A QSPR STUDY OF BOILING POINT OF SATURATED ALCOHOLS USING GENETIC ALGORITHM

Mohsen Kompany-Zareh

Institute for Advanced Studies in Basic Sciences, Zanjan 45195-159, Iran.

Received 28-10-2002

Abstract

A QSPR method is applied to study the boiling point of alcohols by the employment of the following properties: Schulz's index, Randić's connectivity index, Wiener's number, surface area, volume, log *P*, molar refractivity, and polarizability. The idea behind the choice of these topological and physicochemical descriptors is to use realistic molecular quantities, which can, in principle, express all of the topological, electronic and geometric properties of molecules and their interactions. The boiling point values for a set of 44 alkanols were used, and by using a genetic algorithm (GA) coupled with partial least squares (PLS) method, all different possible relations between boiling point (bp) and the molecular properties up to the fourth order were examined and a group of multiple regression models with high fitness scores was generated. Using a backward elimination method on the top descriptors obtained from genetic algorithm, Randić's index, surface area (grid), log of octanol-water partition coefficient, molecular refractivity, and polarizability were selected as significant descriptors. The analysis of computed data, and test of model for a validation set including 10 alcohols, shows that selected descriptors and selected order for each one are extremely well fitted tools for assessing the boiling point of alcohols. In particular, we have verified that using higher level relationships (i.e. square, cubic, and/or quadratic) in several-variable equations give excellent accuracy.

Introduction

In chemistry, anything that can be said about the magnitude of the property and its dependence upon changes in the molecular structure depends on the chemist's capability to establish valid relationships between structure and property. In many physical organic, biochemical and biological areas, it is increasingly necessary to translate those general relations into quantitative associations expressed in useful algebraic equations known as Quantitative Structure-Activity(-property) Relationships (QSA(P)R).[1] To obtain a significant correlation, it is crucial that appropriate descriptors be employed, whether they are theoretical, empirical or derived from readily available experimental features of the structures. Many descriptors reflect simple molecular properties and thus they can provide some meaningful insights into the physico-chemical nature of the activity/property under consideration.[2]

In a relatively recent paper Castro et al.[3] applied three well known topologic indices in the QSPR study of boiling point of saturated alcohols: The Schultz index, the

Wiener number, and a connectivity index of Randić. They analyzed several polynomial correlations between the boiling points and the three topological indices, and found a satisfactory enough agreement between the theoretical and experimental results.

Table 1. Features used in the QSPR analysis of the data set.

MTI :	Schultz index [references 4-6].[a]
χ^v :	Randić's Valence connectivity index [references 9 and 10].[a]
W :	Wiener number [references 7 and 8].[a]
SAG :	Surface area (grid) [references 11 and 12].[b]
V :	Volume.[b]
Log *P* :	log of octanol-water partition coefficient [references 13 and 14].[b]
MR :	Molar refractivity [references 14 and 15].[b]
POL :	Polarizability [reference 16].[b]

[a] Topological descriptor. [b] Molecular descriptor.

Table 2. Experimental and calculated boiling points for the validation set by equations 4 and 10.

	Alkanol	bp(°C) obsd	bp(°C) calcd [a]	Error % [a]	bp(°C) calcd [b]	Error % [b]
	Training set					
1.	Methanol	64.7	63.61	-1.68	65.88	1.82
2.	Ethanol	78.3	81.88	4.58	81.67	4.30
3.	1-propanol	97.2	98.93	1.78	97.98	0.80
4.	2-propanol	82.3	83.13	1.01	86.81	5.49
5.	1-butanol	117.7	116.28	-1.20	113.42	-3.64
6.	2-methyl-1-propanol	107.9	104.90	-2.78	104.53	-3.12
7.	2-methyl-2-propanol	82.4	79.49	-3.53	80.59	-2.20
8.	1-pentanol	137.8	134.25	-2.58	134.09	-2.69
9.	3-pentanol	115.3	122.89	6.58	122.98	6.66
10.	2-methyl-1-butanol	128.7	126.15	-1.98	125.35	-2.60
11.	3-methyl-1-butanol	131.2	125.20	-4.58	124.21	-5.33
12.	3-methyl-2-butanol	111.5	112.67	1.05	111.76	0.23
13.	2,2-dimethyl-1-propanol	113.1	111.19	-1.69	109.35	-3.31
14.	1-hexanol	157.0	155.76	-0.79	156.92	-0.05
15.	2-hexanol	139.9	140.51	0.43	140.42	0.37
16.	2-methyl-1-pentanol	148.0	146.52	-1.00	146.23	-1.19
17.	4-methyl-1-pentanol	151.8	146.64	-3.40	145.50	-4.15
18.	2-methyl-2-pentanol	121.4	125.67	3.52	124.23	2.33
19.	4-methyl-2-pentanol	131.7	133.22	1.16	131.98	0.21

[a] The values of bp were calculated using equation 4. [b] The values of bp were calculated using equation 10.

