J Mol Model (2007) 13: 55–64
DOI 10.1007/s00894-006-0125-z

ORIGINAL PAPER

Georgia Melagraki · Antreas Afantitis · Haralambos Sarimveis · Panayiotis A. Koutentis · John Markopoulos · Olga Igglessi-Markopoulou

A novel QSPR model for predicting *θ* (lower critical solution temperature) in polymer solutions using molecular descriptors

Received: 19 December 2005 / Accepted: 16 March 2006 / Published online: 24 May 2006

Abstract In this study, we present a new model that has been developed for the prediction of θ (lower critical solution temperature) using a database of 169 data points that include 12 polymers and 67 solvents. For the characterization of polymer and solvent molecules, a number of molecular descriptors (topological, physicochemical, steric and electronic) were examined. The best subset of descriptors was selected using the elimination selection-stepwise regression method. Multiple linear regression (MLR) served as the statistical tool to explore the potential correlation among the molecular descriptors and the experimental data. The prediction accuracy of the MLR model was tested using the leave-one-out cross-validation procedure, validation through an external test set and the *Y*-randomization evaluation technique. The domain of applicability was finally determined to identify the reliable predictions.

Keywords Lower critical solution temperature · QSPR · Molecular descriptors

G. Melagraki · A. Afantitis · H. Sarimveis (✉) · O. Igglessi-Markopoulou
School of Chemical Engineering,
National Technical University of Athens,
Athens, Greece
e-mail: hsarimv@central.ntua.gr
Tel.: +30-210-7723237
Fax: +30-210-7723138

A. Afantitis
Department of ChemoInformatics, NovaMechanics Limited,
Nicosia, Cyprus

P. A. Koutentis
Department of Chemistry, University of Cyprus,
P.O. Box 20537,
1678 Nicosia, Cyprus

J. Markopoulos
Department of Chemistry, University of Athens,
Athens, Greece

Introduction

The phase behavior of polymer solutions is an important property involved in the development and design of most polymer-related processes. Partially miscible polymer solutions often exhibit two solubility boundaries, the upper critical solution temperature (UCST) and the lower critical solution temperature (LCST), which both depend on the molar mass and the pressure. At temperatures below LCST, the system is completely miscible in all proportions, whereas above LCST partial liquid miscibility occurs [1, 2]. θ(LCST) is the LCST at infinite chain length, which is not affected by polymer molar mass. θ(LCST) is often regarded as less important than θ(UCST) because it is usually located at high temperature, where the polymer degenerates. Nevertheless, θ(LCST) can serve as an upper temperature limit for polymer processing. Solvent systems that exhibit this LCST behavior have been suggested for applications where partial or complete miscibility above and below LCST offers advantages [3, 4].

There are three groups of methods for correlating and predicting LCSTs. The first group proposes models that are based on a solid theoretical background using liquid–liquid or vapor–liquid experimental data. These methods require experimental data to adjust the unknown parameters, resulting in limited predictive ability [5–7]. Another approach uses empirical equations that correlate θ(LCST) with physicochemical properties such as density, critical properties etc., but suffers from the disadvantage that these properties are not always available [8–10]. A new approach proposed by Liu and Zhong develops linear models for the prediction of θ(LCST) using molecular connectivity indices, which depends only on the solvent and polymer structures [11, 12]. The latter approach has proven to be a very useful technique in quantitative structure–activity/property relationships (QSAR/QSPR) research for polymers and polymer solutions. QSAR/QSPR studies constitute an attempt to reduce the trial-and-error element in the design of compounds with desired activity/properties by establishing mathematical relationships between the activity/property of interest and measurable or computable

parameters, such as topological, physicochemical, stereochemistry, or electronic indices [13–18].

The present study follows the third approach and its goal is to develop a new, more efficient QSPR model for the prediction of θ(LCST) in polymer solutions, using a newly introduced set of molecular descriptors. The model is developed using a rigorous variable-selection technique and the multiple linear regression (MLR) modeling methodology. The accuracy of the produced model is illustrated using numerous model validation techniques.

Materials and methods

A large set of 169 experimental θ(LCST) data was collected from the literature [12] comprising 12 polymers and 67 solvents. The polymers and solvents are shown in Tables 1 and 2. There were 37 topological, physicochemical, steric and electronic descriptors considered as potential descriptors (inputs) to the QSPR model. The descriptors were calculated from the structure of the monomer compound using ChemSar, which is included in the ChemOffice (CambridgeSoft Corporation) suite of programs [19]. All structures were fully optimized using MOPAC (included in ChemOffice) and, more specifically, the AM1 Hamiltonian, which provides balance between speed and accuracy. Eight of the descriptors were the topological descriptors calculated and used by Liu and Zhong [12]. Table 3 shows the remaining 29 descriptors for both the polymers and the solvents.

