ELSEVIER

A TOPS-MODE approach to predict permeability coefficients

Maykel Pérez González[a,b,c,*], Aliuska Morales Helguera[b,d], Humberto González Díaz[b,c]

[a]*Unit of Services, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo 54830, Cuba*
[b]*Department of Drug Design, Chemical Bioactives Center, Central University of Las Villas, Santa Clara, Villa Clara, Cuba*
[c]*Department of Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Santiago de Compostela 15706, Spain*
[d]*Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, Villa Clara, Cuba*

Abstract

The TOPological Sub-Structural Molecular Design (TOPS-MODE) approach has been applied to the study of the permeability coefficient of various compounds through low-density polyethylene at 21.1 °C. A model able to describe closed to 90% of the variance in the experimental permeability of 63 organic compounds was developed with the use of the mentioned approach. In contrast, no one of nine different approaches, including the use of constitutional, topological, BCUT, 2D autocorrelations, geometrical, RDF, 3D Morse, WHIM and GETAWAY descriptors was able to explain more than 73% of the variance in the mentioned property with the same number of descriptors. In addition, genetic algorithms were used in feature selection experiments considering all molecular descriptors in order to obtain mixed models. Although, statistically significant models were derived containing other descriptors than spectral moments still the best one fitted out model was find with these variables. Finally, the TOPS-MODE approach permitted to find the contribution of different fragments to the permeability coefficients giving to the model a straightforward structural interpretability.
© 2003 Published by Elsevier Ltd.

Keywords: Permeability coefficients; QSPR; TOPS-MODE

1. Introduction

The permeability of polymeric materials is an important consideration in the pharmaceutical industry. Many pharmaceutical preparations need to be protected from oxygen, water vapor, carbon dioxide and other penetrants. The polymeric packaging material for the preparation must serve as an effective barrier to gaseous diffusion. In other applications, the polymeric material serves as a core to the controlled release of an active biological agent. Solubility and diffusion of the (usually) small organic biological agent in the polymer matrix is central to the controlled release behavior.

The penetration of molecules through polymer films is named 'permeability'. There are many dimensions and units found in literature for the general expression 'permeability' [1].

The permeability coefficient, in a strict sense, is not only a function of the chemical structure of the polymer. It varies also with the morphology of the polymer and depends on many physical factors such as density, crystallinity and orientation of the polymer chains. However, the chemical structure of a polymer and the permeant can be considered to be the predominant factor that controls the magnitude of the permeability coefficient [2].

If a permeant does not interact with the polymer under investigation, the permeability coefficient is characteristic for the permeant-polymer system. This is the case with the permeation of many gases, such as H_2, He, N_2, O_2 and CO_2, through many polymers. On the other hand, if a permeant interacts with the polymer the permeation coefficient is no longer a constant, and many of them depend on the special conditions of the measurement and on the history of the polymer film. In such cases, a single value of the permeability coefficient does not represent the characteristic permeability of the polymer and it is necessary to know the dependency of the permeability coefficient of all possible variables in order to obtain the complete profile of the permeability of the polymer [3].

In the context of in silico methods for modeling

* Corresponding author. Address: Unit of Services, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo 54830, Cuba. Tel.: +53-534-228-1473; fax: +53-422-811-30.
  E-mail address: mpgonzalez76@yahoo.es (M.P. González).

physicochemical and biological properties of chemicals the topological sub-structural molecular design (TOPS-MODE) approach has been introduced [4–19].

The successful applications of this theoretical approach to the modeling of physical and physico-chemical properties [4–9,13] have inspired us to perform a more exhaustive study in order to test and/or validate the TOPS-MODE applicability in this area. The permeability through low-density polyethylene was previously studied using the physico-chemical properties as molecular for a limited data set of compounds. [3]. Here we will investigate the role that TOPS-MODE and other molecular descriptors calculated from the molecular structure play on the explanation of such property using an extended data set of 63 organic compounds.

