SELECTION OF MOLECULAR DESCRIPTOR SUBSETS FOR PROPERTY PREDICTION

Inga Paster, Chem. Eng. Dept., Ben-Gurion University, Beer-Sheva, Israel

Neima Brauner, School of Engineering, Tel-Aviv University, Tel-Aviv, Israel

Mordechai Shacham, Chem. Eng. Dept., Ben-Gurion University, Beer-Sheva, Israel

Introduction

Pure-compound property data are at present available only for a small fraction of the compounds, pertaining to such diverse areas as chemistry and chemical engineering, environmental engineering and environmental impact assessment, hazard and operability analysis. Therefore, methods for reliable prediction of property data are needed. Current methods used to predict physical and thermodynamic properties can be classified into "group contribution" methods (see, for example, Marrero and Gani, 2001), methods based on the "corresponding-states principle", (Poling et al., 2001), "asymptotic behavior" correlations (Marano and Holder, 1997) and Quantitative Structure Property Relationships (QSPRs, Dearden et al., 2003).

Recently we have developed the Targeted QSPR (TQSPR) method (Shacham et al., 2007, Brauner, et al., 2008) which enables predicting properties within experimental error level. Unlike in the traditional QSPR methods, the TQSPR method is targeted to a particular compound, or a group of compounds, and relies on the identification of a relatively small number of structurally similar compounds. Hence, it can provide accurate predictions and estimates of the prediction error, while avoiding the need to model the highly nonlinear relationships between molecular descriptors and properties that may require large amount of experimental data.

In recent years computer programs that can calculate several thousands of descriptors have emerged. This raised concerns regarding the accuracy and consistency of the molecular descriptors used and the probability of obtaining "chance" correlations (Topliss and Costello, 1972), while applying stepwise regression procedures to large descriptor databases in order to obtain a QSPR or TQSPR for representing a particular property.

It is practically impossible to check the accuracy and consistency of the individual-descriptor values because of the large number of descriptors and compounds involved. However, some of the inconsistencies can be detected when examining the variations of the descriptor values in homologous series. It is well known that most physical properties change in a consistent manner in homologous series when plotted versus the number of carbon atoms (nC, Marano and Holder 1997A and 1997B) and many molecular descriptors follow the same trend as properties (Brauner et al., 2008). In this manuscript we investigate the behavior of molecular descriptors associated with some homologous series in order to identify clear trends in the behavior of the descriptors and to find potential causes of inconsistencies among the descriptors. Descriptor subsets which can be helpful in training set selection and/or prediction of particular properties while minimizing the probability of descriptor inconsistency are identified. These subsets of descriptors are used for predicting seven properties for five homologous series.

The Targeted QSPR Method

The targeted QSPR (TQSPR) technique is described in detail by Brauner et al. (2006), Shacham et al. (2007) and Kahrs et al. (2008). The basic principles of the method are briefly reviewed here .

Let us assume that the vector of properties of the target compound $\mathbf{y}_t$ (the dependent variable) is potentially related to a set of m vectors of properties of predictive compounds (independent variables) $\mathbf{x}_1$, $\mathbf{x}_2$,…$\mathbf{x}m$. The following partition of the $\mathbf{y}_t$ and $\mathbf{x}$ vectors to sub-vectors is used:

$$\mathbf{y}_t = \begin{Bmatrix} \mathbf{y}_{ct} \\ \mathbf{y}_{pt} \end{Bmatrix} \quad ; \quad \mathbf{x}_j = \begin{Bmatrix} \mathbf{x}_{ci} \\ \mathbf{x}_{pi} \end{Bmatrix} \tag{1}$$

where $\mathbf{y}_{ct}$ is an N vector of known properties, $\mathbf{y}_{pt}$ is a K vector of unknown properties. Both the N vector $\mathbf{x}_{ci}$ and the K vector $\mathbf{x}_{pi}$ contain known properties. Typically, the sub-vectors $\mathbf{y}_{ct}$ and $\mathbf{x}_{ci}$ contain properties, which are directly related to the molecular structure and can be calculated with high accuracy (molecular descriptors), while the sub-vectors ypt and xpi contain measured properties with various levels of experimental error.