Stepwise multiple regression

Among the aforementioned indices, the best combinations were determined using a rigorous elimination selection-stepwise regression (ES-SWR) algorithm that was programmed in Matlab. The aim of variable subset selection is to reach optimal model complexity in predicting a response variable using a reduced set of descriptors that are not highly intercorrelated. In particular, the objective of this

Table 1 Polymers used in this work

Polymer ID	Polymer
A	Polyethylene
B	Polypropylene
C	Polybut-1-ene
D	Polyisobutylene
E	Polypent-1-ene
F	Poly(4-methylpent-1-ene)
G	Poly(*cis*-1,4-butadiene)
H	Polystyrene
I	Poly(α-methylstyrene)
J	Poly(p-chrolostyrene)
K	Poly(dimethylsiloxane)
L	Poly(isotactic methylmethacrylate)

Table 2 Solvents used in this work

Solvent ID	Solvent
1	*n*-Butane
2	*n*-Pentane
3	*n*-Hexane
4	*n*-Heptane
5	*n*-Octane
6	*n*-Nonane
7	*n*-Decane
8	*n*-Undecane
9	*n*-Dodecane
10	*n*-Tridecane
11	*n*-Cetane
12	2-Methylbutane
13	2,2-Dimethylbutane
14	2,3-Dimethylbutane
15	2-Methylpentane
16	3-Methylpentane
17	2,4-Dimethylpentane
18	2,3-Dimethylpentane
19	2,2-Dimethylpentane
20	3,3-Dimethylpentane
21	2,2,3-Trimethylbutane
22	2-Methylhexane
23	3-Methylhexane
24	2,2,4-Trimethylpentane
25	2-Methylheptane
26	3-Methylheptane
27	2,2-Dimethylhexane
28	2,4-Dimethylhexane
29	2,5-Dimethylhexane
30	3,4-Dimethylhexane
31	3-Ethylpentane
32	2,2,4,4-Tetramethylpentane
33	2,3,4-Trimethylhexane
34	Cyclopentane
35	Cyclohexane
36	Cycloheptane
37	Cyclooctane
38	Methylcyclopentane
39	Methylcyclohexane
40	Ethylcyclopentane
41	*n*-Propylcyclopentane
42	Benzene
43	Toluene
44	Methyl acetate
45	Ethyl acetate
46	*n*-Propyl acetate
47	*i*-Propyl acetate
48	*n*-Butyl acetate
49	*i*-Butyl acetate
50	*sec*-Butyl acetate
51	*tert*-Butyl acetate
52	*n*-Pentyl acetate
53	*i*-Pentyl acetate
54	*n*-Hexyl acetate

Table 2 (continued)

Solvent ID	Solvent
55	Ethyl *n*-butyrate
56	Methyl ethyl ketone
57	Diethyl ketone
58	Ethyl propyl ketone
59	Dipropyl ketone
60	Diethyl ether
61	Diethyl malonate
62	1-Octanol
63	Ethyl carbinol
64	*n*-Butyl carbinol
65	Propylene oxide
66	Butyl chloride
67	Tetrahydrofuran (THF)

work was to select the subset of variables that produces the most significant linear QSPR model as far as prediction of θ(LCST) is concerned.

ES-SWR is a popular stepwise technique that combines forward selection (FS)-SWR and backward elimination (BE)-SWR. It is basically a forward-selection approach but, at each step, it considers the possibility of deleting a variable, as in the backward-elimination approach, provided that the number of model variables is greater than two. The two basic elements of the ES-SWR method are described next in more detail.

Table 3 Descriptors

ID	Description	Notation
1	Molar refractivity	MR
2	Diameter	Diam
3	Partition coefficient (octanol water)	ClogP
4	Molecular topological index	TIndx
5	Principal moment of inertia *Z*	PMIZ
6	Number of rotatable bonds	NRBo
7	Principal moment of inertia *Y*	PMIY
8	Polar surface area	PSAr
9	Principal moment of inertia *X*	PMIX
10	Radius	Rad
11	Connolly accessible area	SAS
12	Shape attribute	ShpA
13	Connolly molecular area	MS
14	Shape coefficient	ShpC
15	Total energy	TotE
16	Sum of valence degrees	SVDe
17	Electronic energy	ElcE
18	Sum of degrees	SDeg
19	LUMO energy	LUMO
20	Total connectivity	TCon
21	HOMO energy	HOMO
22	Total valence connectivity	TVCon
23	Balaban index	Bindx
24	Wiener index	WIndx
25	Cluster count	ClsC
26	Repulsion energy	NRE
27	Dipole length	DPLL
28	Connoly solvent_excluded volume	SEV
29	Ovality	Ovality

Forward selection

The variable considered for inclusion at any step is the one yielding the largest single degree of freedom *F*-ratio among those eligible for inclusion. The variable is included only if this value is larger than a fixed value F_{in}. At each step, the *j*th variable is consequently added to a *k*-size model if

$$F_j = \max_j \left(\frac{RSS_k - RSS_{k+j}}{s^2_{k+j}} \right) > F_{in} \tag{1}$$

In the above inequality, RSS is the *residual sum of squares* and *s* is the *mean square error*. The subscript $k+j$ refers to quantities computed when the *j*th variable is added to the *k* variables that are already included in the model.

Backward elimination

The variable considered for elimination at any step is the one yielding the minimum single degree of freedom *F*-ratio among the variables included in the model. The variable is eliminated only if this value does not exceed a specified value F_{out}. At each step, the *j*th variable is eliminated from a *k*-size model if

$$F_j = min_j \left(\frac{RSS_{k-j} - RSS_k}{s^2_k} \right) < F_{out} \tag{2}$$

The subscript $k-j$ refers to quantities computed when the *j*th variable is eliminated from the *k* variables included in the model so far.

