J Mol Model (2007) 13:573–578
DOI 10.1007/s00894-007-0180-0

ORIGINAL PAPER

A quantitative structure-property relationship study for refractive indices of conjugated polymers

Jiangang Gao · Jie Xu · Biao Chen · Qijin Zhang

Received: 5 September 2006 / Accepted: 25 January 2007 / Published online: 6 March 2007

Abstract The quantitative structure-property relationship (QSPR) study was performed between descriptors representing the molecular structures and refractive indices for a set of 35 π-conjugated polymers. A seven-descriptor correlation was developed for the prediction of refractive indices with R=0.936 and s=0.028 by stepwise multilinear regression analysis. The average relative error for the calculation of refractive indices was 1.022%. The stability of the proposed model was validated using leave-one-out cross-validation and randomization experiments. Since the model requires only molecular descriptors derived solely from the repeating unit structures of conjugated polymers, it has better predictive capability comparing with the existing group-contribution methods.

Keywords Conjugated polymers · Molecular descriptors · QSPR · Refractive index

J. Gao · J. Xu (✉) · B. Chen · Q. Zhang
Department of Polymer Science and Engineering,
University of Science and Technology of China,
Hefei, Anhui 230026, People's Republic of China
e-mail: xujie0@ustc.edu

Q. Zhang
e-mail: zqjm@ustc.edu.cn

Present address:
J. Xu
Hubei New Texile Material and Its Application Key Lab,
Wuhan University of Science and Engineering,
Wuhan, Hubei 430073, People's Republic of China

Introduction

The optical and nonlinear optical properties of polymers, particularly π-conjugated polymers, are of growing interest in view of their expected applications in photonics, integrated optics, optical communication, and optoelectronics [1–7]. The refractive index (n) is a fundamental optical property of polymers that is directly related to other optical, electrical and magnetic properties [8]. The prediction of this property is valuable due to its application in the design of new optical polymeric materials. For example, the total reflection of light at the boundary between two media with different optical properties is of great importance in the manufacture of waveguides, optical films and optical fibers. Furthermore, the refractive index is also of interest to those studying the physical, chemical, and molecular properties of polymers by optical techniques, e.g., light scattering for the determination of molecular weight, size, and shape [9].

Many semi-empirical group-contribution methods derived from the refractive indices of liquid organic compounds as well as organic polymers have been well-established to give reliable predictions of the refractive indices of nonconjugated polymers [9]. These group-contribution calculations are based on the molar refraction as the additive function and different models of the refractive index such as those due to Lorentz-Lorenz, Gladstone-Dale, Vogel, and Looyenga respectively. The molar refraction values corresponding to these group-contribution models have been collected extensively by van Krevelen [9]. Yang and Jenekhe [10] developed new Lorentz and Lorentz molar refraction (R_{LL}) group contributions for 24 functional groups commonly found in conjugated polymers. They successfully used these new R_{LL} data to calculate the refractive indices at 2500 nm of 33 conjugated polymers (with an average error of 0.9%). Group-contribution methods can sometimes give prediction with reasonable accuracy, but a serious limitation of group-contribution methods is that these methods are only applicable for the polymer containing structural groups previously investigated.

Theoretical QSPR approach is based on the molecular descriptors [11, 12] of polymers instead of group contributions of each component structural group, which has avoided the limitation of group-contribution methods as mentioned above. Bicerano [13] has developed a model using connectivity indices and obtained an accurate prediction of the refractive indices at 589 *nm* (n_D) of nonconjugated polymers. Katritzky et al. [14] used the comprehensive descriptors for structural and statistical analysis (CODESSA) program to obtain a correlation of R=0.970 for a set of the n_D values of 95 nonconjugated polymers with 5 descriptors involved: four quantum-chemical ones and the relative number of F atoms. García-Domenech et al. [15] correlated the n_D values of nonconjugated polymers with their topological indices and got a 10-descriptor correlation with R^2 of 0.962. In our previous work, [16] a four-descriptor QSPR model has been built, based on the structural analysis of polymers, to predict the n_D values of nonconjugated polymers.

