Infrared Spectra Information and Their Correlation with QSAR Descriptors

Romualdo Benigni,*,† Laura Passerini,† David J. Livingstone,‡ Mark A. Johnson,§ and
Alessandro Giuliani†

Istituto Superiore di Sanita’, Laboratory of Comparative Toxicology and Ecotoxicology, Viale Regina Elena
299 - 00161 Rome Italy, ChemQuest, 19-21 Cheyney St., Steeple Morden, UK, Centre for Molecular Design,
University of Portsmouth, U.K., and Pharmacia and Upjohn, Kalamazoo, Michigan 49007-4949

Received October 5, 1998

In principle, the InfraRed (IR) spectra have very appealing properties for QSAR research: they are generated in the range of low energy molecular interactions that play a fundamental role in life (e.g., for molecular recognition), and they are extremely specific fingerprints of the molecules. We compared the information carried by the fingerprint region of the IR spectra (1500−600 cm^{-1}) with that of a range of descriptors presently in use in the QSAR field: (a) classical physical chemical and quantum mechanical properties (logP, MR, HOMO, LUMO); (b) molecular connectivities; (c) 2-D molecular distances; and (d) a novel infrared range vibration based theoretical descriptor (EVA). Much redundancy and overlapping was found among descriptors such as connectivities, 2-D distances, and theoretical spectral EVA descriptors. On the contrary, the complex information carried by IR spectra (fingerprint region) was markedly different from that codified in various molecular descriptors presently in use in QSAR practice, thus pointing to the importance of further studying the potential relevance of IR information for QSAR analysis.

INTRODUCTION

A crucial component of the QSAR research is the identification of molecular descriptors relevant to the chemical or biological activity of interest. The range of descriptors presently available for the investigators is impressive: descriptors closely related to chemical physical theory, substructural descriptors, quantum mechanical descriptors at various approximation levels, theoretical descriptors founded on the topology of the molecules, 3-D descriptors from the calculation of interaction energies or based on transformation of atomic coordinates, experimentally determined properties such as solubility, melting point, half-wave reduction potential, and so on.[1,2] The derivation of such a huge amount and variety of descriptors is due to a number of causes, including the apparent inability of any one type of parameter to describe adequately all features of biologically active compounds; the need to describe different aspects of the chemical/biological system interaction; a requirement for easily interpreted properties; the desire to compute descriptors quickly and easily; the failure of “traditional” physical chemical properties to fully characterize the 3-D aspects of molecules; and the need for parameters which can be applied to noncongeneric series of compounds.

This paper focuses on a “natural” molecular descriptor that hitherto has received very little attention in the QSAR field: the infrared (IR) spectra. In principle, IR spectra have very appealing properties for this type of research: they are generated in the range of low energy molecular interactions that play a fundamental role in life (e.g., molecular recognition),[3] and they are extremely specific fingerprints of the molecules (two molecules with identical IR spectra do not exist, except for optical enantiomorphs). This particularly applies to the so-called fingerprint region (1500−600 cm^{-1}).[4]

A very limited number of QSAR/QSPR studies employing IR spectra have been reported. In these papers, the authors have found a relationship between $\Delta\nu$ (shift of location of a certain peak) values and electronic parameters (e.g., refs 5−8). To the best of our knowledge, no examples have been reported in which the information carried by the IR spectra as a whole has been explored. As a matter of fact, it is the entire pattern of an IR spectrum that it is unique and makes each molecule different from every other molecule. Thus, the specific goal of this paper is to compare the information carried by the fingerprint region of IR spectra with that provided by a range of descriptors presently in use in the QSAR field: (a) descriptors more closely related to chemical physical properties (logP, MR, HOMO, LUMO); (b) molecular connectivities; (c) 2-D molecular distances from bond deletion metric on chemical graphs;[9] and (d) novel theoretical descriptors based on normal coordinate vibrational EigenVAlues (EVA).[10] For the purpose of this study, IR spectral information is represented by numerical intensities over a confined “fingerprint” region.