Table 2. continued.

	Alkanol	bp(°C) obsd	bp(°C) calcd [a]	Error % [a]	bp(°C) calcd [b]	Error % [b]
20	2-methyl-3-pentanol	126.5	131.24	3.75	131.45	3.91
21.	3-methyl-3-pentanol	122.4	127.08	3.82	127.28	3.99
22.	2-ethyl-1-butanol	146.5	148.55	1.40	149.16	1.81
23.	2,2-dimethyl-1-butanol	136.8	136.68	-0.09	134.80	-1.46
24.	2,3-dimethyl-1-butanol	149.0	141.34	-5.14	140.86	-5.46
25.	3,3-dimethyl-1-butanol	143.0	144.98	1.38	143.19	0.13
26.	3,3-dimethyl-2-butanol	120.0	123.13	2.61	121.03	0.86
27.	1-heptanol	176.3	174.83	-0.84	175.47	-0.47
28.	4-heptanol	155.0	156.98	1.28	158.06	1.98
29.	2-methyl-2-hexanol	142.5	145.42	2.05	142.92	0.29
30.	3-methyl-3-hexanol	142.4	144.34	1.36	146.16	2.64
31.	3-ethyl-3-pentanol	142.5	144.12	1.14	149.63	5.00
32.	2,2-dimethyl-3-pentanol	136.0	138.40	1.76	138.96	2.18
33.	2,4-dimethyl-3-pentanol	138.8	137.81	-0.71	139.76	0.69
34.	2-octanol	179.8	179.32	-0.26	179.76	-0.02
35.	2-Ethyl-1-hexanol	184.6	186.18	0.86	192.21	4.12
36.	2,2,3-trimethyl-3-pentanol	152.2	143.08	-5.99	143.37	-5.80
37.	1-Nonanol	213.1	214.39	0.60	216.34	1.52
38.	2-Nonanol	198.5	196.10	-1.21	194.88	-1.82
39.	4-Nonanol	193.0	190.33	-1.38	191.50	-0.78
40.	5-Nonanol	195.1	190.84	-2.18	191.03	-2.09
41.	7-methyl-1-octanol	206.0	206.48	0.23	205.82	-0.09
42.	2,6-dimethyl-4-heptanol	178.0	179.56	0.88	178.48	0.27
43.	3,5,5-trimethyl-1-hexanol	193.0	196.31	1.71	192.87	-0.07
44.	1-Decanol	230.2	232.73	1.10	230.22	0.01
	Validation set					
45.	2-pentanol	119.0	121.35	1.97	121.05	1.72
46.	2-methyl-2-butanol	102.0	105.51	3.44	104.36	2.32
47.	3-hexanol	135.4	140.32	3.63	141.25	4.32
48.	3-methyl-1-pentanol	152.4	148.85	-2.33	148.75	-2.39
49.	3-methyl-2-pentanol	134.2	133.24	-0.72	133.36	-0.62
50.	2,3-dimethyl-2-butanol	118.6	118.78	0.15	117.21	-1.17
51.	3-heptanol	156.8	156.70	-0.07	158.07	0.81
52.	2,3-dimethyl-3-pentanol	139.0	137.34	-1.19	139.52	0.38
53.	1-octanol	195.2	196.17	0.50	199.23	2.06
54.	3-Nonanol	194.7	189.46	-2.69	191.55	-1.62

[a] The values of bp were calculated using equation 4. [b] The values of bp were calculated using equation 10.

Table 3 Top 9 QSPR models generated by genetic algorithm. The number of compounds used in the fit was 44 for all equations.