The practical application of the TQSPR method requires preparation of a bank of potential predictive compounds as a database. The same set of molecular descriptors must be defined for all compounds included in the database, while the span of the molecular descriptors should reflect the difference between the compounds in the database. Having the corresponding molecular descriptors for a target compound, $\mathbf{y}_c$, defined as well, a stepwise regression procedure is applied in order to identify the most appropriate predictive compounds to be included in the training set associated with the particular target compound. The similarity between potential predictive compounds and the target compound is measured by the partial correlation coefficient, r_{ti}, between the vector of the molecular descriptors of the target compound, $\mathbf{y}_{ct}$, and that of a potential predictive compound $\mathbf{x}_{ci}$. The partial correlation coefficient is defined as $r_{ti} = \bar{\mathbf{y}}_{ct}\bar{\mathbf{x}}_{ci}^T$, where $\bar{\mathbf{y}}_{ct}$ and $\bar{\mathbf{x}}_{ci}$ are row vectors, centered (by subtracting the mean) and normalized to unit length (by dividing by the Euclidean norm of the vector). Absolute r_{ti},values close to one ($|r_{ti}|$≈1) indicate high correlation between vectors $\mathbf{y}_{ct}$ and $\mathbf{x}_{ci}$, and thus, high level of similarity between the molecular structures of the target compound and the predictive compound i. The training set is established by selecting the np compounds with highest | r_{ti},| value for which experimental property values $\mathbf{y}_{pi}$ are available. The remaining compounds in the similarity group are used for validation .

For development of a TQSPR for a particular property of the target compound, a linear structure- property relation is assumed of the form:

$$\mathbf{y}_p = \beta_0 + \beta_1\boldsymbol{\zeta}_1 + \beta_2\boldsymbol{\zeta}_2 \dots \beta_m\boldsymbol{\zeta}_m + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{y}_p$ is a n_p-dimensional vector of the respective property values, $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ … $\boldsymbol{\zeta}_m$ are n_p -dimensional vectors of predictive molecular descriptors, $\beta_0, \beta_1, \beta_2 \dots \beta_m$ are the corresponding model parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a n_p -dimensional vector of random measurement errors .

A stepwise regression program is used to determine which molecular descriptors should be included in the TQSPR to best represent the measured property data of the training set and to calculate the TQSPR parameter values. In each step one molecular descriptor that reduces the prediction error most strongly is added into the model. The descriptors are selected to the model in a stepwise manner according to the value of the partial correlation coefficient, $|\rho_{yj}|$ between the vector of the property values $\mathbf{y}_p$, and that of a potential predictive descriptor ζ_j. The partial correlation coefficient is defined as $\rho_{yj} = \bar{\mathbf{y}}_p\bar{\zeta}_j^T$, where $\bar{\mathbf{y}}$ and $\bar{\zeta}_j$ are row vectors, centered (by subtracting the mean) and normalized to a unit length. Values close to one indicate high correlation between the molecular descriptor and the property. The TQSPR so-obtained can be subsequently employed for calculating estimated property values for the target compound by

$$\tilde{y}_{pt} = \beta_0 + \beta_1\zeta_{t1} + \beta_2\zeta_{t2} \ldots \beta_m\zeta_{tm} \qquad (3)$$

where $\tilde{y}_{pt}$ is the estimated unknown property value of the respective compound and ζ_{t1}, ζ_{t2} … ζ_{tm} are its corresponding molecular descriptor values.

The selection of a suitable set of predictive molecular descriptors for Eq (2) is a challenging problem, since the number of candidates is in the order of θ(10^3), which prohibits the determination of the best of all possible sets of predictive molecular descriptors by a full search procedure. The stepwise regression program SROV (Shacham and Brauner, 2003) is used, which selects in each step one molecular descriptor that reduces the prediction error most strongly. Addition of descriptors to the TQSPR stops when all the signal-to-noise ratios in the correlation between the residuals of the property and the *j*-th candidate descriptor get below a threshold value.