2. The TOPS-MODE approach

TOPS-MODE is based on the computation of the spectral moments of the bond matrix, whose mathematical basis was described in previous reports by Estrada [4–6]. The TOPS-MODE approach has been recently reviewed in the literature [17], given a methodological explanation of how to use it as well as a software description [18].

According to Estrada the application of the TOPS-MODE approach to the study of quantitative structure–property relationships (QSPR) can be resumed in the following set of steps:

1. To draw the hydrogen-depleted molecular graphs for each molecule of the data set.
2. To use appropriated bond weights in order to differentiate the molecular bonds, e.g. bond distance, bond dipoles, bond polarizabilities, etc.
3. To compute the spectral moments of the bond matrix with the appropriated weights for each molecule in the data set, generating a table in which rows correspond to the compounds and columns correspond to the spectral moments of the bond matrix. Spectral moments are defined as the trace of the different powers of the bond matrix [16].
4. To find a QSPR by using any appropriated linear or non-linear multi-variate statistical technique, such as multi-linear regression analysis (MRA), etc.:

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + \cdots + a_k\mu_k + b \quad (1)$$

where P is the property measurement, μ_k is the kth spectral moment, and a_k's are the coefficients obtained by the MRA.

5. To test the predictive capability of the QSPR model by using cross-validation techniques.
6. To compute the contributions of different groups of interest in order to determine their quantitative contribution to the permeability of molecules under study.

The computation of fragment contributions to the permeability property under study is probably the most important advance of the TOPS-MODE approach to the study of permeability variables compared to the traditional QSAR and QSPR methods. The procedure consists of calculating the spectral moment for all the fragments contained in a given substructure, and by difference of these moments we obtain the contribution of the sub-structure. The general algorithm for this computational approach is as follows.

First, we select the substructure whose contribution to the moments we would like to determine. Then, we generate all the fragments, which are contained in the corresponding substructure, and calculate the spectral moments for both, the substructure and all their fragments. The contribution of the substructure to the spectral moments is finally obtained as the difference between the spectral moments of the substructure and all those from their fragments. Once, the contributions of the different structural fragments are obtained, we only need to substitute these contributions into the quantitative model developed to describe the property studied.

3. Data sets and computational strategies

A data set of 63 compounds for which the permeability coefficients were reported in the literature was selected [2]. The parameter studied is the log(p) where p is the permeability coefficient through low-density polyethylene. These kinds of studies usually have a variation coefficient of 3–5% for this method [2]. The names of the compounds, as well as the calculated and experimental values of log(p) are shown in Table 1.

TOPS-MODE [18] and DRAGON [21] computer softwares were employed to calculate the molecular descriptors. In the case of TOPS-MODE software, only the polar surface was used as bond weight for making differentiations of heteroatoms. The selection of only this type of descriptors from the whole pool of six types included in TOPS-MODE methodology was carried out on the sake of simplicity and on the belief that polarity parameters influence the permeability of compounds through polymer layers. The total number of descriptors used for obtaining this model was 15 spectral moments. On the other hand, we carry out geometry optimization calculations for each compound used in this study using the quantum chemical semi-empirical method AM1 [22] included in MOPAC 6.0 [23]. Other nine models were developed using the computer software Dragon [21], calculating the Constitutional, Topological, BCUT, 2D autocorrelations, Geometrical, RDF, 3D-MORSE, WHIM and GETAWAY descriptors were [20]. The statistical processing to obtain the QSAR models was carried out by using the forward stepwise regression methods. In addition to the models considering one specific family of

Table 1
The observed, predicted, and residual values of permeability coefficients of 63 compounds through low-density polyethylene used to derive the QSPR