Model validation

A reliable and predictive QSPR model should (1) be statistically significant and robust, (2) provide accurate predictions for external data sets not used during the model development, and (3) have its application boundaries defined. The approaches used in this work to ensure the significance and predictive power of the QSPR model are described below.

Cross-validation technique

To explore the reliability of the proposed method, we used the cross-validation method. Based on this technique, a

number of modified data sets are created by deleting, in each case, one or a small group (leave-some-out) of objects [20–22]. For each data set, an input–output model is developed, based on the modeling technique used. Each model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones not used to develop the model). In particular, the leave-one-out (LOO) procedure was used in this study. It produces a number of models by deleting each time one object from the training set. The number of models produced by the LOO procedure is obviously equal to the number of available examples n. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the R^2_{CV} and S_{PRESS} values can be calculated easily. The formulae used to calculate all the aforementioned statistics are presented below (Eqs. 3 and 4):

$$R^2_{CV} = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i-1}^{n} \left(y_{\exp} - y_{pred}\right)^2}{\sum_{i=1}^{n} \left(y_{\exp} - \bar{y}\right)^2} \tag{3}$$

$$S_{PRESS} = \sqrt{\frac{PRESS}{n}} \tag{4}$$

Y-randomization test

This technique ensures the robustness of a QSPR model [23, 24]. The dependent variable vector (property) is randomly shuffled and a new QSPR model is developed using the original independent variable matrix. The new QSPR models (after several repetitions) are expected to have low R^2 and R^2_{CV} values. If the opposite happens, then an acceptable QSPR model cannot be obtained for the specific modeling method and data.

Estimation of the predictive ability of a QSPR model

According to Tropsha et al. [24], the predictive power of a QSPR model can be estimated conveniently by an external $R^2_{CV,ext}$ (Eq. 5).

$$R^2_{CV,ext} = 1 - \frac{\sum_{i=1}^{test} \left(y_{\exp} - y_{pred}\right)^2}{\sum_{i=1}^{test} \left(y_{\exp} - \bar{y}_{tr}\right)^2} \tag{5}$$

where $\bar{y}_{tr}$ is the averaged value for the dependent variable for the training set.

Furthermore, the same group [24, 25] considered a QSPR model predictive if the following conditions are satisfied:

$$R^2_{CV,ext} > 0.5 \tag{6}$$

$$R^2_{pred} > 0.6 \tag{7}$$

$$\frac{\left(R^2 - R^2_o\right)}{R^2} < 0.1 \quad \text{or} \quad \frac{\left(R^2 - R'^2_o\right)}{R^2} < 0.1 \tag{8}$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{9}$$

The mathematical definitions of R^2_o, R'^2_o, k and k' are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in [25] and are not repeated here for brevity.

Defining the model-applicability domain

In order for a QSPR model to be used for the prediction of θ(LCST) of new systems, its domain of application [24, 26] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of extrapolation* [24] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [27] for each chemical, where the QSPR model is used to predict its activity:

$$h_i = x_i^T \left(X^T X\right) x_i \tag{10}$$

In Eq. 10, x_i is the descriptor-row vector of the query compound and X is the $k \times n$ matrix containing the k descriptor values for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large and means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

Results and discussion

The ES-SWR variable-selection technique was used to select a subset of the available chemical descriptors that is most meaningful and statistically significant in terms of correlation with the θ(LCST). The variable-selection procedure identified nine descriptors that significantly influence θ(LCST) and characterize both the polymer and the solvent.

In particular, for the polymer, the following descriptors were selected: HOMO energy, shape coefficient (ShpC), dipole length (DPLL), radius (Rad), and the polymer third-order connectivity index contributed by the side groups ($^3\chi^{SG}$-topological). For the solvent, four descriptors were chosen: sum of degrees (SDeg), electronic energy, dipole length, and solvent third-order connectivity index ($^3\chi_p$-topological).

The above descriptors are defined as follows: HOMO energy (HOMO) is the energy of the highest occupied molecular orbital. According to frontier-orbital theory, the shapes and symmetries of the highest occupied molecular orbital HOMO are crucial in predicting the molecule's reactivity. The electronic energy (ElcE) is the total electronic energy given in electron volt at 0 °C. Dipole length is the electric dipole moment divided by the elementary charge. Electric dipole is a vector quantity that encodes displacement with respect to the center of gravity of positive and negative charges in a molecule. The radius is the minimum such value and is held by the most central atom(s). The shape coefficient is given by: ShpC= (D−Rad)/Rad, where the diameter (D) is the maximum such value for all atoms and is held by the most outlying atom(s). The sum-of-degrees is the sum of the degrees of every atom. The polymer third-order connectivity index contributed by the side groups ($^3\chi^{SG}$) and solvent third-order connectivity index ($^3\chi_p$) are described in [12].