It is seen that all the above studies are regarding QSPRs for correlating the refractive index of nonconjugated polymers, where monomer structure or repeating unit end-capped by hydrogen were used as representative structures to calculate the descriptors. This implies that existing QSPR models which completely neglect the cooperative phenomenon of π-electron delocalization between repeating units cannot be accurate for predicting the refractive indices of conjugated polymers. For example, the refractive indices of conjugated polymers predicted by our previous model [16] can have average deviations from experimental values as much as 11.5%. The other models were equally as bad or worse in predictions. The source of these discrepancies is believed to be large optical dispersion and π-electron delocalization effects in conjugated polymers which are not taken into account in the currently available QSPR models.

The aim of the present work is (1) to develop a QSPR model, which is expected to predict the refractive indices of conjugated polymers with very different chemical and structural characteristics, (2) to help to understand the physical mechanisms determining the n of conjugated polymers.

Materials and methods

The molecular structures of conjugated polymers (Fig. 1) and the corresponding experimental refractive index data at 2500 *nm* (Table 1) were taken from [10, 17]. Unlike nonconjugated polymers which are transparent with negligible absorption at 589 *nm* (sodium D line), most conjugated polymers are highly absorbing at 589 *nm*, and the 2500-*nm* data can be regarded as nonresonant values of refractive indices. Thus in the present work, the n_{2500} values were chosen to develop a model as [17]. A total of 35 conjugated polymers were selected as the data set, including aromatic polyimines, polyquinolines, polyanthrazolines, and polybenzobisazoles.

It is impossible to calculate descriptors directly for entire molecule because all the polymers possess high molecular weights. There are two approaches adopted to derive descriptors for polymers in present QSPR studies: (1) using the repeating unit or monomer structure as representative of the corresponding polymer [14–16], and (2) introducing a normalization by extrapolation method [18]. In this work, the repeating unit structure was used to derive descriptors due to the relative ease in calculating.

The software Dragon [19] was used to calculate 305 descriptors, including constitutional descriptors, topological descriptors, 2D autocorrelations, edge adjacency indices, topological charge indices. Most of these descriptors are reviewed in the recent textbook by Todeschini and Consonni [20].

Objective feature selection was done to remove those descriptors that provide minimal or redundant information. Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two indices with a correlation greater than 0.90 was removed to reduce redundant information), thus 61 descriptors for each structure underwent subsequent variable selection for the modeling.

To develop QSPR models, stepwise multilinear regression analysis (MLRA) [21] with leave-one-out (LOO) cross-validation was applied to the data set. Stepwise MLRA produces a multiple-term linear equation; however, not all independent descriptors are used. Step-by-step descriptors are added to the equation, and a new regression is performed. If the new descriptor contributes significantly to the regression equation, the descriptor is retained; otherwise, the descriptor is discarded, hence preventing overfitting. F-to-enter and F-to-remove were 4 and 3, respectively. The following statistical characteristics of the models were used: correlation coefficient R, adjusted R (R_{adj}), the F ratio values, the standard error of estimates (s) and p-values all corresponding to a 95 percent confidence level.

Randomization experiments were also performed to prove the possible existence of fortuitous correlations. To do this, the n_{2500} values were randomly permuted and used in the experiment. Models were then investigated with all