1. $\text{b.p.} = 25.55 - 6.460\text{MTI} + 25.911\text{W} - 2.371\text{MR} + 41.540\text{POL} - 2.80\times10^{-3}\text{SAG}^2 + 0.6029\text{MR}^2 - 5.341\text{POL}^2 + 5.09\times10^{-9}\text{V}^3 - 0.2703(\chi^{\text{v}})^4 + 9.15\times10^{-9}\text{SAG}^4$
   (fitness = 0.9928, SD = 5.09, F = 227.97)
2. $\text{b.p.} = 136.53 + 44.551\chi^{\text{v}} + 55.460\log P - 13.953\text{POL} - 2.55\times10^{-3}\text{SAG}^2 + 1.249\text{MR}^2 - 7.535\text{POL}^2 + 7.88\times10^{-6}\text{SAG}^3 - 9.80\times10^{-9}\text{SAG}^4$
   (fitness = 0.9934, SD = 4.76, F = 326.53)
3. $\text{b.p.} = 249.94 + 50.520\chi^{\text{v}} + 0.3407\text{W} - 1.279\text{SAG} + 56.11\log P + 1.343\text{MR}^2 - 8.824\text{POL}^2 + 5.87\times10^{-7}\text{SAG}^3 + 9.02\times10^{-7}\text{V}^3 - 0.2707(\chi^{\text{v}})^4$
   (fitness = 0.9946, SD = 4.35, F = 346.94)
4. $\text{b.p.} = 331.42 - 0.070\text{MTI} + 116.93\chi^{\text{v}} - 1.994\text{SAG} + 61.125\log P - 23.838\text{POL} - 17.143(\chi^{\text{v}})^2 + 3.145\times10^{-3}\text{SAG}^2 + 1.400\text{MR}^2 - 7.878\text{POL}^2 - 1.3\times10^{-9}\text{SAG}^4$
   (fitness = 0.9959, SD = 3.86, F = 396.98)
5. $\text{b.p.} = 49.186 - 4.015\text{MTI} + 16.595\text{W} - 0.2345\text{SAG} + 26.873\log P + 48.436\text{MR} - 101.455\text{POL} - 1.099\text{POL}^2 - 9.6\times10^{-7}\text{SAG}^3 - 3.2\times10^{-7}\text{V}^3$
   (fitness = 0.9931, SD = 4.90, F = 272.88)
6. $\text{b.p.} = 156.53 + 53.810\chi^{\text{v}} + 58.269\log P - 16.170\text{POL} - 3.04\times10^{-3}\text{SAG}^2 + 1.329\text{MR}^2 - 7.845\text{POL}^2 + 5.2\times10^{-6}\text{SAG}^3 - 0.2303(\chi^{\text{v}})^4$
   (fitness = 0.9952, SD = 4.07, F = 447.89)
7. $\text{b.p.} = 119.88 + 44.779\chi^{\text{v}} + 55.481\log P - 18.436\text{POL} - 3.5\times10^{-4}\text{SAG}^2 + 1.234\text{MR}^2 - 7.284\text{POL}^2 - 2.0\times10^{-9}\text{SAG}^4$
   (fitness = 0.9932, SD = 4.73, F = 376.78)
8. $\text{b.p.} = 160.66 + 92.165\chi^{\text{v}} - 0.3031\text{SAG} + 58.231\log P - 27.640\text{POL} - 11.625(\chi^{\text{v}})^2 + 1.285\text{MR}^2 - 7.144\text{POL}^2$
   (fitness = 0.9950, SD = 4.06, F = 514.01)
9. $\text{b.p.} = 171.45 + 96.680\chi^{\text{v}} - 0.4099\text{SAG} + 57.925\log P - 26.939\text{POL} - 12.745(\chi^{\text{v}})^2 + 1.282\text{MR}^2 - 7.132\text{POL}^2 + 4.01\times10^{-7}\text{SAG}^3$
   (fitness = 0.9951, SD = 4.10, F = 440.26)

However, it seems also necessary to use molecular structures to make a correlation study of a property like boiling point because of possibility of taking into consideration the electronic and geometric properties of molecules and their corresponding interactions, in addition to topological properties. Accordingly, and in addition to the three topological properties reported in the previous study,[3] we have chosen as molecular descriptors the following five molecular properties: surface area (grid) (SAG), volume (V), log *P* (log of the octanol-water partition coefficient, which is a measure of hydrophobicity), molar refractivity (MR), and polarizability (POL). In the next step, second- to fourth- orders from each of eight main descriptor was generated. The 32 (=4x8) subfeatures obtained in this way was used as initial set of data from which, and based on a GA procedure, crucial descriptors for performing a proper quantitative structure-property relationship (QSPR) analysis were selected.

Calculation of physicochemical properties:

Calculation of the Schultz index[4-6] is explained by Nikolić et al,[6] the Wiener number by Hosoya,[7,8] and the valence connectivity index by Randić and also by Kier and Hall.[9,10]

The grid calculation of surface area is rather accurate for a given set of atomic radii, and was described by Bodor et al.[11] using the atomic radii of Gavezotti.[12] The volume calculation is very similar to the SAG calculations and employs a grid method described by Bodor. Calculation of log*P* is carried out using atomic parameters first derived by Ghose et al.[13] and extended later by Ghose and coworkers.[14] The molar refractivity is estimated by the same method of computing log *P*. Ghose and Crippen presented atomic contributions to the refractivity in exactly the same way as the hydrophobicity.[14,15] The polarizability is estimated from an additivity scheme given by Miller[16] where different increments are associated with different atom types.