The full linear equation for the prediction of θ(LCST) for all systems in the data set is the following:

$$\begin{aligned}\theta(LCST)/K = {} & 31.5(\pm 3.7) * DPLL(solvent) \\ & - 38.7(\pm 6.6) *^3 \chi^{SG} \\ & + 49.3(\pm 4.9) * Rad \\ & - 92.0(\pm 8.0) * DPLL(polymer) \\ & + 99.9(\pm 7.3) * ShpC \\ & + 46.7(\pm 5.9) *^3 \chi_p \\ & + 0.0351(\pm 0.00332) * ElcE \\ & - 105.6(\pm 6.4) * HOMO \\ & + 30.6(\pm 2.3) * SDeg \\ & - 931.6(\pm 68.88)\end{aligned} \tag{11}$$

To evaluate the performance of the QSPR model presented in this work, the data set was split randomly into a training and a validation set in a ratio of approximately 65:35% (112 and 57 systems, respectively) according to [12]. The training and validation compounds are clearly indicated in Tables 4 and 5. The validation set was not involved in any way during the training phase. The

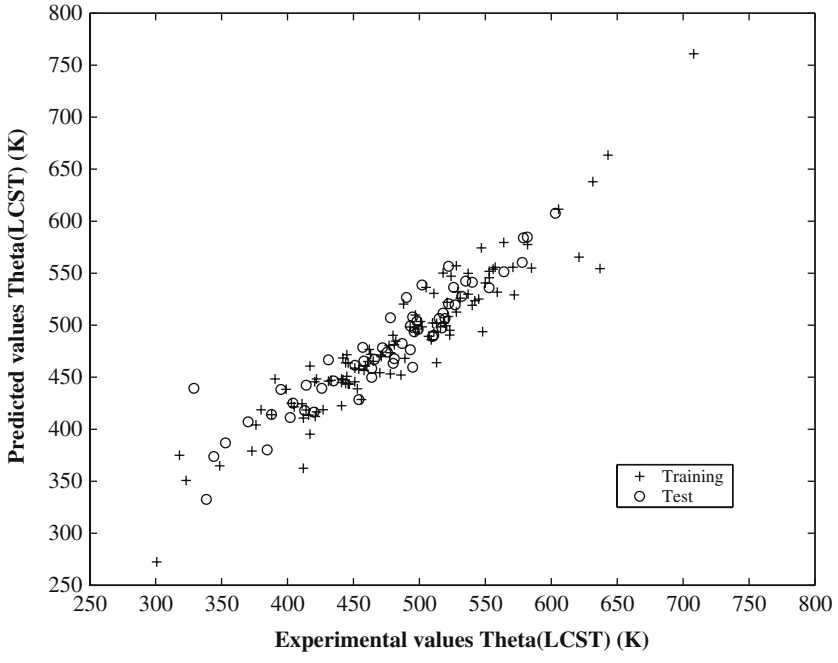

Fig. 1 Predicted vs experimental values for the training and test sets

Table 4 Experimental and predicted values, absolute/relative errors for the training set

System ID	Experimental values θ(LCST)/K	Predicted values θ(LCST)/K	Absolute error (K)	Relative error (%)
A3	411	424.59	13.59	3.31
A4	459	460.67	1.67	0.36
A6	531	526.00	5.00	0.94
A7	557.55	555.67	1.88	0.34
A9	605.55	611.55	6.00	0.99
A10	631.55	637.99	6.44	1.02
A19	399	438.33	39.33	9.86
A21	444	464.03	20.03	4.51
A18	463	472.02	9.02	1.95
A31	471	469.77	1.23	0.26
A32	513	464.01	48.99	9.55
A33	545	525.12	19.88	3.65
A38	488	520.33	32.33	6.62
A39	537	549.80	12.80	2.38
A62	621.1	565.53	55.57	8.95
A52	528	557.13	29.13	5.52
B2	422	416.73	5.27	1.25
B3	470	454.35	15.65	3.33
B5	542	523.58	18.42	3.40
B6	571	555.76	15.24	2.67
B13	441	448.43	7.43	1.68
B14	465	465.22	0.22	0.05
B19	489	468.09	20.91	4.28
B21	511	493.79	17.21	3.37
B18	513	501.78	11.22	2.19
B31	520	499.53	20.47	3.94
B32	548	493.77	54.23	9.90
B33	585	554.89	30.12	5.15
B38	518	550.09	32.09	6.19
B39	564	579.57	15.57	2.76
B60	420.6	445.61	25.01	5.95
C2	421	412.31	8.69	2.06
C4	509	486.01	22.99	4.52
C5	540	519.16	20.84	3.86
C12	416	413.62	2.38	0.57
C13	444	444.01	0.01	0.00
C21	507	489.37	17.63	3.48
C29	519	502.28	16.72	3.22
C31	523	495.11	27.89	5.33
C30	559	531.82	27.18	4.86
D34	461	465.11	4.11	0.89
D35	516	498.41	17.59	3.41
D36	572	529.16	42.84	7.49
D37	637	554.45	82.55	12.96
D4	442	447.41	5.41	1.22
D5	477	480.56	3.56	0.75
D9	582	577.59	4.41	0.76
D12	318	375.03	57.03	17.93
D15	376	404.15	28.15	7.49
D16	405	421.55	16.55	4.09
D23	446	443.38	2.62	0.59

Table 4 (continued)