It should be stressed that this paper is not aimed at further understanding or predicting IR spectra but rather at considering raw IR spectral information for its potential use within QSAR practice. In particular, we checked if, and to what extent, the information contained in the fingerprint region of IR is different from (hence potentially complementary to) that of other QSAR descriptors.

* Corresponding author: fax: + 39 - 06 - 49387139; e-mail: rbenigni@iss.it.

† Istituto Superiore di Sanita’.

‡ ChemQuest.

§ Pharmacia and Upjohn.

10.1021/ci980223x CCC: $18.00
Published on Web 02/25/1999

Table 1.

chemical	CAS no	chemical	CAS no
tert-butylhydroquinone	1948-33-0	naphthalene	91-20-3
1,2,3-benzotriazole	95-14-7	nitromethane	75-52-5
acetonitrile	75-05-8	propyl gallate	121-79-9
diallyl phthalate	131-17-9	pentachloroethane	76-01-7
benzyl alcohol	100-51-6	pentachlorophenol	87-86-5
p,*p*′-dichlorodiphenylethylene	72-55-9	resorcinol	108-46-3
chlorodibromomethane	124-48-1	styrene	100-42-5
2-chloroethanol	107-07-3	propylene	115-07-1
chlorobenzene	108-90-7	phenol	108-95 2
coumarin	91-64-5	*o*-phenylphenol	90-43-7
bromodichloromethane	75-27-4	piperonyl butoxide	51-03-6
1,3-butadiene	106-99-0	1-phenyl-3-methyl-5-pyrazolone	89-25-8
benzaldehyde	100-52-7	1,1,1,2-tetrachloroethane	630-20-6
butyl benzyl phthalate	85-68-7	tetrafluoroethylene	116-14-3
benzofuran	271-89-6	tetranitromethane	509-14-8
butyl benzyl phthalate	85-68-7	4-vinylcyclohexene	100-40-3
butylated hydroxytoluene	128-37-0	1,1,2,2-tetrachloroethane	79-34-5
n-butylchloride	109-69-3	tetrachloroethylene	127-18-4
acetamide	60-35-5	tribromomethane	75-25-2
benzene	71-43-2	toluene	108-88-3
butyrolactone gamma	96-48-0	1,1,1-trichloroethane	71-55-6
caprolactam	105-60-2	trichloroethylene	79-01-6
d-carvone	2244-16-8	vinylidene chloride	75-35-4
methylmethacrylate	80-62-6	1,2,3-trichloropropane	96-18-4
dl-menthol	153567-00-4	triethanolamine	102-71-6
ethyleneglycol	107-21-1	*p*-chloroaniline	20265-96-7
1,2-dichlorobenzene	95-50-1	4-aminobiphenyl	92-67-1
1,1-dichloroethane	75-34-3	*m*-cresidine	102-50-1
dichloromethane	75-09-2	3-chloro-*p*-toluidine	95-74-9
eugenol	97-53-0	2,4-diaminotoluene	950-80-7
3,4-dihydrocoumarin	119-84-6	*N*,*N*-dimethylaniline	121-69-7
isophorone	78-59-1	2,4-dimethoxyaniline	54150-69-5
dimethyl terephthalate	120-61-6	*N*-phenyl-2-naphthylamine	135-88-6
isobutyl nitrite	542-56-3	*o*-toluidine	636-21-5
1,4-dioxane	123-91-1	*p*-nitroaniline	100-01-6
epoxyhexadecane	7320-37-8	2,4-dinitrotoluene	121-14-2
furan	110-00-9	*o*-nitroanisole	91-23-6
furfural	98-01-1	parathion	56-38-2
geranyl acetate	105-87-3	*p*-nitrophenol	100-02-7
malathion	121-75-5	1-nitronaphthalene	86-57-7
1,4-dichlorobenzene	106-46-7	bis(2-chloro-l-methyl ethyl) ether	108-60-1
diethylphthalate	84-66-2	3-chloro-2-methylpropene	563-47-3
hydroquinone	123-31-9	allyl isothiocynate	57-06-7
heptachlor	76-44-8	chloroethane	75-00-3
lindane	58-89-9	2-chloroacetophenone	532-27-4
2,3-dibromo-1-propanol	96-13-9	allyl glycidyl ether	106-92-3
hexachlorocyclopentadiene	77-47-4	bromoethane	74-96-4
hexachloroethane	67-72-1	1,2-epoxybutane	106-88-7
methyl bromide	74-83-9	glycidol	556-52-5
hexachloroethane	67-72-1	dichlorvos	62-73-7
di(2-ethylhexyl) phthlate	117-81-7	1,2-dichloropropane	78-87-5
4-hexylresorcinol	136-77-6	1,2-dichloroethane	107-06-2
d-limonene	5989-27-5	ethyl acrylate	140-88-5
2,4-dichlorophenol	120-83-2	telone II	542-75-6
iodoform	75-47-8	propylene oxide	75-56-9
monochloroacetic acid	79-11-8	1,1,2-trichloroethane	79-00-5