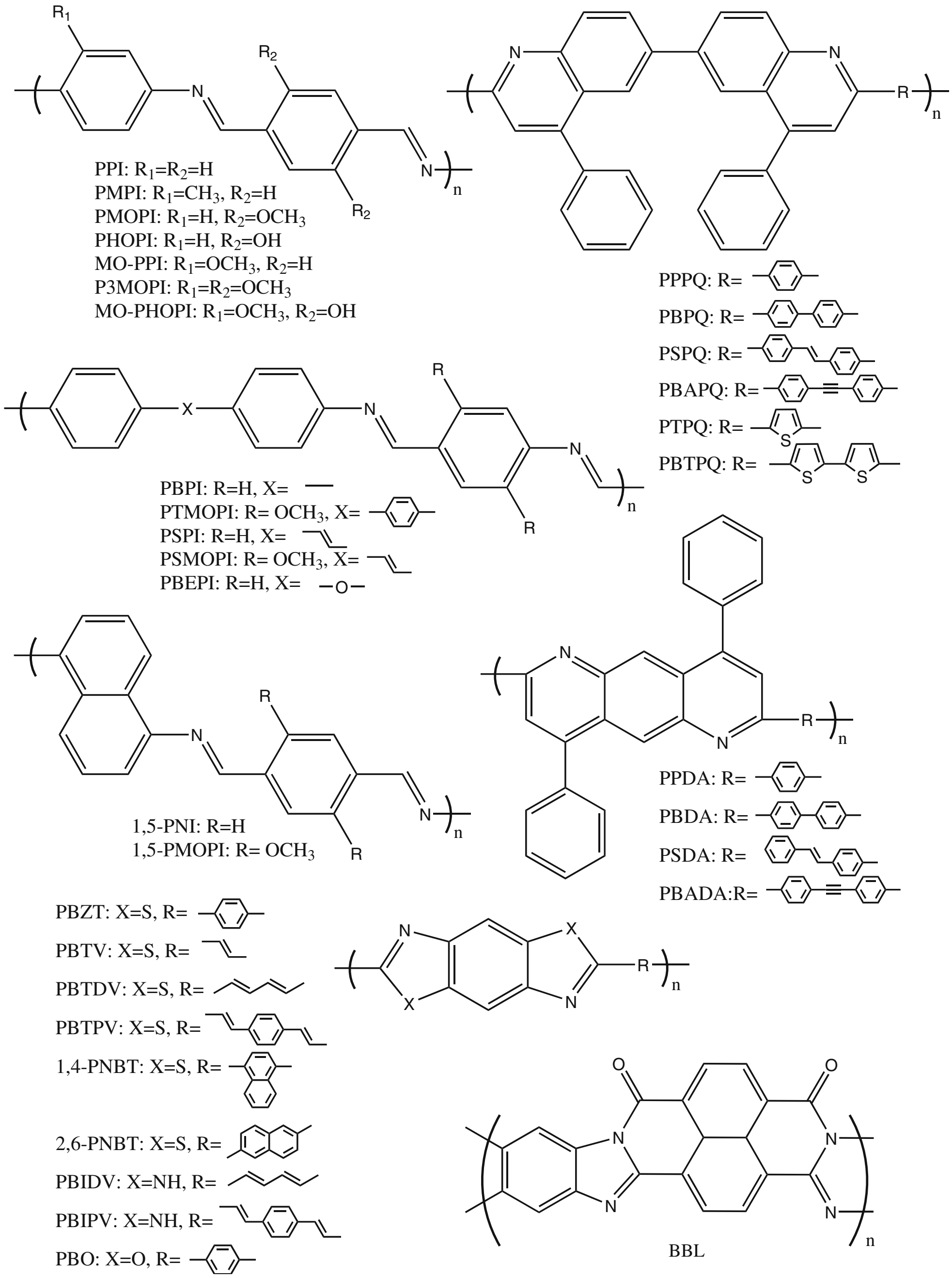

Fig. 1 The structures of conjugated polymers included in the data set

members in the descriptor pool to find the most predictive models. The *s* and correlation coefficients found using random n_{2500} values should be very poor if the original model did accurately represent the relationship between chemical structure and n_{2500}.

Results and discussion

Stepwise MLRA with LOO cross-validation was used to select the descriptors for the best model and the number of descriptors in the final QSPR model was determined on the basis of the data set size and on the basis of the correlation coefficient *R*, the adjusted *R*, the significance test *F* and the standard error *s*. It is clear that univariant correlations between n_{2500} and the different descriptors have a small value for the correlation coefficient. This indicates that n_{2500} is not linearly correlated with any of the molecular descriptors.