Number	Compounds	Observed	Predicted	Deleted Residuals
1	Acetaldehyde	0.373	1.114	−0.756
2	Acetic anhydride	−0.495	−0.486	−0.010
3	Acetone	0.427	1.022	−0.607
4	Allyl alcohol	−0.244	−0.152	−0.094
5	Amyl acetate	0.534	0.344	0.196
6	*i*-Amyl alcohol	−1.108	−0.425	−0.705
7	*i*-Amyl propionate	0.591	0.253	0.350
8	*n*-Amyl propionate	1.072	0.253	0.847
9	Aniline	−0.174	−0.288	0.159
10	Benzaldehyde	0.427	0.929	−0.513
11	Benzene	2.238	2.123	0.121
12	Benzoic acid	−1.553	−1.455	−0.106
13	Benzyl alcohol	−0.796	−0.336	−0.473
14	*n*-Butyl acetate	0.771	0.435	0.345
15	*n*-Butyl alcohol	−0.745	−0.334	−0.423
16	*s*-Butyl alcohol	−0.620	−0.334	−0.294
17	*t*-Butyl alcohol	−1.000	−0.334	−0.685
18	Butyraldehyde	0.594	0.932	−0.344
19	Carbon tetrachloride	2.380	2.307	0.078
20	Chloroacetic acid	−0.921	−1.271	0.374
21	*m*-Chloroaniline	−0.201	−0.288	0.122
22	Chlorobenzene	2.253	2.123	0.137
23	Chloromaleic anhydride	−0.553	−0.395	−0.166
24	*p*-Chloro toluene	2.076	2.032	0.046
25	Cyclohexane	1.994	1.853	0.147
26	Decane	1.447	1.488	−0.042
27	1,2-Dibromoethane	1.740	2.216	−0.503
28	Dibutyl ether	1.526	1.355	0.176
29	*o*-Dichlorobenzene	1.785	2.123	−0.356
30	Diethyl ether	2.090	1.719	0.383
31	Diethyl oxalate	−0.553	−0.982	0.473
32	Ethyl acetate	0.813	0.617	0.200
33	Ethyl alcohol	−0.553	−0.152	−0.411
34	Ethylene glycol monobutyl ether	−0.585	−0.830	0.258
35	Ethyl formate	0.771	0.708	0.064
36	Ethyl mercaptan	2.196	2.259	0.062
37	Ethyl propionate	0.991	0.526	0.477
38	Formamide	−0.699	−1.208	0.704
39	Formic acid	−0.585	−1.180	0.632
40	Furfuryl alchol	−1.046	−0.848	−0.208
41	*n*-Heptane	2.025	1.761	0.274
42	*n*-Heptene	2.025	1.851	0.181
43	*n*-Heptyl acetate	0.851	0.162	0.716
44	Heptyl alcohol	−0.409	−0.607	0.205
45	*n*-Hexane	2.140	1.852	0.299
46	Methyl acetate	0.740	0.708	0.033
47	Methyl alcohol	−0.319	−0.061	−0.264
48	Methyl ethyl ketone	0.695	0.931	−0.241
49	Methylcyclohexane	2.033	1.762	0.281
50	Nitrobenzene	0.286	−0.096	0.392
51	Nitroethane	0.025	−0.003	0.029
52	Nitromethane	−0.081	0.088	−0.174
53	Octyl alcohol	−0.699	−0.698	−0.001
54	*i*-Pentane	2.025	1.943	0.086
55	*n*-Pentane	2.314	1.943	0.387
56	Phenol	−0.699	−0.264	−0.447
57	*n*-Propil alcohol	−0.699	−0.243	−0.468
58	Tetradecane	0.756	1.124	−0.379
59	Tetradecene	0.672	1.214	−0.558
60	Toluene	2.299	2.032	0.280
61	1,1,1-Trichloroethane	2.009	2.216	−0.219
62	*o*-Xylene	2.004	1.941	0.066
63	*p*-Xylene	2.281	1.941	0.355

Table 2
The statistical parameters of the lineal regressions models obtained for the ten kinds of descriptors

Descriptors	No. of variables	S	R^2	F	p	q^2
Spectral moments	3	0.366	0.907	192.271	0.000	0.877
Constitutional	4	0.923	0.421	10.547	0.000	0.378
Topological	8	0.542	0.814	29.579	0.000	0.785
BCUT	7	0.845	0.539	9.194	0.000	0.412
2D autocorrelations	10	0.721	0.683	11.202	0.000	0.612
Geometrical	5	0.885	0.477	10.402	0.000	0.425
RDF	12	0.469	0.870	28.079	0.000	0.825
3D-MORSE	12	0.541	0.825	19.690	0.000	0.762
WHIM	8	0.846	0.546	8.131	0.000	0.501
GETAWAY	12	0.659	0.745	12.205	0.000	0.703

descriptors mixed models with the entire pool of descriptors were seek. In this experiment feature selection was carry out by means of genetic algorithm. All the parameters such as population size mutation probabilities, cross-over probabilities, smoothing and so on where fixed at their default values [20].