DATA AND METHODS

The chemicals used for this study are listed in Table 1. As evident, the sample is relatively small, but very heterogenous, thus allowing for a general characterization of IR information independently from individual chemical series.

IR Spectra. The spectra, all derived in gas phase, were retrieved from (a) the Aldrich compilation[11] and (b) the National Toxicology Program technical reports (U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health). The printed spectra were scanned with an Epson GT-6500 scanner. With the program Spectrum (Delta Sistemi, Roma), written and hoc for this study, we sampled 91 values in the range 1500−600 cm^{-1} (fingerprint region) (interval 10 cm^{-1}). Each spectrum was normalized in the range 0.0−100.0, by setting the minimum value to zero and the maximum to 100 and then linearly scaling the other values between these two extremes. This normalization scheme was aimed at correcting for different backgrounds and experimental singularities.

Classical Descriptors. Calculated logP and Molecular Refractivity (MR) values were taken from a previous paper.[12] The frontier orbital energies, HOMO and LUMO, computed using the PM3 semiempirical quantum mechanics method with full geometry optimization,[13] were also obtained from a prior study.[14] For those interested in the carcinogenic properties of the chemicals (not considered in the present paper), the revised carcinogenicity classification has been published.[15]

Molecular Connectivities. A total of 40 molecular connectivity descriptors was calculated with the program TSAR (Oxford Molecular Ltd., Oxford OX4 4GA, England, Version 3.0), by selecting all possible molecular connectivity options from the properties menu, with the exception of two electrotopological indices valid only for substituents and not for whole molecules. The parameters were Randić, Balaban, and Wiener topological indices; electrotopological indices summed over all atoms; six shape indices and molecular flexibility indices; χ and χV indices for atoms, bonds, path, cluster, path/cluster, and rings up to a path length of 6.

2-D Distances (Bond Deletion Metric on Chemical Graphs). The bond-deletion metric defined on the chemical graphs of molecules is defined in detail in an earlier reference.[9] Briefly, it is the sum of the number of bonds in two molecular structures minus twice the number of bonds in their largest (as defined by the number of bonds) common substructure. In this definition, a largest common substructure may, and often does, consist of more than one common structural fragment, i.e., the largest common substructure need not be connected.