Methodology

The Dragon program (version 5.4, DRAGON is copyrighted by TALETE srl, http://www.talete.mi.it) was used to calculate 1664 descriptors for the compounds in the database from minimized energy molecular models. The molecular geometries were optimized using the CNDO (Complete Neglect of Differential Overlap) semi-empirical method implemented in the HyperChem package (Version 7.01, Hyperchem is copyrighted by Hypercube Inc.). The number of descriptors was reduced to 1280 by removing the descriptors that had the same value for all compounds. Property data (measured and predicted) were taken DIPPR (Rowley et al., 2006) and NIST (National Institute of Standards, 2005) databases. A modified version of the stepwise regression program (SROV, Shacham and Brauner 2003) was used for the identification of the most appropriate QSPRs.

Analysis of the Molecular Descriptor Database

The Dragon program generates the following types of descriptors: 0-D (e.g., molecular weight), 1-D (e.g., numbers of functional groups, atoms, bonds), 2-D (e.g., topological descriptors), 3-D (e.g., geometrical descriptors) charge and molecular properties (e.g., drug-like index, toxicity) descriptors .

The calculation of molecular descriptors is a complex process which may lead to different types of inconsistencies. A major source of inconsistencies is the optimization of the 3-D molecular geometries. The software used may yield different results depending on the tolerances used and may even converge to a local minimum, and the calculated values of some of the descriptors are highly sensitive to small variations in the (optimized) molecular geometry. Another source of difficulty is when dealing with descriptors which are undefined for some of the compounds. In some cases the descriptor values are categorized as "missing data" (marked by a value of -999 by Dragon) or assigned a value of zero. Obviously, an assigned zero value (e.g., a descriptor value of n-alkane with n1 carbon atoms, which is defined for n>n1) may not have the same role as a zero value of a physical significance (such as the number of oxygen atoms in a hydrocarbon).

To check the accuracy and consistency of the descriptors we plot them versus n_C. This procedure will be demonstrated for the 1-alkene homologous series. The descriptor *ADDD* (defined by Dragon as "average distance/distance degree" from the 3-D, geometrical descriptors category) versus n_C for the 1-alkene series is depicted in Figure 1. The descriptor value increases linearly with n_C, following the relationship *ADDD* = $3.364n_C - 3.388$ with R^2 = 0.995. A similar linear (or nearly linear) change can be observed, for example, for the liquid molar volume of *n*-paraffins and *n*-olefins (Marano and Holder, 1997A).
The normalized values of the descriptors *AGDD* (3-D, average geometric distance degree), *ASP* (3-D, asphericity) and H4m (3-D H-autocorrelation of lag 4/ weighted by atomic masses) for the 1-alkene series are plotted versus n_C in Figure 2. In all three cases the descriptor value increases nonlinearly with n_C. For *AGDD* the change is monotonic with an increasing slope. Similar behavior can be observed for critical volume (V_c) of *n*-paraffins and *n*-olefins (Marano and Holder, 1997A). For the descriptor *ASP* the change is monotonic with a decreasing slope, and the descriptor approaches a constant value for high n_C. Similar behavior can be observed for normal boiling (T_b) and critical temperatures (T_c) of *n*-paraffins and *n*-olefins (Marano and Holder, 1997A). The variation of the descriptor *H4m* with n_C is similar to that of ASP except that there is an abrupt change in the slope (and the level of the nonlinearity) at a particular value of n_C. On the other hand, the plot of the descriptor *Gm* (3-D descriptor, G total symmetry index/weighted by atomic masses) versus n_C (Figure 3) does not reveal any consistent variation, and such a behavior will be designated as "random".

A similar descriptor analysis and characterization was carried out for additional descriptor types and also members of the *n*-alkane, 1-alkene, cycloalkane, alkyl-cyclohexane, alkyl-cycloheptane *n*-alkylbenzene, *n*-alcohol, and *n*-alyphatic acid homologous series for 1280 descriptors of the database. The results of this study are summarized in Table 1.