System ID	Experimental values θ(LCST)/K	Predicted values θ(LCST)/K	Absolute error (K)	Relative error (%)
D31	458	456.51	1.49	0.32
D18	451	458.76	7.76	1.72
D17	403	424.90	21.90	5.43
D21	445	450.78	5.78	1.30
D40	524	547.11	23.11	4.41
D26	478	473.57	4.43	0.93
D27	454	457.81	3.81	0.84
D29	446	463.68	17.68	3.96
D30	497	493.22	3.78	0.76
D41	547	574.31	27.31	4.99
E2	433	447.07	14.07	3.25
E3	482	484.70	2.70	0.56
E5	556	553.93	2.07	0.37
E12	422	448.39	26.39	6.25
E17	493	498.27	5.27	1.07
E19	502	498.44	3.56	0.71
E18	529	532.13	3.13	0.59
E31	537	529.88	7.12	1.33
F1	388	414.05	26.05	6.71
F2	441	444.97	3.97	0.90
F4	522	508.32	13.68	2.62
F5	553	551.82	1.18	0.21
F12	431	446.28	15.28	3.55
F13	462	476.66	14.66	3.17
F19	499	496.33	2.67	0.53
F21	521	522.03	1.03	0.20
F34	505	536.37	31.37	6.21
G3	373	379.07	6.07	1.63
G21	414	418.52	4.52	1.09
G5	390.5	448.30	57.80	14.80
G58	510	502.26	7.74	1.52
G57	481	481.45	0.45	0.09
H34	427	418.82	8.18	1.91
H35	486	452.13	33.87	6.97
H38	417	460.78	43.78	10.50
H39	480	490.26	10.26	2.14
H42	523	490.51	32.49	6.21
H43	550	540.60	9.40	1.71
H45	412	410.75	1.25	0.30
H46	451	445.60	5.40	1.20
H47	380	418.73	38.73	10.19
H55	471	471.41	0.41	0.09
H49	445	471.62	26.62	5.98
H50	442	468.47	26.47	5.99
I34	417	395.30	21.70	5.20
I35	456	428.60	27.40	6.01
I66	412	362.52	49.48	12.01
I48	446.9	443.47	3.43	0.77
I54	500.9	503.49	2.59	0.52
J49	348.5	364.88	16.38	4.70
J63	300.8	272.48	28.32	9.42

Table 4 (continued)

System ID	Experimental values θ(LCST)/K	Predicted values θ(LCST)/K	Absolute error (K)	Relative error (%)
J64	323.1	350.73	27.63	8.55
K2	453	438.87	14.13	3.12
K4	528	512.58	15.42	2.92
K5	553	545.73	7.27	1.31
K9	643	663.46	20.46	3.18
K11	708	760.90	52.90	7.47
L44	441	422.60	18.40	4.17
L45	478	453.12	24.88	5.20
L58	511	530.60	19.60	3.84
L57	497	509.79	12.79	2.57

Table 5 Experimental and predicted values, absolute/relative errors for the test set

System ID	Experimental values θ(LCST)/K	Predicted values θ(LCST)/K	Absolute error (K)	Relative error (%)
A2	353	386.97	33.97	9.62
A5	496	493.82	2.18	0.44
A8	581.75	584.72	2.97	0.51
A17	395	438.16	43.16	10.93
A24	495	459.63	35.37	7.15
A30	515	506.48	8.52	1.65
A34	472	478.37	6.37	1.35
A35	518	511.67	6.33	1.22
A48	490	526.54	36.54	7.46
B4	511	490.43	20.57	4.03
B12	413	418.04	5.04	1.22
B17	481	467.92	13.08	2.72
B24	510	489.39	20.61	4.04
B30	553	536.24	16.76	3.03
B34	495	508.13	13.13	2.65
B35	540	541.43	1.43	0.27
C3	464	449.93	14.07	3.03
C6	564	551.34	12.66	2.25
C17	480	463.50	16.50	3.44
C18	517	497.36	19.64	3.80
C34	498	503.71	5.71	1.15
D2	344	373.71	29.71	8.64
D3	402	411.33	9.33	2.32
D38	478	507.07	29.07	6.08
D22	426	439.48	13.48	3.16
D19	404	425.07	21.07	5.22
D20	451	461.52	10.52	2.33
D39	526	536.55	10.55	2.01
D25	466	467.20	1.20	0.26
D28	458	465.55	7.55	1.65
D24	435	446.37	11.37	2.61
D7	535	542.41	7.41	1.39
E4	522	520.77	1.23	0.24
E13	457	478.77	21.77	4.76

Table 5 (continued)

System ID	Experimental values θ(LCST)/K	Predicted values θ(LCST)/K	Absolute error (K)	Relative error (%)
E24	527	519.73	7.27	1.38
E34	502	538.47	36.47	7.27
F3	487	482.59	4.41	0.91
F6	579	583.99	4.99	0.86
F17	499	496.16	2.84	0.57
F31	532	527.77	4.23	0.80
G22	370	407.22	37.22	10.06
G24	388	414.11	26.11	6.73
G65	414	442.30	28.30	6.84
H51	329	439.43	110.43	33.56
H61	578	560.49	17.51	3.03
H53	493	499.27	6.27	1.27
H44	384.5	380.23	4.27	1.11
H56	420	416.33	3.67	0.87
I39	431	466.74	35.74	8.29
I52	475.8	474.06	1.74	0.36
J51	338.4	332.69	5.71	1.69
K3	493	476.50	16.50	3.35
K7	603	607.58	4.58	0.76
L66	454	428.41	25.59	5.64
L59	522	556.65	34.65	6.64
L56	464	458.70	5.30	1.14
L67	519.5	506.20	13.30	2.56