The statistical significance of the models was determined by examining the regression coefficient, the standard deviation, the number of variables, the cross validation leave-one-out statistics and the proportion between the cases and variables in the equation.

4. Quantitative structure permeation relations

The best QSPR model obtained with the TOPS-MODE descriptors is given below together with the statistical parameters of the regression.

$$\log(p) = 2.3985 - 0.00\mu_2^{PS} + 1.87 \times 10^{-5}\mu_4^{PS}$$
$$- 3.50 \times 10^{-7}\mu_5^{PS}$$

$$N = 63 \ \ S = 0.366 \ \ R^2 = 0.907 \ \ F = 192.271$$
$$p = 0.000 \ \ q^2 = 0.877 \ \ S_{cv} = 0.391$$

(2)

Where N is the number of compounds included in the model, R^2 is the correlation coefficient, S the standard deviation of the regression, F the Fisher ratio, q^2 the correlation coefficient of the cross-validation, p is the significance of the variables in the model and S_{cv} is the standard deviation of the cross-validation.

The variables included in the model are the designed as follow: the sub-index represents the order of the spectral moment and the super-index the type of bond weight used, i.e. PS for polar surface.

The structural significance of this model will be more evident later when we analyze the contribution of the different structural fragments to the permeability through the polyethylene layer. From the statistical point of view this model is a robust one as can be seen from the statistical parameters of the cross-validation.

As we advance previously one of the objectives of the current work is to compare the reliability of the TOPS-MODE approach to describe the property under study as compared with other different descriptors and methods. Consequently, we have developed other nine models using the same data set that was included in the TOPS-MODE QSPR model. The results obtained with the use of Constitutional, Topological, BCUT, 2D autocorrelations, Geometrical, RDF, 3D Morse, WHIM, and GETAWAY descriptors are given in Table 2.

As can be seen there are remarkable differences concerning the explanation of the experimental variance given by these models compared to the TOPS-MODE one. While the TOPS-MODE QSPR model explains more than 90% of permeability the rest of the models are unable to explain more than 87% of such variance. The Fig. 1 shows a linear regression between the predicted and observed values for log P for the Eq. (2).

The model obtained using the RDF descriptors explain the 87% of the data variance, but this model need 12 variables for this. In the case of TOPS-MODE only need three variables, this shown the great superiority of this model over the best generate with the Dragon software.

On the other hand, models with these kinds, limiting the number of variables was carry out. This is carried out with the objective of compare the statistic parameters of these models when all of them have the same variables number. As can be seen, in the Table 3, the results obtained with the model using the spectral moments are much better to the rest of the using methodologies, which are unable of explain more than of the 73% of the data variance, besides of present important statistic parameters of higher quality that all the models obtained, such as the Fischer ratio (F) and the standard deviation (S).