EVA. The structures in the database were optimized with the semiempirical quantum mechanics program MOPAC (Version 6.0); force calculations were then carried out on the minimized structures. The normal coordinate frequencies were processed to produce an EVA descriptor for each molecule. This processing was carried out using an in-house program (Centre for Molecular Design, University of Portsmouth, U.K.) which projected the eigenvalues (normal coordinate frequencies) onto a bounded frequency scale of 0−4000 cm^{-1}. Each vibration on this scale had a Gaussian curve superimposed on it, so that where vibrations were close together the Gaussian would overlap to produce a summed "intensity". The width of the Gaussian employed is controlled by an adjustable parameter (σ) which, in this case, was set at 20 cm^{-1}. The resulting "spectrum" was sampled at intervals of 10 cm^{-1} to produce a descriptors consisting of 400 "intensities" for each compound. For this work, only 91 intensities in the fingerprint range 1500−600 cm^{-1} were considered. Further details on the computation of EVA descriptors and the effect of changes in the Gaussian width and sampling interval are given in the literature.[10,16]

RESULTS AND DISCUSSION

The aim of this paper is the comparison of the information content of IR spectra (fingerprint region) with that of other descriptors used in QSAR practice. The comparison was based on a set of 112 chemicals (Table 1), from diverse classes, selected from a carcinogenicity database for which we had already calculated several descriptors,[15] and for which IR spectra determined in the same experimental conditions (gas phase) were available. Even though the number of chemicals is limited, the generality of the comparison was assured by (a) the presence of diverse chemical functionalities and substructures and (b) the substantial orthogonality of the descriptors logP, MR, HOMO, and LUMO (see Table 2). The analysis was carried out at two different levels: (a) a global comparison of the information carried by the different sets of descriptors, by considering together the local and general information, and (b) isolation of large trends, and comparison of "principal properties".

Table 2. Nine Distance Matrices Among Chemicals (112 × 112) Were Calculated, According to the Nine Properties Indicated in the Table[a]

	IR	EVA	CON	CHEM	logP	MR	HOMO	LUMO	2D
IR	1.00								
EVA	−0.02	1.00							
CON	0.03	**0.66**	1.00						
CHEM	−0.04	0.47	0.50	1.00					
logP	−0.03	0.46	0.42	0.57	1.00				
MR	0.08	0.50	**0.67**	0.54	0.29	1.00			
HOMO	−0.05	0.14	0.09	0.59	0.08	0.14	1.00		
LUMO	−0.05	0.02	0.09	0.50	0.10	0.06	−0.04	1.00	
2D	−0.00	**0.65**	**0.66**	0.41	0.43	0.50	0.03	0.07	1.00

[a] Each matrix (actually, one corner) was then correlated with all the other matrices. The table reports the correlation coefficients: CON: distance matrix based on 40 connectivity variables; CHEM: distance matrix based on logP, MR, HOMO, LUMO; and 2-D: distance matrix calculated according to the bond deletion metric on chemical graphs.

Comparison of Descriptors: Global Analysis. For each set of descriptors, separately, we calculated a distance matrix among the 112 chemicals under study. This led to the generation of the following nine 112 × 112 distance matrices, based respectively on (I) IR spectra; (II) EVA descriptors; (III) molecular connectivities (CON); (IV) classical descriptors (logP, MR, HOMO, and LUMO together) (CHEM); (V) logP; (VI) MR; (VII) HOMO; (VIII) LUMO; and (IX) 2-D distances. The 2-D distances were generated as described in Methods. For the other sets of descriptors, we calculated the Euclidian distances after standardization of the variables (i.e. subtraction of the mean and division by the standard deviation). The distances based on the IR spectra and the EVA descriptors were calculated from the original data. The classical descriptors were considered both in combination (IV) and individually (V, VI, VII, VIII), because of the clear different mechanistic significance of each of them.

The next step was the calculation of the correlation coefficients among distance matrices. Since the distance matrices are symmetrical, we considered only one lower diagonal half for each of them ($n = 6216$). Table 2 reports the correlation coefficients. The inspection of the table shows that the pattern of relationships among chemicals based on IR spectra was uncorrelated with all the other patterns. Beside the obvious intercorrelations (CHEM with the individual classical descriptors), the strongest relationships were (a) those of EVA with connectivities and 2-D distances and (b) those of connectivities with EVA, MR, and 2-D distances. The table also shows that the classical descriptors in this database were not intercorrelated.