full linear equation that was developed using only the 112 training data is the following:

$$\begin{aligned}\theta(LCST)/K = {} & 33.0(\pm 4.8) * DPLL(solvent) \\ & - 39.8(\pm 8.4) *^3 \chi^{SG} \\ & + 47.3(\pm 6.3) * Rad \\ & - 92.1(\pm 10.4) * DPLL(polymer) \\ & + 95.7(\pm 9.4) * ShpC \\ & + 41.4(\pm 7.5) *^3 \chi_p \\ & + 0.0371(\pm 0.0041) ElcE \\ & - 99.1(\pm 6.4) * HOMO \\ & + 32.2(\pm 2.7) * SDeg \\ & - 860.7(\pm 89.9) \\ & n = 112,\ R^2_{CV} = 0.8546,\ R^2_{tr} = 0.8860, \\ & R^2_{pred} = 0.8738,\ F = 88.04, \\ & RMS_{tr} = 23.4806,\ RMS_{pred} = 23.7893, \\ & S_{PRESS} = 25.5095 \end{aligned} \tag{12}$$

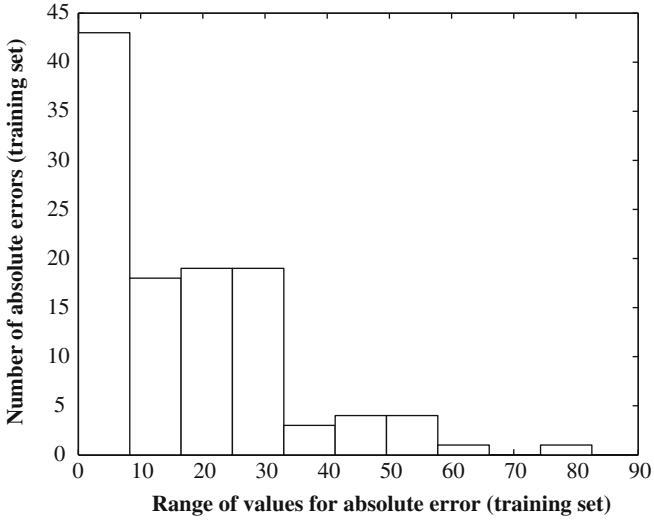

Fig. 2 Distribution of absolute errors for the training set

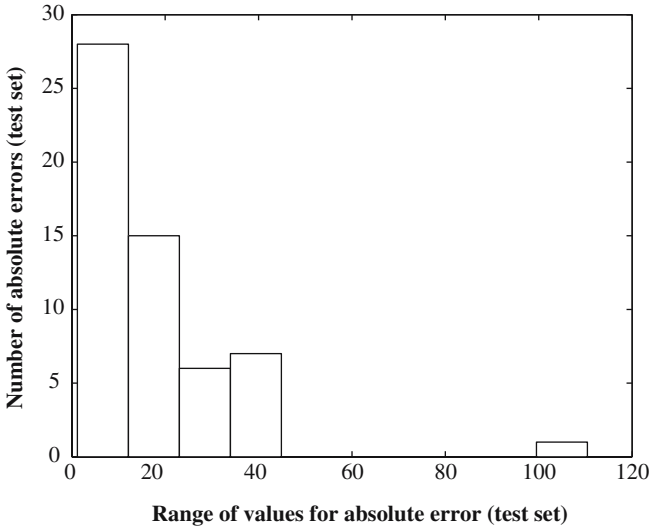

Fig. 4 Distribution of absolute errors for the test set

The results are shown in Tables 4 and 5, where the prediction of the QSPR model is shown for both the training and the external examples. The corresponding absolute and relative errors are also indicated in Tables 4 and 5. The experimental vs predicted values for the training and test sets are shown graphically in Fig. 1. Figures 2, 3, 4, and 5 show the distributions of absolute and relative errors for the training and the test sets. The average absolute (relative) error in predicting θ(LCST) for the training set is 17.57K (3.73%) and the corresponding value for the validation set is 16.59K (3.82%). There is a clear improvement compared to the model described in [12], where the corresponding average relative errors were 5.44 and 5.63% for the training and the validation sets, respectively.

The min/max values of the absolute errors are 0/82.55 and 1.20/110.43 for the training and validation sets, respectively. As far as the relative error is concerned, the corresponding statistics are 0/17.93 and 0.23/33.56 for the training and validation sets, respectively. Among the systems in the training set, 41 out of 112 have an absolute error of more than 20K. System D37 (PIB-cyclooctane) has the maximum absolute error (82.54K). This can be explained by the fact that the solvent was not included in any other system in the training set. Among the systems in the test set, 18 out of 57 have an absolute error of more than 20K. System H51 (PS-*tert*-butyl acetate) has the maximum absolute error (110.42K). The large error is due to the fact that the particular solvent was not included in any system during the training procedure. Only six systems in the training set have relative errors greater than 10% and there are no systems with a relative error greater than 20%. In the test set, there are only two systems having a relative error greater than 10% and only one system (PS-*tert*-butyl acetate) has a relative error of 33.5%.