Methods

QSPRs /based on genetic algorithms:

Recently, some published papers suggested that genetic algorithm (GA) is useful in data analysis, especially in the task of reducing the number of features for regression models.[17-23] Roger and Hopfinger first applied this method in QSA(P)R analysis[17] and

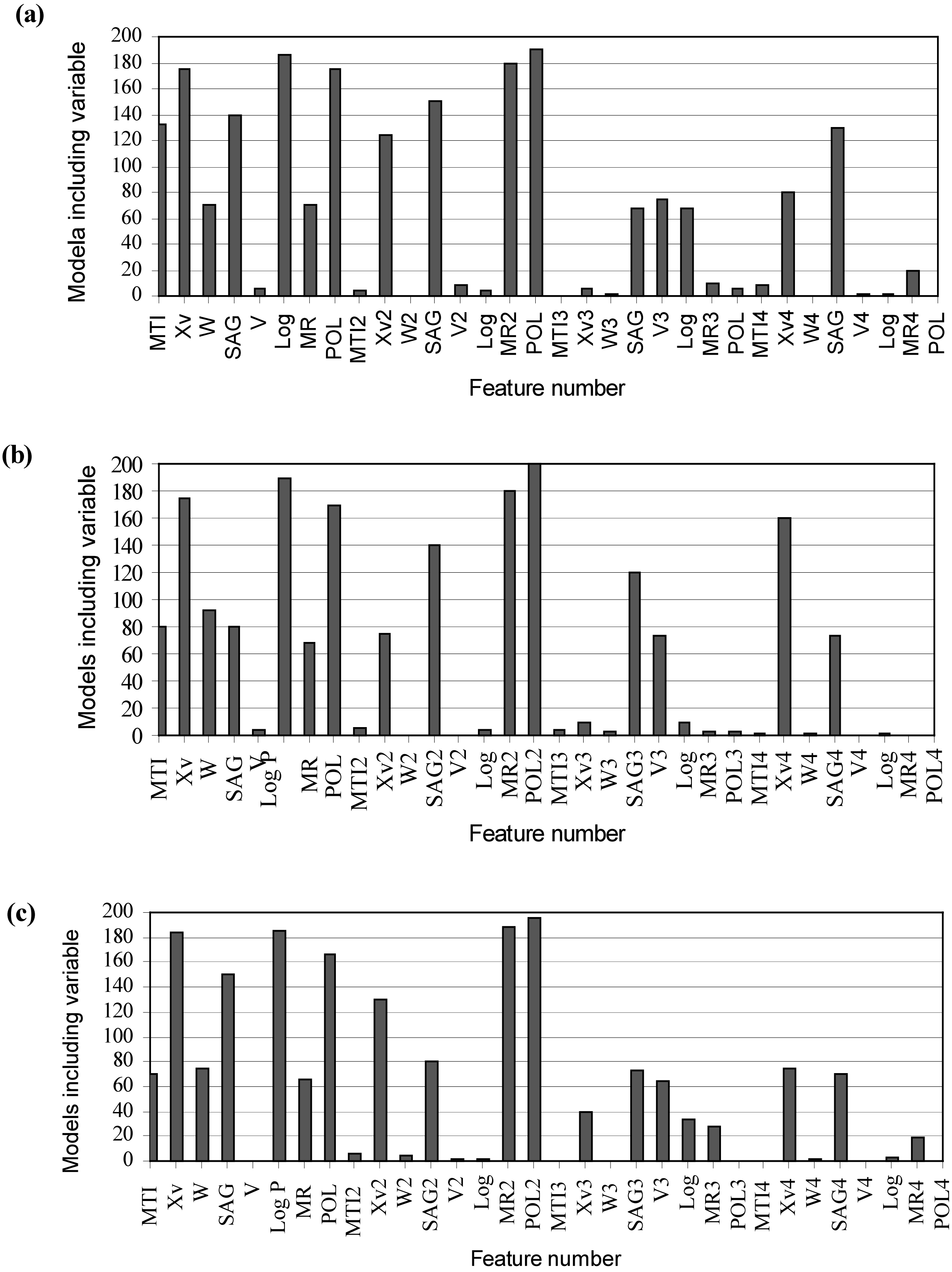

Figure 1. Number of models including each of the features in the elite population, obtained from runing the genetic algorithm for the 4th time(a), 2nd time(b), and 8th time(c) that result to equations 4, 2, and 8 in Table 3.

proved GA a very effective tool and had many merits that other methods did not have. Compared to other traditional methods, QSPRs based on GAs find a group of reliable QSPR models from a large number of sample polynomials. Moreover, from the analysis of the variables used in the evolution procedure, we might obtain the crucial physicochemical properties related to the property.