The best correlation model obtained for the entire data set of 35 conjugated polymers contains seven descriptors. From a statistical viewpoint the ratio of the number of samples (N) to the number of descriptors (M) should not be too low. Usually, it is recommended that N/M≥5. In the

Table 1 Predicted results of n_{2500} for 35 conjugated polymers

Polymer	n_{2500}					Δn^{a}	
	Exp.	Calc.	CV.	Jenekhe et al.	Calc.	CV	Jehekhe et al.
1,5-PMONI	1.75	1.72	1.71	1.728	−0.03	−0.04	−0.022
1,5-PNI	1.79	1.78	1.78	1.817	−0.01	−0.01	0.027
1,4-PNBT	1.75	1.80	1.80	1.752	0.05	0.05	0.002
MO-PHOPI	1.74	1.71	1.70	1.726	−0.03	−0.04	−0.014
MO-PPI	1.73	1.74	1.74	1.719	0.01	0.01	−0.011
P3MOPI	1.62	1.62	1.62	1.638	0.00	0.00	0.018
PBADA	1.76	1.78	1.78	1.806	0.02	0.02	0.046
PBAPQ	1.87	1.83	1.82	1.824	−0.04	−0.05	−0.046
PBDA	1.78	1.75	1.74	1.742	−0.03	−0.04	−0.038
PBEPI	1.76	1.78	1.80	1.757	0.02	0.04	−0.003
PBIDV	1.82	1.81	1.80	1.789	−0.01	−0.02	−0.031
PBIPV	1.75	1.76	1.76	1.781	0.01	0.01	0.031
PBO	1.70	1.68	1.67	1.692	−0.02	−0.03	−0.008
PBPI	1.79	1.78	1.78	1.791	−0.01	−0.01	0.001
PBPQ	1.79	1.79	1.80	1.765	0.00	0.01	−0.025
PBTDV	1.85	1.87	1.87	1.799	0.02	0.02	−0.051
PBTPQ	1.90	1.87	1.86	1.904	−0.03	−0.04	0.004
PBTPV	1.92	1.90	1.90	1.897	−0.02	−0.02	−0.023
PBTV	1.74	1.74	1.74	1.708	0.00	0.00	−0.032
PBZT	1.69	1.69	1.69	1.684	0.00	0.00	−0.006
PHOPI	1.77	1.76	1.76	1.781	−0.01	−0.01	0.011
PMOPI	1.66	1.66	1.66	1.666	0.00	0.00	0.006
PMPI	1.66	1.71	1.71	1.668	0.05	0.05	0.008
PPDA	1.68	1.67	1.65	1.660	−0.01	−0.03	−0.02
PPI	1.79	1.78	1.78	1.772	−0.01	−0.01	−0.018
PPPQ	1.69	1.74	1.74	1.696	0.05	0.05	0.006
PSDA	1.81	1.78	1.78	1.811	−0.03	−0.03	0.001
PSMOPI	1.64	1.67	1.67	1.623	0.03	0.03	−0.017
PSPI	1.69	1.68	1.69	1.686	−0.01	0.00	−0.004
PSPQ	1.79	1.83	1.83	1.832	0.04	0.04	0.042
PTMOPI	1.75	1.75	1.75	1.751	0.00	0.00	0.001
PTPQ	1.78	1.79	1.79	1.776	0.01	0.01	−0.004
2,6-PNBT	1.80	1.80	1.80	1.752	0.00	0.00	−0.048
BBL	1.88	1.89	1.89	n.a.	0.01	0.01	n.a.
PPI/PMPI	1.73	1.74	1.74	1.720	0.01	0.01	−0.010
Average					0.018 (1.022%)	0.021 (1.199%)	0.018 (1.052%)

[a] $\Delta n = n_{2500}(\text{calc.}) - n_{2500}(\text{exp.})$

situation of this work, with 35 samples, seven descriptors were selected. The final correlation equation is the following:

$$\begin{aligned} n_{2500} &= 2.324 - 0.515J - 0.137EEig12x \\ &\quad + 0.177Lop + 0.102EEig08d - \\ &\quad 0.597MATS3v + 0.05927nR10 - 10.754JGI9 \\ R &= 0.936,\ R_{adj} = 0.919,\ s = 0.028,\ F = 27.196,\ N = 35. \end{aligned} \tag{1}$$

Here, J is the Balaban distance connectivity index; [22–24] $EEig12x$ is the eigenvalue 12 from edge adjacency matrix weighted by edge degrees; [25–29] Lop is the Lopping centric index; [30] $EEig08d$ is the eigenvalue 08 from edge adjacency matrix weighted by dipole moments; [25–29] $MATS3v$ is the Moran autocorrelation - lag 3 / weighted by atomic van der Waals volumes; [20] $nR10$ is the number of 10-membered rings; $JGI9$ is the mean topological charge index of order 9, [31] respectively.