The TOPS-MODE model not only overtakes the other nine models in the statistical parameters of the regression but more importantly in the stability to the inclusion–exclusion of compounds as measured by the correlation coefficient and standard deviation of the cross-validation. Because of the structural variability of the compounds in the data set these statistics of the leave-one-out cross validation
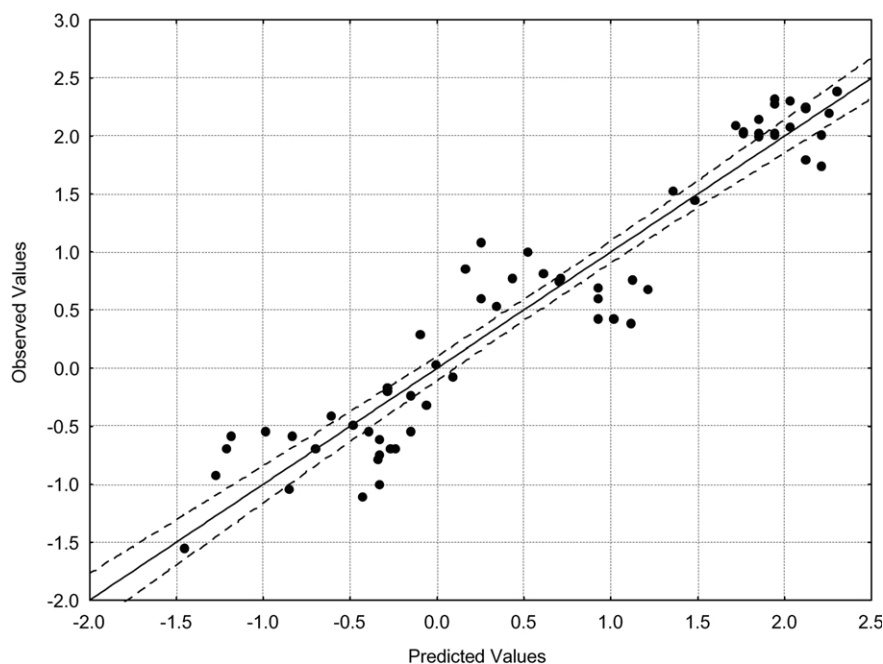
Fig. 1. The linear relation between observed and predicted permeability for the compounds of the training set.

might be considered as a good measurement of the predictability of the models. As can be seen in the Table 3 the value of the determination coefficient of leave-one-out cross validation for the model obtained with the spectral moments ($q^2 = 0.877$) was the highest for the all analyses model proving the high predict power of this approach and the high stability of the model.

However, in all previous studies we only consider models with a specified family of molecular descriptors. Thence, in order to complete the demonstration of the potentialities of TOPS-MODE over the remnant ones mixed models considering all the molecular descriptors at the same time must be developed. The total number of molecular descriptors considered here is higher than 1000. Thus, a strategy for feature selection is necessary. In this sense, we performed a genetic algorithm previous to forward stepwise regression analysis. Table 4 depicts the results of this study. It is interesting to note that spectral moments are selected in almost models regardless of the number of generation used in the generic algorithm feature selection and the number of variables in the model. Anyhow, in our opinion the most interesting result is that the best model found coincides with the one reported in Eq. (2). These results have shown that the TOPS-MODE approach not only explains the experimental data, but seems to be the best one in doing so.

5. Fragments contributions

One of the most important advantages that TOPS-MODE brings for the study of QSPR and QSAR is that concerned with the structural interpretability of the models. This interpretability comes from the fact that the spectral moments can be expressed as linear combinations of structural fragments. In such a way, we can learn what fragments are making a positive or negative contribution to the property under study, which can be interpreted in terms of the physicochemical or biological processes influencing

Table 3
The statistical parameters of the lineal regressions models obtained white three variables for the ten kinds of descriptors

Descriptors	Variables	S	R^2	F	p	q^2
Spectral moments	μ_2^{PS}, $\mu >_4^{PS}$, μ_5^{PS}	0.366	0.907	192.271	0.000	0.877
Constitutional	MW, nN, nO	0.918	0.416	14.048	0.000	0.378
Topological	ZM1V, IC1, SEigp	0.808	0.548	23.871	0.000	0.517
BCUT	BEHm3, BELp2, BELe2	0.948	0.379	12.005	0.000	0.325
2D autocorrelations	ATS1p, GATS3e, GATS2e	0.913	0.423	14.458	0.000	0.402
Geometrical	G(O⋯Cl), MAXDP, FDI	0.897	0.443	15.663	0.000	0.413
RDF	RDF010u, RDF010m, RDF020e	0.623	0.731	53.473	0.000	0.698
3D-MORSE	Mor23m, Mor32m, Mor28p	0.926	0.407	13.512	0.000	0.346
WHIM	Gm, Gu, Ds	1.000	0.308	8.780	0.000	0.232
GETAWAY	ISH, R1e, RTp	0.911	0.426	14.641	0.000	0.351