The QSPR in this study was based on the a GA, coupled with a partial least squares procedure (PLS), which was obtained from PLS-Toolbox of MATLAB.[24] One advantage of using PLS instead of multiple linear regression (MLR) beside GA is the possibility of selecting a number of variables more than the number of samples, that is important when only a small number of samples are utilized in the modeling. The second advantage is the possibility of preparation of models free of errors from the high degrees of collinearity between variables, as will be discussed more in the next section. The brief basic steps of the module are as follows:

Creation of the Initial Population. According to the genetic algorithm, an individual should be represented as a linear string of randomly chosen subfeatures, which plays the role of the DNA for the individuals. The initial population is generated by randomly selecting some number of subfeatures from the data set. Then these individuals are scored according to their fitness score. An elite population is used to retain the best different individuals.

a. Crossover Operation. Once all the models in the population have been rated using the fitness scores, the crossover operation is performed repeatedly. In the operation, two good models are probabilistically selected as "parents" with the likelihood of being chosen proportional to a model fitness score; a pair of children are produced by dividing both parents at a randomly chosen point and then joining the pieces together.

b. Mutation operation. After crossover operation, mutation operation may randomly alter all individuals in the new population, and new model fitness is determined.

c. Comparison Operation. After the crossover and mutation operation, the newly created population and the elite population are compared. If there are some individuals in the newly created population that are better than some individuals in the elite

Table 4 Top fifteen features derived from Figure 1 and Table 3, for the validation set.

Alkanol	MTI	χ^v	W	SAG	log$P$	MR	POL	$(\chi^v)^2$	SAG^2	MR^2	POL^2	SAG^3	V^3	$(\chi^v)^4$	SAG^4
2-pentanol	64.75	1.9509	17.125	245.61	0.96	21.95	8.75	3.806011	60324.27	481.8025	76.5625	14816244	40533878	14.48572	3639017800
2-methyl-2-butanol	114.00	2.4509	30.875	280.88	1.36	26.55	10.59	6.006911	78893.57	704.9025	112.1481	22159627	65032705	36.08298	6224196100
3-hexanol	101.75	2.2843	26.875	265.42	1.04	26.59	10.59	5.218027	70447.78	707.0281	112.1481	18698249	59000163	27.2278	4962889200
3-methyl-1-pentanol	177.25	2.9889	48.625	309.41	1.82	31.08	12.42	8.933523	95734.55	965.9664	154.2564	29621227	91605768	79.80784	9165103700
3-methyl-2-pentanol	177.50	2.9172	48.625	302.92	1.67	31.28	12.42	8.510056	91760.53	978.4384	154.2564	27796099	88103304	72.42105	8419994200
2,3-dimethyl-2-butanol	163.25	2.8616	44.625	296.51	1.76	31.02	12.42	8.188755	87918.18	962.2404	154.2564	26068620	84917111	67.0557	7729606400
3-heptanol	151.00	2.6670	40.625	279.71	1.44	31.06	12.42	7.112889	78237.68	964.7236	154.2564	21883863	72610875	50.59319	6121135200
2,3-dimethyl-3-pentanol	270.50	3.4889	74.375	341.17	2.22	35.68	14.26	12.17242	116397.0	1273.062	203.3476	39711154	128581170	148.1679	13548254000
1-octanol	220.25	3.2276	60.375	301.62	1.91	35.59	14.26	10.41740	90974.62	1266.648	203.3476	27439766	99555414	108.5223	8276382300
3-Nonanol	432.00	4.0233	118.125	372.62	2.53	40.54	16.09	16.18694	138845.7	1643.492	258.8881	51736671	177893100	262.0171	19278119000
2-pentanol	551.00	4.4889	150.875	398.63	3.01	44.88	17.93	20.15022	158905.9	2014.214	321.4849	63344650	230211150	406.0315	25251078000

population, these better individuals are copied to the elite population. When the total fitness of the elite population cannot be improved and about 80% of individuals containing the same subfeatures, “convergence “ is achieved.

Upon completion, from the elite population, the models with the highest fitness scores can be obtained. For a population of 200 models, 20- 50 operations are enough when the data set contains 32 subfeatures. This process takes about 2 min on a PC (Pentium 200).

Reliability of the Models obtained from GA. Most of the models in the elite population had similar fitness scores, after convergence. In this study, the fitness function was defined as the multiple linear regression coefficient (*r*). The reliability of the models were mainly tested with their F-values of their coefficients, as will be discussed in the next part.[25]

Results and Discussion

Construction of the polynomial QSPR models.

The training and validation data sets contained 44 and 10 compounds (Table 2), respectively, and 8 main topological and molecular descriptors. The abbreviations for these descriptors are given in Table 1. For this data set populations with 200 individuals were used. The genetic operator was applied until the total fitness score of the elite populations no longer improved significantly and 70% of individuals include similar subfeatures. The convergence criterion was met after less than 50 operations.

After nine times repeating the GA calculations, nine top seven- to ten-term multiple linear regression models were obtained and are listed in Table 3 as equations 1 to 9. Because a model could not be properly evaluated only by its multiple linear regression coefficient, the quality of the models was tested statistically by the standard error of mean (SD), and overall F statistic for multiple linear regression modeling. The values for the 15 top subfeatures obtained from the nine top models in Table 3 are listed in Table 4.