The calculated results from Eq. (1) and the LOO cross-validation are shown in Table 1 and Fig. 2. The s and R during stepwise MLRA are given in Fig. 3, which indicate that equations with fewer descriptors showed poorer modeling results. The deviations between experimental and calculated refractive indices are also plotted in Fig. 4. The absolute average errors (average relative errors) by Eq. (1) and for the cross-validation are 0.018 (1.022%) and

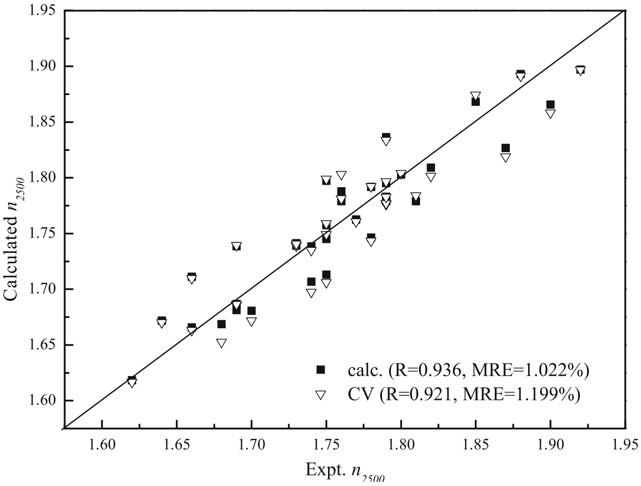

Fig. 2 Plot of the calculated, the cross-validated vs. experimental n_{2500} of conjugated polymers

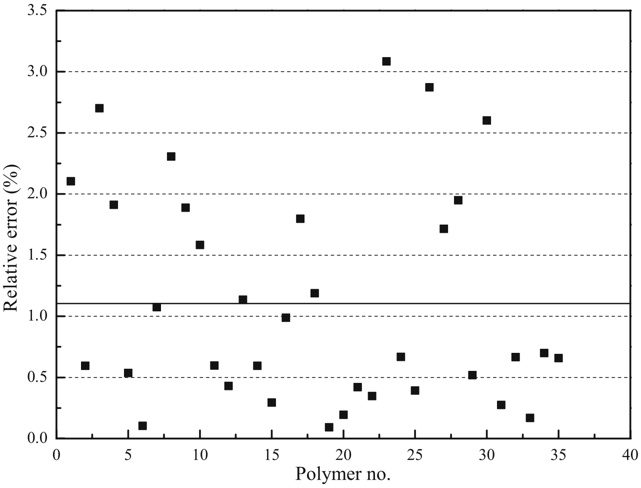

Fig. 4 The deviations between calculated and experimental n_{2500} of conjugated polymers

0.021 (1.199%), respectively. The discrepancies between Δn obtained by Eq. (1) and for cross-validation are small for most of the studied polymers, which reveals the quality of the present model for the prediction of refractive index of conjugated polymers. The characteristics of the best seven descriptors in Eq. (1) are shown in Table 2. The *t*-values indicate that the all the descriptors are highly significant. Variance inflation factor (VIF) was used to test for multi-collinearity of these 7 descriptors. If there is VIF≥10 in a model, the descriptor is strongly correlated with the others and it is not significant to explain that model, which is not reliable. All the values of VIF are less than 10 (see Table 2). Therefore the 7 descriptors are weakly correlated with each other and the QSPR model can be regarded as an optimal regression equation.

The calculated n_{2500} using the R_{LL} data proposed by Yang and Jehekhe [10] are also reported in Table 1, where the absolute relative error is 1.052%. Although it gives comparable accuracy with the present model, calculations are not possible for one polymer listed in Table 1 due to the lacking of the required R_{LL} data. This illustrates the common drawback of the group-contribution methods, as well as the necessity of developing QSPR models based only on structural descriptors, such as Eq. (1) developed in this work.