Table 4
The models and statistical parameters obtained using a genetic algorithm feature selection previous to forward stepwise analysis

Generations	Variables	S	R^2	F	p	q^2
100	μ_2^{PS}, μ_4^{PS}, RDF010m	0.560	0.784	71.414	0.000	0.72
	RDF010m, RDF010u	0.634	0.717	76.043	0.000	0.63
1000	μ_2^{PS}, μ_4^{PS}, μ_5^{PS}	0.366	0.907	192.271	0.000	0.87
	IC1, μ_2^{PS}	0.862	0.477	27.414	0.000	0.38
5000	μ_2^{PS}, μ_4^{PS}, Ms	0.555	0.786	72.515	0.000	0.471
	μ_2^{PS}, Ms	0.906	0.477	21.964	0.000	0.39

it. In Table 5 and the Fig. 2 we show the fragments and their contributions to the permeation coefficient across the low-density polyethylene as calculated from the Eq. (2).

Here, we have only study some small fragments present in the structures of the compounds in the data set. However, the extension of this study to other fragments in such molecules or even to fragments in molecules not contained in this data set is straightforward and it has been shown for other particular cases elsewhere [8,10,11,19].

According to the contributions of the fragments F1 to F4 (see Fig. 2) the number of carbons increases in a chain the permeation coefficient decreases. This is a logical size effect involving the length of the chains in compounds as a factor limiting the permeability through polyethylene layers.

On the other hand, it is observed from the contributions of fragments F5 to F7 that an increase of the polarity of the heteroatom produces a decrease of the permeation. However, the contributions of the heteroatom are also dependent on its volume as can be observed in the smaller difference in contribution between the fragments F6 and F7. Sulfur has a bigger atomic volume than nitrogen, but the nitrogen has higher polarity than the sulfur, and thence the result of this effect is a delay in the permeation process. A similar trend is observed when we analyze the fragments F8 to F10.

Finally, when we compare fragments F6, F11 and F12, we can observe that F11 delays approximately 44 times more the process of permeation than the fragment F6 and 186 times than the fragment F12, in spite of the fact that the former are smaller than the later.

This should be caused by the negative influence exerted by the methyl group over the nitrogen in the F6 and F12 in his interaction with the polymeric matrix. Several other interesting relations can be obtained by analyzing the contributions of the fragments in this table.

6. Concluding remarks

We have shown that the TOPS-MODE approach is able to describe the permeability of different compounds through low-density polyethylene at 21.1 °C. In fact, we have developed a model for predicting the permeability coefficient of a data set of 63 permeants, which is both statistically and chemically sounded. This model explains more than 90% of the variance in the experimental permeability coefficients with a good predictive power. These features are significantly better than that obtained from nine other different methodologies.

On the other hand, the main advantage of using TOPS-MODE approach in QSPR/QSAR has been confirmed again in this work. This approach is able to derive group contributions and gives simultaneously a valuable capability

Table 5
The contribution of different groups to the permeability coefficients through low-density polyethylene