Generally, for the analysis by MLR, the data must be reduced to fewer and less correlated variables. The cross-correlated descriptors would mislead the QSPR model in uncovering the actual relationship between the property and these descriptors. The correlation study of these subfeatures in the top 9 models are listed in Table 5 and many equations in Table 3 were proven to contain descriptors that were highly cross correlated. Considering the significance of the F-value (at 95% confidence level) for each of the coefficients in the polynomial model, all nine equations from GA were unsatisfactory. The results for coefficients in equations 4 and 6, i.e., the *F* values at the 0.95 confidence level, are listed in Table 6.

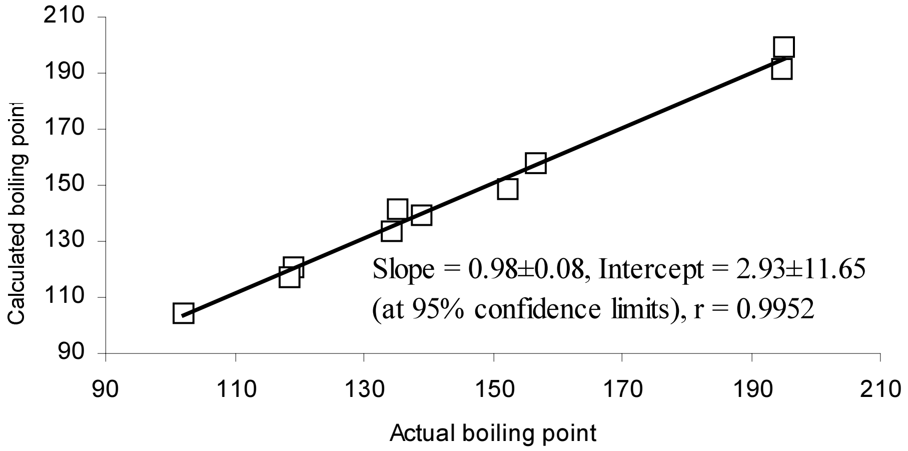

Figure 2. Comparison of actual boiling points with calculated obtained from equation 10 for validation set.

To modify the models into a unique and satisfactory polynomial, and deletion of less importants from the correlated variables, a backward elimination procedure[25] was carried out to a polynomial containing all of the top 15 subfeatures from Tables 3 and 4. The procedure was based on the significance of the F-value (at 95% confidence level) for each of the coefficients in the polynomial model in each step. In this way, equation 10 was obtained, as the most suitable polynomial QSPR model, with all coefficients statistically significant. The predicted boiling point values for all 54 alkanols (44 training set and 10 from validation set) using equation 10 are listed in Table 2. The results for the validation set are also shown in Figure 2.

Principal Features Determined. Figure 1 shows the number of models including each of the subfeatures in the elite population after the convergence for different runs of GA. As illustrated, the appearance frequency of subfeatures in the models due to the final elite populations were quite different from that at the beginning, which was almost equal frequency of appearance for all of the subfeatures. After running GA nine times

Table 5 Squared correlation matrix for top fifteen features in the study.

	MTI	χ^v	W	SAG	log *P*	MR	POL	$(\chi^v)^2$	$(SAG)^2$	$(MR)^2$	$(POL)^2$	$(SAG)^3$	$(V)^3$	$(\chi^v)^4$	$(SAG)^4$
MTI	1.0000														
χ^v	0.9329	1.0000													
W	1.0000	0.9345	1.0000												
SAG	0.9391	0.9830	0.9400	1.0000											
log *P*	0.9023	0.9748	0.9039	0.9642	1.0000										
MR	0.9127	0.9847	0.9141	0.9755	0.9883	1.0000									
POL	0.9090	0.9835	0.9104	0.9727	0.9892	0.9999	1.0000								
$(\chi^v)^2$	0.9868	0.9738	0.9878	0.9679	0.9424	0.9523	0.9496	1.0000							
$(SAG)^2$	0.9732	0.9706	0.9738	0.9906	0.9442	0.9551	0.9515	0.9858	1.0000						
$(MR)^2$	0.9687	0.9741	0.9698	0.9670	0.9677	0.9800	0.9788	0.9857	0.9733	1.0000					
$(POL)^2$	0.9655	0.9736	0.9667	0.9648	0.9696	0.9808	0.9799	0.9837	0.9702	0.9998	1.0000				
$(SAG)^3$	0.9881	0.9457	0.9882	0.9687	0.9135	0.9238	0.9195	0.9858	0.9934	0.9634	0.9596	1.0000			
$(V)^3$	0.9943	0.9493	0.9947	0.9615	0.9249	0.9360	0.9327	0.9919	0.9871	0.9803	0.9778	0.9942	1.0000		
$(\chi^v)^4$	0.9877	0.8810	0.9872	0.8928	0.8436	0.8522	0.8476	0.9621	0.9426	0.9291	0.9251	0.9719	0.9767	1.0000	
$(SAG)^4$	0.9904	0.9154	0.9901	0.9410	0.8789	0.8885	0.8836	0.9751	0.9778	0.9446	0.9402	0.9953	0.9896	0.9868	1.0000