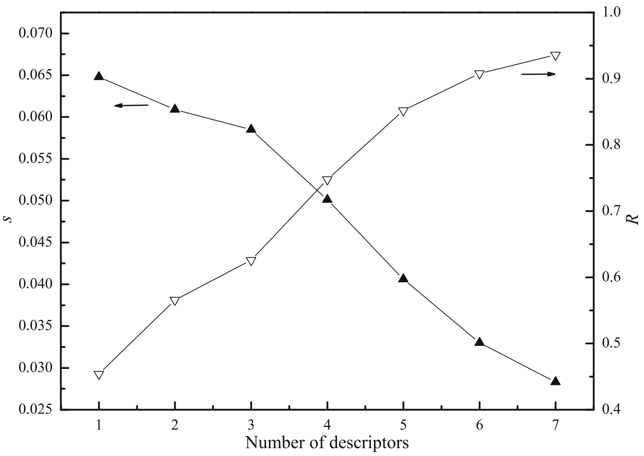

Fig. 3 *s* and *R* vs. number of descriptors in the best MLRA equation

By interpreting the descriptors in the regression model, it is possible to gain some insights into factors that are likely to relate to the n_{2500} values of conjugated polymers. The Balaban index, *J*, is the average-distance sum connectivity, [22–24] taking into consideration both heteroatoms and multiple bonds. For a given connected molecular graph G,

$$J = \frac{M}{\mu + 1} \sum \left(D_i D_j\right)^{-0.5} \tag{2}$$

where M is the number of edges in G, and μ denotes the ring number of G. In a polycyclic graph, μ is the minimum number of edges that must be removed before G becomes acyclic. $D_i = \sum_{j=1} D_{ij}$ and D_{ij} is a distance matrix of the shortest paths between any two vertices for *N* vertices. D_{ij}= l_{ij} if $i \neq j$, otherwise equal to zero. l_{ij} is the shortest distance between vertices *i* and *j*. The presence of *J* and *nR*10 in Eq. (1) reflect the influence of the number of rings on the value of n_{2500}. *J* decreases with increased number of rings.

Table 2 Descriptors involved in the best seven-parameter correlation derived for n_{2500} of conjugated polymers

Descriptor	Standard error	*t*-Test	*t*-Probability	VIF
Constant	0.096	24.211	0.000 000	
J	0.066	−7.819	0.000 000	4.405
EEig12x	0.014	−10.069	0.000 000	8.536
Lop	0.021	8.494	0.000 000	4.204
EEig08d	0.019	5.410	0.000 010	5.575
MATS3v	0.155	−3.860	0.000 640	4.881
nR10	0.008	7.059	0.000 000	3.301
JGI9	3.233	−3.326	0.002 547	2.806

The negative sign of J and the positive sign of $nR10$ indicate that the more rings in the repeating unit of conjugated polymers, the higher refractive index is, which is similar to that for nonconjugated polymers [16].

The centric index Lop [30] is calculated by Eq. (3), where n_g is the number of the terminal vertices removed at the g th step, A is the number of graph vertices and K is the number of steps to remove all graph vertices:

$$Lop = -\sum_{g=1}^{K} \frac{n_g}{A} \cdot \log_2 \frac{n_g}{A}. \quad (3)$$

The pruning algorithm, used for the calculation of Lop, is conceived in such a way to take into account, iteratively, all the atoms with vertex degree equal to 1. The atoms not reaching a vertex degree =1 during the pruning procedure give no contribution to Lop. The index decreases with increased chain branching. The coefficient for Lop in Eq. (1) is positive, meaning that as branching increases, n_{2500} decreases.

The positive sign of $EEig08d$ in Eq. (1) indicates that the n_{2500} values of conjugated polymers would increase with decreased dipole moments. This can be understood as that π-electrons in the conjugated polymers with smaller dipole moments are weakly attracted and the polarization of these electrons under the electromagnetic fields of light would become easier, resulting in the increase of the n_{2500} values.