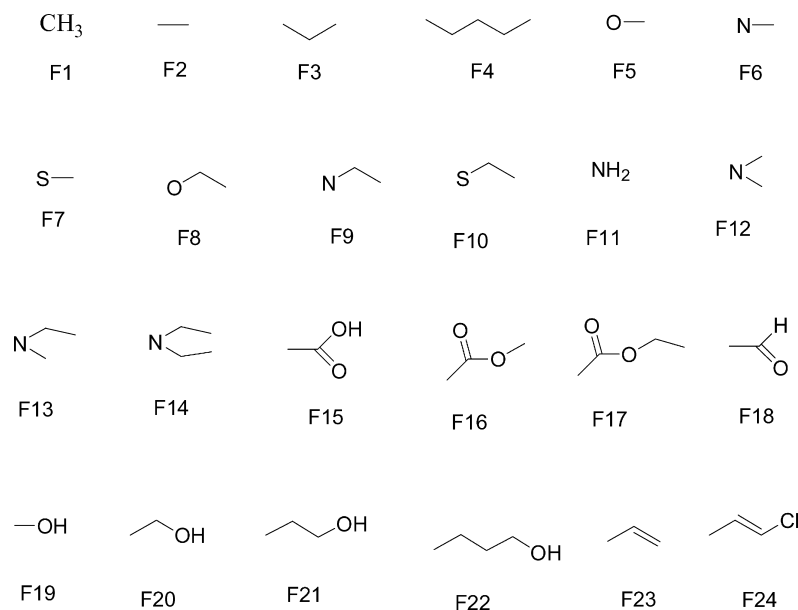| Studied fragments	Group contribution	Studied fragments	Group contribution
F1	0.810	F13	0.453
F2	0.657	F14	0.367
F3	0.533	F15	0.0002
F4	0.432	F16	0.025
F5	0.016	F17	0.020
F6	0.132	F18	0.064
F7	0.160	F19	0.003
F8	0.011	F20	0.002
F9	0.114	F21	0.0022
F10	0.150	F22	0.001
F11	0.003	F23	0.650
F12	0.559	F24	0.551

Fig. 2. Structures of selected fragments for which their contributions to the permeability coefficient.

of interpretation, contributing to understanding the physico-chemical or biological processes involved.

Acknowledgements

References

[1] Minoura N, Tani S, Nakagawa T. J Polym Sci 1978;(22):833.
[2] Brandrup J, Immergut EH, 3rd ed. Polymer handbook, New York: Wiley; 1989. p. V122.
[3] Patel H, Tokarski J, Hopfinger A. Pharm Res 1997;(10):1349–54.
[4] Estrada E. J Chem Inf Comput Sci 1996;(36):844–9.
[5] Estrada E. J Chem Inf Comput Sci 1997;(37):320–8.
[6] Estrada E. J Chem Inf Comput Sci 1998;(38):23–7.
[7] Estrada E. J Chem Soc, Faraday Trans 1998;(94):1407–11.
[8] Estrada E, Gutiérrez Y. J Chromatogr, A 1999;(858):187–99.
[9] Estrada E, Gutiérrez Y, González H. J Chem Inf Comput Sci 2000; (40):1386–99.
[10] Estrada E, Peña A. Bioorg Med Chem 2000;(8):2755–70.
[11] Estrada E, Uriarte E. SAR QSAR Environ Res 2001;(12):309–24.
[12] Estrada E, Uriarte E, Gutierrez Y, Gonzalez H. SAR QSAR Environ Res 2003;(14):145–63.
[13] Estrada E, Gonzalez H. J Chem Inf Comput Sci 2003;(42):75–84.
[14] Estrada E. J Mol Graphics Modell 2001;(20):54–74.
[15] Estrada E. SAR QSAR Environ Res 2000;(11):55–73.
[16] Estrada E. J Chem Inf Comput Sci 1995;(35):31–3.
[17] Estrada E. SAR QSAR Environ Res 2000;(11):55–73.
[18] Gutierrez Y, Estrada E. TOSS-MODE (Topological Sub-Structural Molecular Design) for Windows Version 4.0, Universidad de Santiago de Compostela, Spain; 1997.
[19] González MP, González HD, Molina RR, Cabrera MA, Ramos RA. J Chem Inf Comput Sci 2003;(43):1192–99.
[20] Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley; 2000.
[21] Todeschini R, Consonni, V, Pavan M. Dragon. Software version 2.1; 2002.
[22] Michael JS, Dewar E, Zoebisch G, Eamonn F, Stewart JP. J Am Chem Soc 1985;(107):3902–9.
[23] Stewar JJP. MOPAC manual, 6th ed. Colorado Springs, CO: Frank J. Seiler Research Laboratory, U.S. Air Force academy; 1990. p. 189.