Most of the descriptors (Categories II, IIIA, IIIB, 46.2 % of the total) increase monotonically with n_C. Additional 21.9 % are defined at zero for lower region of n_C and increase monotonically for the rest. Thus, approximately 68 % of the descriptors are correlated with n_C. Approximately 8.5% of the descriptors (type I) have constant value for a particular series. A large portion of the descriptors (22.9 %, type IV descriptors) appears to vary inconsistently with n_C, in an apparently random manner. There are only 0.4% descriptors with separate curves for odd and even carbon number compounds (type VI) and 0.2 % descriptors with a periodic (sinusoidal) change (type VII).

In the last column of Table 3 the percentages of the 3D descriptors in each category are listed. Amongst type II and IIIA descriptors, which are the important ones for most properties, only 41.7% and 32.1 % respectively, are 3D descriptors. In contrast, all the type VI descriptors, that are used for T_m prediction, are 3D. The large percentage of the 3D descriptors in category IV of descriptors with inconsistent trends raises the suspicion that some of the inconsistency may result from the route adopted for obtaining the 3-D representation of the molecule.

Selecting Members of the Homologous Series to the Training Set

For the derivation of a reliable TQSPR (which can provide a reliable estimate for the target-compound property value), it is essential that the training set contain predictive compounds which are "similar" to the target. A typical example of groups which contain "similar" members, are the homologous series. In such series most properties change with n_C according to specific rules, which can be often approximated by ABC correlations.

Because of the similarity between the members of homologous series it is important to be able to include in the similarity group all the members of the series to which the target compound belong and for whom the pertinent property values are available. In order to achieve that we attempted to identify the subset of descriptors which can be used to select the members of training sets so that the maximal number of homologous series members are included. Several clustering algorithms tested by Kahrs et al (2008) were able to achieve this objective only for particular target compounds, depending on their location in the homologous series.

We have tried to use different subsets of the descriptors to achieve maximal utilization of the available predictive compounds belonging to the target compounds' homologous series. The most successful option turned out to be the use of a two step procedure. In the first step Category I descriptors, which have constant values for homologous series, only used. After removing the 3-D descriptors 61 of this type of descriptors remain in the database. Using Type I descriptors yields $|r_{ti}|$ = 1 values for all the members of the homologous series to which the target compound belongs, thus, this step enables isolation of the homologous series subset from the rest of the database. It however does not arrange the members of homologous series according to their similarity with the target compound. To achieve that, the 2nd step of the similarity group selection is carried out where only Category II descriptors (after excluding the 3-D descriptors) are used to identify the training set from the subset of compounds that were identified in step 1.

In Table 2 the members of the training set for the target compound: *n*-dodecane (n_C = 12) that were identified using the two-step procedure are shown. Observe that the $|r_{ti}|$ values are such that the predictive compounds are arranged according to their distance (in terms of n_C) from the target compound, while the compound with the higher n_C always precedes the compound with the lower n_C at the same distance.

This type of training set is the most appropriate for predicting gas and liquid properties, however it may not be the best for predicting solid properties, such as the normal melting temperature: T_m, as for some series there are different similarity rules for odd and even n_C compounds. In such cases it may be preferable to remove from the training set compounds whose n_C type (odd or even) does not match that of the target compound.

Comparison of Property Prediction with and without 3-D descriptors

In order to check the assumption that property predictions can be carried out with the TQSPR method in a similar level of accuracy with or without the use of 3-D descriptors, seven properties were predicted for members of five homologous series using the two alternatives. The properties predicted included critical temperature (T_C), critical pressure (P_C), critical volume (V_C), normal melting temperature T_m, normal boiling temperature (T_b), liquid molar volume (V_m) and refractive index (*RI*). The following homologous series were considered: *n*-alkanes, from *n*-butane (n_C = 4) to *n* – hexatriacontene (n_C = 36); 1-alkenes, from 1-butene (n_C = 4) to 1-eicosene (n_C = 20); n-alkylbenzenes, from butylbenzene (n_C = 10) to n-octadecylbenzene (n_C = 24) ; 1-alcohols, from 1-butanol ((n_C = 4) to 1-docosanol (n_C = 22); and aliphatic acids, from butanoic-acid (n_C = 4) to eicosanoic-acid (n_C = 20).