The negative sign of $MATS3v$ in Eq. (1) indicates that conjugated polymers containing atoms with larger van der Waals volumes would possess higher refractive indices, because this descriptor increases with increased atomic van der Waals volumes. The importance of the transfers of intramolecular charge on the n_{2500} value is apparent due to the presence of $JGI9$ in Eq. (1).

The same model size and algorithm that produced the best model for the standard experiment were tested with the randomized n_{2500} values. The most predictive model with s of 0.053 (R=0.709, R_{adj}=0.671) was obtained. The s and R-value indicate that a poor correlation was found between structure and n_{2500}, which proves the validity of the real model.

Conclusions

In this paper, a successful correlation model for the prediction of n_{2500} of a variety of conjugated polymers was reported. The R of the correlation is 0.936 and the average relative error for the prediction is 1.022%. Considering the uncertainty that accompanies the experimental determination of n_{2500} for each case, these values are acceptable. Since molecular descriptors can be calculated as long as the repeating unit structure of the polymer in question is known, the proposed correlation is predictive. Therefore, this QSPR model should be useful in design and development of new conjugated polymers.

Acknowledgements This work was supported by the National Natural Science Foundation of China (50025309). Special thanks were given to Prof. Todeschini and other members in Milano Chemometrics and QSAR Group for providing Dragon package for use in this research.

References

1. Olshavsky M, Allcock HR (1997) Macromolecules 30:4179–4183
2. Zhan X, Liu Y, Zhu D, Huang W, Gong Q (2002) J Phys Chem B 108:1884–1888
3. Kuang L, Chen Q, Sargent EH, Wang ZY (2003) J Am Chem Soc 125:13648–13649
4. You W, Cao S, Hou Z, Yu L (2003) Macromolecules 36:7014–7019
5. Thomas SW, Swager TM (2005) Macromolecules 38:2716–2721
6. Yang L, Feng JK, Ren AM (2005) J Org Chem 70:5987–5996
7. Liu B, Bazan GC (2006) J Am Chem Soc 128:1188–1196
8. Knoll W (1998) Annu Rev Phys Chem 49:569–639
9. van Krevelen DW (1976) Properties of polymers: correlation with chemical structure. Elsevier, Amsterdam
10. Yang CJ, Jenekhe SA (1995) Chem Mater 7:1276–1285
11. Devillers J, Balaban AT (1999) Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands
12. Karelson M (2000) Molecular Descriptors in QSAR/QSPR. Wiley-Interscience, New York
13. Bicerano J (1996) Prediction of polymer properties. Marcel Dekker Inc, New York
14. Katritzky A, Sild S, Karelson M (1998) J Chem Inf Comput Sci 38:1171–1176
15. García-Domenech R, Julián-Ortiz JV (2002) J Phys Chem B 106:1501–1507
16. Xu J, Chen B, Zhang Q, Guo B (2004) Polymer 45:8651–8659
17. Yang C-J, Jenekhe SA (1994) Chem Mater 6:196–203
18. Balaban TS, Balaban AT, Bonchev D (2001) J Mol Struct (THEOCHEM) 535:81–92
19. http://www.talete.mi.it/dragon_exp.htm.
20. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors. Wiley-VCH, Weinheim
21. Jurs PC (1996) Computer Software Applications in Chemistry. Wiley, New York
22. Balaban AT (1982) Chem Phys Lett 89:399–404
23. Balaban AT (1983) Pure Appl Chem 55:199–206
24. Balaban AT (1986) Math Chem (MATCH) 21:115–122
25. Estrada E (1995) J Chem Inf Comput Sci 35:31–33
26. Estrada E (1995) J Chem Inf Comput Sci 35:701–707
27. Estrada E (1996) J Chem Inf Comput Sci 36:844–849
28. Estrada E, Ramirez A (1996) J Chem Inf Comput Sci 36:837–843
29. Estrada E (1997) J Chem Inf Comput Sci 37:320–328
30. Balaban AT (1979) Theor Chim Acta 53:355–375
31. Gálvez J, García-Domenech R, Salabert MT, Soler R (1994) J Chem Inf Comput Sci 34:520–525