root mean square error (RMSE) is a measurement of the average difference between predicted and experimental values, at the prediction step. RMSE can be interpreted as the average prediction error, expressed in the same units as the original response values. Its small value indicates that the model predicts better than chance and can be considered statistically significant. The RMSE was obtained by the following formula:

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2]^{\frac{1}{2}} \quad (6)$$

The other statistical parameter was relative error (RE) that shows the predictive ability of each component, and is calculated as:

$$RE(\%) = 100 \times \left[\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i^{\wedge} - y_i)}{y_i} \right] \quad (7)$$

The predictive ability was evaluated by the square of the correlation coefficient (R^2) which is based on the prediction error sum of squares and was calculated by following equation:

$$R^2 = \frac{\sum_{i=1}^{n} (y_i^{\wedge} - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})} \quad (8)$$

Where y_i is the experimental RT in the sample i, $\hat{y}_i$ represented the predicted RT in the sample i, $\bar{y}$ is the mean of experimental RT in the prediction set and n is the total number of samples used in the validation set [40, 41].

The main aim of the present work was to assess the performances of GA-PLS, GA-KPLS and L-M ANN for modeling the retention time of compounds. The procedures of modeling including descriptor generation, splitting of the data, variable selection and validation were the same as those performed for modeling of the retention time of doping agents.

Conclusion

The GA-PLS, GA-KPLS, and L-M ANN modeling were applied to predict the retention time of 103 doping agents. High correlation coefficients and low prediction errors confirmed the good predictability of models. Application of the developed model to a validation set of 23 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Three methods seemed to be useful, although a comparison between these methods revealed the slight superiority of the L-M ANN over the models. The QSRR procedure allowed us to achieve a precise and relatively fast method for determination of RT of different series of doping agents to predict with sufficient accuracy the RT of new compound derivatives. To the best of our knowledge, this is the first study for the prediction of retention time of doping agents using GA-PLS, GA-KPLS and L-M ANN.

References

[1] Parssinen M., Seppala T. *Sports Med.*, 2002, **32**:83

[2] Gruber A.J., Pope H.G. *Psychother Psychosom.,* 2000, **69**:19

[3] Scarth J.P., Clarke A.D., Teale Ph., Pearce C.M. *Steroids*, 2010, **75**:643

[4] World Anti-Doping Agency (WADA), *The World Anti-Doping Code*, The 2008 Prohibited List, Montreal, 2009, www.wada-ama.org (accessed May 2009).

[5] World Anti-doping Agency (WADA), *The World Anti-Doping Code. Minimal Required Performance Limits, Technical Document TD2004MRPL*, Montreal, **2004**, www.wada-ama.org (accessed May 2009).

[6] Van Eenoo P., Delbeke F.T. *Chromatographia*., 2004, **59**:39

[7] World Anti-Doping Agency (WADA), *International Standard for Laboratories* V5.0, Montreal, **2008**, www.wada-ama.org (accessed May 2009).

[8] Rivier L. *Anal. Chim. Acta*., 2003, **492**:69

[9] Hadef Y., Kaloustian J., Portugal H., Nicolay A. *J. Chromatogr. A.*, 2008, **1190**:278

[10] Schappler J., Guillarme D., Rudaz S., Veuthey J.L. *Electrophoresis*., 2008, **29**:11

[11] Caslavska J., Thormann W. *J. Chromatogr. B*. 2002, **770**:207

[12] Mateus-Avois L., Mangin P., Saugy M. *J. Chromatogr. B*., 2003, **791**:203

[13] Giancotti V., Medana C., Aigotti R., Pazzi M., Baiocchi C. *J. Pharm. Biomed. Anal*., 2008, **48**:462

[14] Deventer K., Pozo O.J., Van Eenoo P., Delbeke F.T. *J. Chromatogr. B.***,** 2009, **877**:369

[15] Gika H.G., Macpherson E., Theodoridis G.A., Wilson. *J. Chromatogr. B.,* 2008, **871**:299