Each compound in these series was targeted individually. The two step approach for selection of the training set, which was described in the previous section, was used for predicting all the properties except T_m. For T_m the first compounds to enter the training set were the ones that matched the n_C type (odd or even) of the target. TQSPRs including one or maximum two descriptors were used, (depending on the signal to noise ratio after introducing the first descriptor into the model). After predicting a property $\tilde{y}_{pt}$ for the target compound its value was compared with the property value recommended by DIPPR ($y_{t,publ}$) and the absolute difference (in %) between the predicted and the recommended values was calculated:

$$\delta = 100\left|\tilde{y}_{pt} - y_{t,publ}\right| / y_{t,publ} \quad (4)$$

In Table 3 the mean, the median and the standard deviation of the δ values are summarized for the five homologous series and the seven properties, for the case where the 3-D descriptors are excluded from the TQSPRs. Observe that T_C, T_b, V_m and *RI* are predicted for essentially all the compounds (except for V_m of n-aliphatic acids) with difference from the DIPPR recommended values δ < 1%. For P_C and V_C, there are some cases where δ > 1%. The larger differences are caused in this case by large uncertainties in the DIPPR recommended property values, which may reach 10% - 25 %, especially for high n_C compounds. For T_m there are also cases where δ > 1%. As the uncertainty in the recommended T_m values is usually < 1% (or even <0.2%), the conclusion from these results is that linear TQSPR with two descriptors can be insufficient to represent T_m within experimental error, for some homologous series.

The same study was carried out without excluding the 3-D descriptors. There were only four cases where the difference between the numbers shown in Table 3 and the ones obtained using the full database of descriptors exceeded 0.5%. With inclusion of the 3-D descriptors, the standard deviation for P_c of 1-alkenes was 2.12 (1.48 in Table 3), the standard deviation of T_m of 1-alkenes was 2.26 (1.21 in Table 3), the mean of V_C for 1-alcohols was 3.02 (3.53 in Table 3), and the standard deviation of V_C for 1-alcohols was 2.69 (6.17 in Table 3).

Thus, it can be concluded that exclusion of 3-D descriptors did not degrade the precision of the TQSPR prediction of properties (except for Tm) in the homologous series tested.

Conclusions

To study the suitability of various descriptors to represent different properties, 1256 Dragon descriptors were divided into seven categories according to the trend of their change as function of n_C in homologous series. For the selection of the TQSPR training set, a two-step procedure was adopted, which includes the use of descriptors from the categories "Constant" and "Linear or nearly linear increase". This procedure enables the selection of training sets that include only members of the target-compound homologous series (if enough such compounds are available in the database).

Seven properties (critical temperature, critical pressure, critical volume, normal melting temperature, normal boiling temperature, liquid molar volume and refractive index) were predicted for five homologous series using 1- D or 2-D descriptor TQSPRs with 3-D descriptors included or excluded. It has been shown that all these properties can be predicted within experimental uncertainty, regardless of the use of 3-D descriptors. However, for predicting the normal melting temperature, T_m, the use of 3-D descriptors may be essential and often more than two descriptors will be needed for representing this property within the experimental uncertainties.

As expected, for predicting T_C, P_C, T_m, T_b and *RI* of homologous series, the most frequently selected dominant (of highest correlation with a particular property) descriptors are of category IIIA (nonlinear monotonic increase or decrease with decreasing slope). Descriptors of Category II (linear or nearly linear increase) had the highest correlation with V_C and V_m.

This study has shown that for prediction of properties in homologous series using the TQSPR method, most properties can be predicted on experimental error level using maximum two (non 3-D) descriptors. The exclusion of the 3-D descriptors increases the reliability of the prediction and extends the possible use of the developed QSPR by the scientific community. Moreover, the use of small number of descriptors reduces considerably the probability of obtaining "Chance correlations".