We should point out that the experimental data were collected from different sources that might have used different measuring methods (with different measurement accuracies and systematic errors). As stated by Liu and

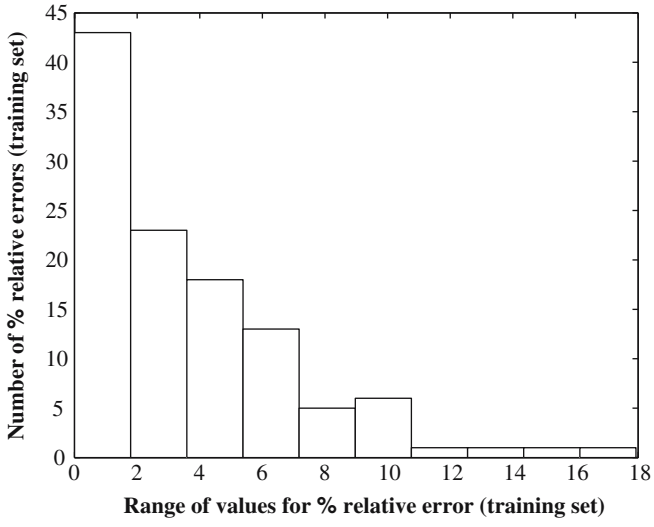

Fig. 3 Distribution of percent of relative errors for the training set

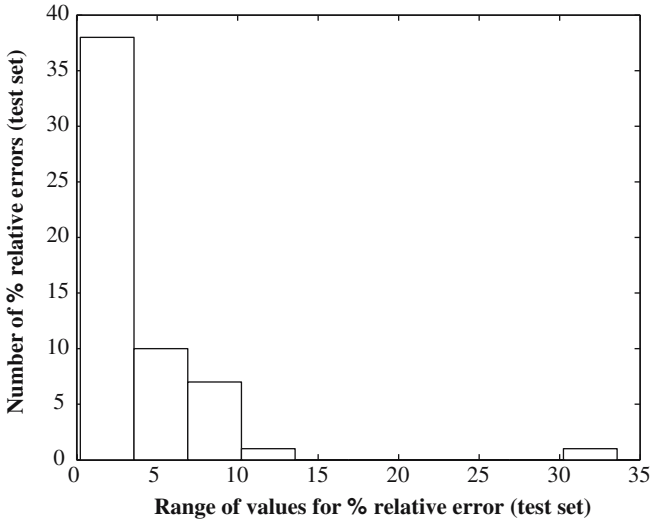

Fig. 5 Distribution of percent of relative errors for the test set

Table 6 Results of the *Y*-randomization test

Iteration	R^2	R^2_{CV}
1	0.1307	0.00
2	0.0248	0.00
3	0.0763	0.00
4	0.0634	0.00
5	0.0857	0.00
6	0.0586	0.00
7	0.0473	0.00
8	0.0493	0.00
9	0.1289	0.00
10	0.0995	0.00

Table 7 Leverages for the test set

System ID	Leverages
A2	0.1156
A5	0.0662
A8	0.0776
A17	0.0934
A24	0.0986
A30	0.1125
A34	0.1161
A35	0.1055
A48	0.1183
B4	0.0304
B12	0.0456
B17	0.0450
B24	0.0636
B30	0.0895
B34	0.0667
B35	0.0668
C3	0.0561
C6	0.0561
C17	0.0630
C18	0.0640
C34	0.0891
D2	0.0629
D3	0.0482
D38	0.0692
D22	0.0457
D19	0.0587
D20	0.0572
D39	0.0698
D25	0.0461
D28	0.0462
D24	0.0807
D7	0.0626
E4	0.0386
E13	0.0407
E24	0.0661
E34	0.0778
F3	0.0438
F6	0.0591
F17	0.0515
F31	0.0666
G22	0.1435
G24	0.1752
G65	0.2903
H51	0.1098
H61	0.1685
H53	0.0992
H44	0.0947
H56	0.1597
I39	0.0868
I52	0.0836
J51	0.3002
K3	0.0868
K7	0.0991
L66	0.1716
L59	0.2087
L56	0.2095
L67	0.1979

Zhong [12] the experimental uncertainties in θ(LCST) can be large, depending on the measurement methods, polymer samples used and the techniques followed by different researchers. Taking these facts into account, we can conclude that our approach models θ(LCST) successfully and has a significant predictive potential.

Several other statistics calculated based on Eq. 12 illustrate the efficiency of the QSPR model. The coefficients of determination (R^2 values) given above indicate a high correlation between experimental and predicted values. R^2_{CV} (the result of the LOO cross-validation procedure) is particularly high ($R^2_{CV} = 0.8546 > 0.5$), showing that the model has a high predictive ability and is also robust. As mentioned before, the calculation of this statistic is based on a number of modified data sets created by deleting, in each case, one object from the data. An MLR model is developed based on the remaining data and is validated using the deleted object. For our particular training set, 112 MLR models were obviously built by deleting each time one compound from the training set.