and preparation of models based on the most frequent subfeatures for each run, as shown in Table 3 and Figure 1, Table 4 was obtained which is the list and values of top 15 subfeatures accounted for nearly all the features in the top 9 QSPR models in Table 3. These top 15 subfeatures present all eight features (MTI, χ^V, W, SAG, V, log*P*, MR, and POL) as important factors affecting the boiling point. The appearance frequencies of the other 17 subdescriptors were very low in the elite population and show that these 17 forms of 8 main features are not the effective forms. In the last step, according to a backward elimination procedure the significance of presence for each of fifteen selected variables was tested using F-values of their coefficients in the multiple regression model and a QSPR equation was obtained in which F-values calculated for all of the coefficients were significant. The final model is:

$$bp = 172.87 + 49.23\chi^v - 0.40SAG + 57.84\log P - 18.08POL + 1.30MR^2 - 7.66POL^2 - 0.119(\chi^v)^4 \quad (10)$$

$$(n = 44, r = 0.9945, F = 464.92, SD = 4.26)$$

Statistical results from this final model compare to some of the polynomials obtained from GA are in Table 6 and well illustrate the significance of the final model.

According to the log*P* definition, which stood for lipophilicity of the compound, the positive coefficient of it pointed out that more lipophile alcohols contributed to high boiling points. The positive high value coefficient of Randić's connectivity index χ^v in equation 10 is similar to the previous study by Nikolić et al.[6] It was also suggested from equation 10 that MR, which was the molar refractivity of the molecule, was a necessary contributor to the boiling point. A positive sign of the coefficient for this term indicate that molecular volume and polarizability of the molecules were very vital to the boiling point, in addition to the topology of them. Polarizability (POL) was assigned as an effective variable on boiling point, but with a negative coefficient. Totally the resulting equation illustrates that boiling point could be satisfactorily explained by one topological and four molecular descriptors.

From the GA results in Table 4, the parameters MTI and W and V^3 seem also important to the boiling point. But the correlation studies, listed in Table 5, showed that they are not independent features. MTI and W were highly crosscorrelated with $(\chi^v)^4$, with the correlation coefficient of more than 0.98, and V^3 was highly correlated with

Table 6 The 95% confidence level and F statistics for the coefficient of variables in equations 4, 6, and 10.

Eqn	Variable	Coeff	95% Conf	t-statistic	*F*	Significance
4	MTI	-0.07	±0.22	-0.65	0.43	NS [a]
	χ^v	+116.93	±44.32	+5.37	28.8	S [a]
	SAG	-1.99	±1.38	-2.94	8.65	S [a]
	log*P*	+61.13	±14.05	+8.85	78.30	S [a]
	POL	-23.84	±12.31	-3.94	15.52	S [a]
	$(\chi^v)^2$	-17.142	±8.99	-3.88	15.04	S [a]
	SAG^2	+0.0031	±0.0027	+2.3467	15.46	S [a]
	MR^2	+1.400	±0.317	+8.995	80.91	S [a]
	POL^2	-7.878	±1.641	-9.768	95.41	S [a]
	SAG^4	-1.25×10^{-9}	$\pm6.41\times10^{-9}$	-0.398	0.16	NS [a]
6	χ^v	+53.81	± 10.30	+10.60	112.40	S [b]
	log*P*	+58.27	± 12.72	+9.30	86.52	S [b]
	POL	-16.17	± 7.21	-4.55	20.71	S [b]
	SAG^2	-0.0030	± 0.0020	-2.9785	3.89	NS [b]
	MR^2	+1.329	± 0.215	+12.550	157.49	S [b]
	POL^2	-7.845	± 1.302	-12.233	149.65	S [b]
	SAG^3	$+5.2\times10^{-6}$	$\pm4.4\times10^{-6}$	2.4	149.96	S [b]
	$(\chi^v)^4$	-0.2304	±0.1240	-3.7704	150.05	S [b]
10	χ^v	+49.23	± 9.78	+10.20	104.13	S [c]
	SAG	-0.40	± 0.19	-4.29	18.44	S [c]
	log*P*	+57.84	± 13.28	+8.83	77.99	S [c]
	POL	−18.07	± 6.38	-5.75	33.04	S [c]
	MR^2	+1.301	±0.224	+11.795	139.13	S [c]
	POL^2	-7.665	±1.363	-11.403	130.02	S [c]
	$(\chi^v)^4$	-0.1188	±0.0623	-3.8686	14.97	S [c]

[a] Not significant, compared to one-tailed F(0.05; 1, 33) = 4.35.
[b] Not significant, compared to one-tailed F(0.05; 1, 35) = 4.35.
[c] Significant, compared to one-tailed F(0.05; 1, 36) = 4.35.