References

1. Brauner, N; Stateva, R. P.; Cholakov, G. St.; Shacham, M. Structurally "Targeted" Quantitative Structure-Property Relationship Method for Property Prediction. *Ind. Eng. Chem. Res*. 2006, 45, 8430-8437.
2. Brauner, N.; Cholakov, G. St.; Kahrs, O.; Stateva, R. P.; Shacham, M. Linear QSPRs for Predicting Pure Compound Properties in Homologous Series. *AIChE J*. 2008, 54(4), 978-990.
3. Dearden, J. C. "Quantitative Structure–Property Relationships for Prediction of Boiling Point, Vapor Pressure, and Melting Point", Environmental Toxicology and Chemistry, 22(8), 1696–1709 (2003).
4. Kahrs, O.; Brauner, N; Cholakov, G. St.; Stateva, R. P.; Marquardt, W.; Shacham, M. Analysis and Refinement of the Targeted QSPR Method. Computers Chem. Engng. 2008, 32 (7) 1397-1410.
5. Marano, J.J.; Holder, G.D. General Equations for Correlating the Thermo-physical Properties of n-Paraffins, n-Olefins and other Homologous Series. 2. Asymptotic Behavior Correlations for PVT Properties. *Ind. Eng. Chem. Res.* 1997A, 36, 1895.
6. Marano, J.J.; Holder, G.D. General Equations for Correlating the Thermo-physical Properties of n-Paraffins, n-Olefins and other Homologous Series. 3. Asymptotic Behavior Correlations for Thermal and Transport Properties, *Ind. Eng. Chem. Res*. 1997B, 36, 2399.
7. Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibrium.* 2001, 183.
8. NIST Chemistry WebBook, NIST Standard Reference Database Number 69, June 2005, Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology, Gaithersburg MD, 20899 (http://webbook.nist.gov).
9. Poling, B.E., Prausnitz, J. M., O'Connel, J. P., Properties of Gases and Liquids, 5th Ed., McGraw-Hill, New York (2001).
10. Rowley, R.L.; Wilding, W.V.; Oscarson, J.L.; Yang, Y.; Zundel, N.A. DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties, (http//dippr.byu.edu), Brigham Young University Provo Utah, 2006.
11. Shacham, M.; Brauner, N. The SROV Program for Data Analysis and Regression Model Identification. *Computers Chem. Engng*. 2003, 27, 701.
12. Shacham, M.; Kahrs, O.; St Cholakov, G.; Stateva, R.; Marquardt, W.; Brauner, N. The Role of the Dominant Descriptor in Targeted Quantitative Structure Property Relationships, *Chem. Eng. Sci*. 2007, 62, (22), 6222-6233.
13. Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *Journal of Medicinal Chemistry*. 1972, 15(10), 1066.

Table 1. Trend of change of descriptors with n_C for homologous series

Category	Trend of change of the descriptor with n_C	% of descriptors in the database	% of 3D descriptors
I	Constant	8.5	7.3
II	Linear or nearly linear increase	10.9	41.7
IIIA	Nonlinear monotonic increase or decrease with decreasing slope	25.2	32.1
IIIB	Nonlinear monotonic increase or decrease with increasing slope	10.1	66.7
IV	Inconsistent, no particular trend or different trends for different homologous series	22.9	83.6
V	Zero value for some $n_{C,}$ nonlinear monotonic increase for others	21.9	62.9
VI	Separate curves for odd and even n_C	0.4	100.0
VII	Periodic	0.2	100.0

Table 2. Members of the training set for the target compound: *n*-dodecane (n_C = 12)

No.	Compound	n_C	$\lvert r_{ti}\rvert$*
1	*n*-tridecane	13	0.99948
2	*n*-undecane	11	0.99946
3	*n*-tetradecane	14	0.99798
4	*n*-decane	10	0.99779
5	*n*-pentadecane	15	0.99555
6	*n*-nonane	9	0.99493
7	*n*-hexadecane	16	0.99227
8	*n*-octane	8	0.99082
9	*n*-heptadecane	17	0.98822
10	*n*-heptane	7	0.98533

*Type II descriptors (3 – D descriptors excluded) used for calculation of $\lvert r_{ti}\rvert$

Table 3. Statistics of *δ* (absolute difference (in %) between the predicted and DIPPR recommended values) using TQSPRs without 3-D descriptors