The proposed model also passed all the tests defined by Eqs. 6, 7, 8, and 9):

$$R^2_{CV,ext} = 0.8634 > 0.5$$

$$R^2_{pred} = 0.8738 > 0.6$$

$$\frac{\left(R^2 - R^2_o\right)}{R^2} = -0.2834 < 0.1 \text{ or}$$

$$\frac{\left(R^2 - R'^2_o\right)}{R^2} = -0.2969 < 0.1$$

$$k = 0.9885 \text{ and } k' = 1.0093$$

The model was validated further by applying the *Y*-randomization of response test (in this work, the θ(LCST) values). It consists of repeating the calculation procedure several times after shuffling the *Y* vector randomly. If all models obtained by the *Y*-randomization test have relatively high values for both R^2 and R^2_{CV} statistics, this is due to a chance correlation and implies that the current modeling method cannot lead to an acceptable model using the available data set. This was not the case for the data set and methodology used in this work. Several random shuffles of the *Y* vector were performed and the results are shown in Table 6. The low R^2 and R^2_{CV} values show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

It needs to be emphasized that, no matter how robust and accurate a QSPR model proves to be, it cannot be expected to predict the modeled property reliably for the entire universe of chemicals. Therefore, for the QSPR model, the domain of applicability must be defined and predictions for only those chemicals that fall in this domain can be considered as reliable. The method was applied to the compounds that constitute the test set. The leverages for all 57 test systems were computed (Table 7). Two systems (G65 and J51) were found to fall slightly outside the domain of the model (warning leverage limit 0.2679).

Conclusions

In this work, we have presented a novel MLR model to predict θ(LCST) using nine molecular descriptors. For the development and the validation of the model, 169 polymer–solvent systems were used. The methodologies used in this work illustrated the accuracy of the model, not only by calculating its fitness on sets of training data but also by testing the predicting abilities of the model. In terms of various validation techniques and statistical indicators, the MLR model produced proved to have significant predictive potential. Using the proposed model, experimental time and effort can be reduced significantly as reliable estimates of θ(LCST) for polymer solutions can be obtained before they are actually synthesized in the laboratory.

Acknowledgements G.M. wishes to thank the Greek State Scholarship Foundation for a doctoral assistantship. A.A. wishes to thank Cyprus Research Promotion Foundation (grant no. PENEK/ENISX/0603/05) and A.G. Leventis Foundation for their financial support.

References

1. (a) Charlet G, Delmas G (1981) Polymer 22:1181–1189; (b) Charlet G, Ducasse R, Delmas G (1981) Polymer 22:1190–1198
2. Christensen SP, Donate FA, Frank TC, LaTulip RJ, Wilson LC (2005) J Chem Eng Data 50:869–877
3. Kavanagh CA, Rochev YA, Gallagher WM, Dawson KA, Keenan AK (2004) Pharmacol Ther 102:1–15
4. Kopecek J (2003) Eur J Pharm Sci 20:1–16
5. Chang BH, Bae CY (1998) Polymer 39:6449–6454
6. Pappa GD, Voutsas EC, Tassios DP (2001) Ind Eng Chem Res 40:4654–4663
7. Bogdanic G, Vidal J (2000) Fluid Phase Equilib 173:241–252
8. Wang F, Saeki S, Yamaguchi T (1999) Polymer 40:2779–2785
9. Vetere A (1998) Ind Eng Chem Res 37:4463–4469
10. Imre AR, Bae YC, Chang BH, Kraska Th (2004) Ind Eng Chem Res 43:237–242
11. Liu H, Zhong C (2005) Eur Polym J 41:139–147
12. Liu H, Zhong C (2005) Ind Eng Chem Res 44:634–638
13. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Supuran CT (2006) Bioorg Med Chem 14:1108–1114
14. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2005) Mol Divers (In press) DOI: 10.1007/s11030-005-9012-2
15. Afantitis A, Melagraki G, Makridima K, Alexandridis A, Sarimveis H, Igglessi-Markopoulou O (2005) J Mol Struct Theochem 716:193–198
16. Melagraki G, Afantitis A, Makridima K, Sarimveis H, Igglessi-Markopoulou O (2005) J Mol Model 12:297–305
17. Al-Fahemi JH, Cooper DL, Allan NL (2005) J Mol Struct Theochem 727:57–61
18. Villanueva-Garcıa M, Gutierrez-Parra RN, Martınez-Richa A, Robles J (2005) J Mol Struct Theochem 727:63–69
19. CambridgeSoft Corporation (http://www.cambridgesoft.com)
20. Efron B (1983) J Am Stat Assoc 78:316–331
21. Efroymson MA (1960) Multiple regression analysis. In: Ralston A, Wilf HS (eds) Mathematical methods for digital computers. Wiley, New York, pp 191–203
22. Osten DW (1998) J Chemom 2:39–48
23. Wold S, Eriksson L (1995) Statistical validation of QSAR results. In: van de Waterbeemd H (ed) Chemometrics methods in molecular design. VCH, Weinheim, pp 309–318
24. Tropsha A, Gramatica P, Gombar VK (2003) Quant Struct-Act Relatsh 22:1–9
25. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269–276
26. Shen M, Beguin C, Golbraikh A, Stables J, Kohn H, Tropsha A (2004) J Med Chem 47:2356–2364
27. Atkinson A (1985) Plots, transformations and regression. Clarendon, Oxford, p 282