MR^2 with the correlation coefficient of 0.98. That is to say, the change of the values of MTI, W, and V^3 were mainly caused by the changes of the $(\chi^v)^4$ and MR^2.

Compared with these seven subfeatures at equation 10, other subfeatures, with high frequencies in the elite populations obtained from GA, contributed a little to the value of boiling point. Addition of these descriptors to equation 10 not only results in no improvements in r value, but also would cause a decrease in the F value of the regression from that of final model (F=464.92). From the correlation study, it could be found that $(SAG)^2$ was highly cross-correlated with SAG, and was not selected in the final polynomial in spite of its high frequency of presence in the elite populations from GA.

Conclusion

In this study we attempted to correlate boiling points of 54 alkanols with toplogical and molecular properties. By using a GA, the polynomial regression models were constructed. These derived models were tested from the viewpoint of statistical significance and from the final statistically significant obtained QSPR polynomial (equation 10), five principal features relevant to the boiling point of alcohols, including χ^V, SAG, log*P*, MR, and POL were obtained. Considering SAG, log*P*, MR, and POL as molecular descriptor in the final model, it could be concluded that the molecular effects, such as surface area, lipophilicity, molecular volume and polarizability, would influence the boiling point of alkanols in addition to the topology of them.

References

1. L. H. Hall, L. B. Kier, *Rev. Comp. Chem.* **1991**, *2*, 367–422.
2. M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1043.
3. E. A. Castro, M. Tueros, http://preprint.chemweb.com/cps/physchem/0110012.
4. H. P. Schultz, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.
5. R. Todeschini, V. Consoni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**, p. 381.
6. S. Nikolić, N. Trinajstić, Z. Mihalić, *J. Math. Chem.* **1993**, *12*, 251–264.
7. H. Wierner, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
8. H. Hosoya, *Bull. Chem. Soc. Japan* **1971**, *44*, 2332–2339.
9. M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
10. L. B. Kier, L. H. Hall, *J. Pharm. Sci.* **1976**, *65*, 1806–1809.
11. N. Bodor, Z. Gabanyi, C. Wong, *J. Am. Chem Soc.* **1989**, *111*, 3783–3786.
12. A. Gavezotti, *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
13. A. K. Ghose, P. Pritchett, G. Crippen, *J. Comput. Chem.* **1988**, *9*, 80–90.
14. V. N. Visvanadhan, A. K. Ghose, G. Revankar, R. K. Robins, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
15. A. K. Ghose, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.
16. K. J. Miller, *J. Am. Chem. Soc.* **1990**, *112*, 8543–8551.
17. D. Roger, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
18. R. Leardi, R. Boggia, M. Terrile, *J. Chemom.* **1992**, *6*, 267–281.
19. R. Leardi, *J. Chemom.* **1994**, *8*, 65–79.
20. T. Hou, X. Xu, *Chemom. Intel. Lab. Syst.* **2001**, *56*, 123–132.
21. B. M. Smith, P. J. Gemperline, *Anal. Chim. Acta* **2000**, *423*, 167–177.
22. R. Leardi, M. B. Seasholtz, R. J. Pell, *Anal. Chim. Acta* **2002**, *461*, 189–200.

23. S. Agatanović-Kuštrin, I. G. Tucker, M. Zečević, L. J. Živanović, *Anal. Chim. Acta* **2000**, *418*, 181–195.
24. B. M. Wise, N. B. Gallagher, *PLS-Toolbox*, ver. 2.0, Eigenvector Research, Inc., Natick, MA, **1995**.
25. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam, **1997**, p. 280.

Povzetek

Pri QSPR študiji vrelišč alkoholov so bili uporabljeni Schulzov indeks, Randićev indeks, Wienerjevo število, površina molekule, njena prostornina, logP, molska refrakcija in polarizabilnost. Za vrelišča 44 alkoholov so bile s pomočjo genetskega algoritma v povezavi s PLS metodo preizkušene vse možne povezave med vrelišči in lastnostmi molekul do četrte potence. Ustvarjena je bila skupina modelov na osnovi multiple regresije z visoko stopnjo ujemanja z vrelišči. Z metodo vzvratnega odstranjevanja so bili izbrani kot pomembni deskriptorji Randićev indeks, površina molekule, logP, molska refrakcija in polarizabilnost. Izbrani deskriptorji in njihov pravi vrstni red zelo dobro ocenijo vrelišča alkoholov, zlasti pri uporabi višje (druge, tretje in/ali četrte) potence.