[16] Touber M.E., van Engelen M.C., Georgakopoulos C., van Rhijn J.A., Nielen M.W.F. *Anal. Chim. Acta*., 2007, **586**:137

[17] Williamson L.N., Bartlett M.G. *Biomed. Chromatogr*., 2007, **21**:567

[18] Gupta V.K., Khani H., Ahmadi-Roudi B., Mirakhorli Sh., Fereyduni E., Agarwal S. *Talanta*, 2011, **83**:1014

[19] Matteis C.I.D., Simpson D.A., Doughty S.W., Euerby M.R., Shaw P.N., Barrett D.A. *J. Chromatogr. A.*, 2010, **1217**:6987

[20] Liu T., Nicholls I.A., Öberg T. *Anal. Chim. Acta.,* 2011, **702**:37

[21] Riahi S., Pourbasheer E., Ganjali M.R., Norouzi P. *J. Hazard. Mater.*, 2009, **166**:853

[22] Bodzioch K., Durand A., Kaliszan R., Bączek T., Vander Heyden Y. *Talanta*, 2010, **81**:1711

[23] Leardi R. *Comper. Chemom*., 2009, **120**:631

[24] Ferrand M., Huquet B., Barbey S., Barillet F., Faucon F., Larroque H., Leray O., Trommenschlager J.M., Brochard M. *Chemom. Intell. Lab. Syst*., 2011, **106**:183

[25] Badoud F., Grata E., Perrenoud L., Avois L., Saugy M., Rudaz S., Veuthey J.L. *J. Chromatogr. A.* 2009, **1216**:4423

[26] Devos O., Duponchel L. *Chemom. Intell. Lab. Syst*., 2011, **107**:50

[27] Hemmateenejad B., Shamsipur M., Zare-Shahabadi V., Akhond M. *Anal. Chim. Acta*., 201۱, **704**:57

[28] Pourbasheer E., Riahi S., Ganjali M.R., Norouzi P. *Eur. J. Med. Chem.,* 2009, **44**:5023

[29] Singh K.P., Ojha P., Malik A., Jain G. *Chemom. Intell. Lab. Syst*., 2009, **99**:150

[30] Jančić-Stojanović B., Ivanović D., Malenović A., Medenica M. *Talanta*, 2009, **78**:107

[31] Jalali-Heravi M., Asadollahi-Baboli M., Shahbazikhah P. *Eur. J. Med. Chem*., 2008, **43**:548

[32] Xuefeng Y. *Chemom. Intell. Lab. Syst*., 2010, **103**:152

[33] Singh K.P., Basant N., Malik A., Jain G. *Anal. Chim. Acta*, 2010, **658**:1

[34] Kim K., Lee J.M., Lee I.B. *Chemom. Intell. Lab. Syst*., 2005, **79**:22

[35] D'Archivio A.A., Maggi M.A., Mazzeo P., Ruggieri F. *Anal. Chim. Acta***,** 2008, **628**:162

[36] Jančić B., Medenica M., Ivanović D., Janković S., Malenović A. *J. Chromatogr. A*, 2008, **1189**:366

[37] Todeschini R., Consonni V. *Handbook of Molecular Descriptors*, Wiley/VCH, Weinheim, 2000.

[38] Deeb O. *Chemom. Intell. Lab. Syst.,* 2010, **104**:181

[39] Kishore D.P., Balakumar C., A.R. Rao, Roy P.P., Roy K. *Bioorg. Med. Chem. Lett*., 2011, **21**:818

[40] Arab Chamjangali M., Beglari M., Bagherian G. *J. Mol. Graphics Modell*, 2007, **26**:360

[41] Hemmateenejad B., Javadnia K., Elyasi M. *Anal. Chim. Acta*, 2007, **592**:72

How to cite this manuscript: Mehrdad Shahpar*, Sharmin Esmaeilpoor. Approach to Chemometrics Models by Artificial Neural Network for Structure: First Applications for Estimation Retention Time of Doping Agent. Chemical Methodologies 1(2), 2017, 98-120. DOI: 10.22631/chemm.2017.96397.1008.