Group	Statistics	T_C	P_C	V_C	T_m	T_b	V_m	*RI*
n-alkanes	mean	0.04	0.48	0.35	1.80	0.22	0.29	0.01
	median	0.03	0.36	0.22	0.11	0.03	0.22	0.00
	STDEV	0.05	0.38	0.39	6.86	0.81	0.24	0.01
1-alkenes	mean	0.14	0.97	1.78	0.89	0.19	0.31	0.06
	median	0.07	0.56	0.19	0.16	0.03	0.09	0.04
	STDEV	0.24	1.48	4.83	1.21	0.40	0.47	0.06
n-alkylbenzenes	mean	0.13	0.29	0.51	0.17	0.03	0.08	0.00
	median	0.08	0.27	0.29	0.09	0.02	0.08	0.00
	STDEV	0.15	0.23	0.76	0.20	0.04	0.05	0.00
1-alcohols	mean	0.15	1.74	3.53	2.52	0.17	0.50	0.01
	median	0.07	0.91	1.37	0.17	0.17	0.55	0.01
	STDEV	0.23	2.60	6.17	5.38	0.15	0.31	0.01
aliphatic acids	mean	0.30	0.97	2.83	0.17	0.18	0.75	0.02
	median	0.22	0.70	0.40	0.11	0.10	0.23	0.01
	STDEV	0.36	0.87	5.05	0.13	0.19	1.89	0.02

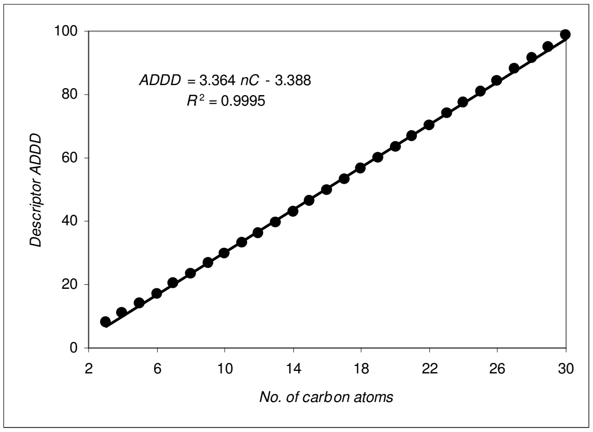

Figure 1. Plot of the descriptor *ADDD* versus the number of carbon atoms for the 1-alkene homologous series.

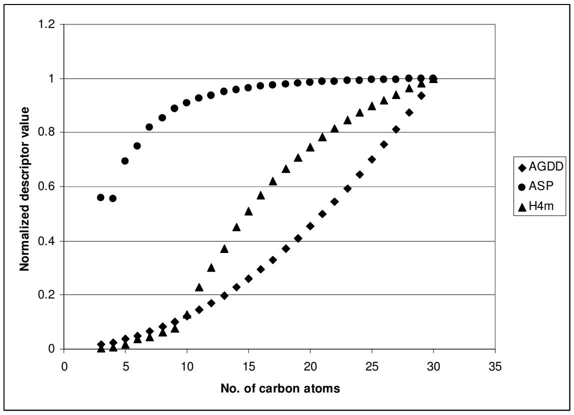

Figure 2. Plot of the normalized values of the descriptors *AGDD*, *ASP* and *H4m* versus the number of carbon atoms for the 1-alkene homologous series.

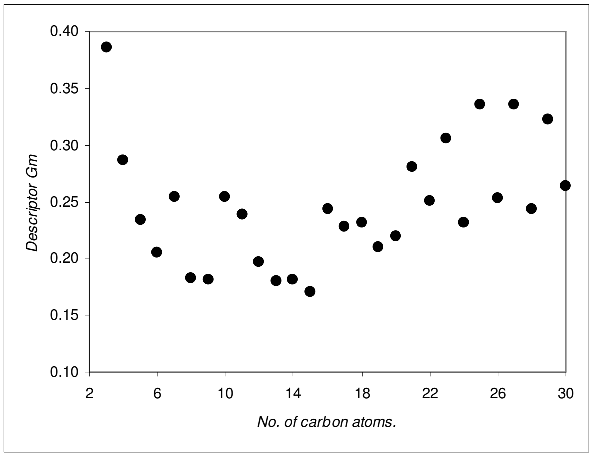

Figure 3. Plot of the descriptor *Gm* versus the number of carbon atoms for the 1-alkene homologous series.