The Table 2 correlation matrix was subsequently analyzed with Principal Component Analysis (PCA), by extracting its eigenvectors. Table 3 reports the factor loadings. PC1 collected information from connectivities, CHEM, EVA, 2-D distances, MR, and a component of logP. The electronic aspects were loaded on PC2 and PC3, whereas another component of logP was loaded on PC5. On the contrary, IR was loaded alone on PC4, thus confirming the diverse character of its information content.

Comparison of Descriptors: Analysis of Principal Properties. The above analysis was based on the use of

Table 3. Factor Loadings of the Principal Components Extracted from Table 2[a]

	PC1	PC2	PC3	PC4	PC5
CON	**0.84**	−0.27	0.00	0.00	0.00
CHEM	**0.81**	**0.56**	0.00	0.00	0.00
EVA	**0.80**	−0.25	0.00	0.00	0.00
2D	**0.77**	−0.33	0.00	0.00	0.00
MR	**0.75**	0.00	0.00	0.00	−0.40
HOMO	0.31	**0.63**	**0.69**	0.00	0.00
LUMO	0.00	**0.57**	**−0.69**	0.30	0.00
IR	0.00	−0.25	0.00	**0.92**	0.00
logP	**0.65**	0.00	0.00	0.00	**0.70**
expl var	0.41	0.15	0.12	0.11	0.08

[a] Explained variability: expl var.

distance matrices that included both the general and local components of information. It should be remembered that a high correlation coefficient between two distance matrices points to a very close resemblance between the two data spaces, since it implies a resemblance at both the very detailed and the coarse-grain scales.[17] To check for weaker but still significant relationships we carried out a further analysis in which we considered only the general trends. To this aim, we first identified the "principal properties" of each set of descriptors by PCA and then compared them with each other.

The 91 IR points were reduced to 15 PCs (explained variability 0.20, 0.1 1, 0.08, 0.07, 0.05, 0.05, 0.04, 0.04, 0.04, 0.03, 0.03, 0.02, 0.02, 0.02, 0.02, respectively). The 91 "fingerprint" EVA points were summarized to six PCs (expl. var. 0.46, 0.11, 0.09, 0.05, 0.04, 0.03). The 40 connectivity variables were summarized by five PCs (expl. var. 0.53, 0.13, 0.11, 0.06, and 0.05). The 2-D distance matrix was reduced to three PCs (expl. var. 0.51, 0.35, and 0.05 respectively). HOMO, LUMO, logP, and MR were considered separately, since they were poorly interrelated.

The various "principal properties" were compared to each other by a further PCA. Table 4 gives the factor loadings of the principal properties on the resulting global PCs. Table 4 confirms the results of Table 2 and shows that the IR components were substantially uncorrelated with all the other parameters. The first three general PCs explained each about 10% of variance and collected mainly the information from connectivities, 2D distances, and EVA descriptors, thus pointing to their high degree of overlapping (and redundancy). Except IR1, all the other IR components showed a rather unique information content. The high contribution of MR to PC1 should be noted; this permits the identification of PC1 as a size or steric component.[18] PC1 also collected the first component of connectivities, 2D distances, and EVA, whereas IR is devoid of any size-related component.

CONCLUSION

Overall, the analyses presented in this paper indicated that the complex information carried by IR spectra (fingerprint region) is markedly different from that codified in various molecular descriptors presently in use in QSAR practice.

Table 4. Relationships Pattern Among Principal Properties: Sorted Factor Loadings (See Details in the Text)[a]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
CO1	**0.94**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.00**	0.00	0.00	0.00	0.00
2D1	**0.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MR	**0.88**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV1	**0.88**	0.26	−0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
logP	**0.55**	0.00	0.00	0.00	0.00	0.43	0.00	−0.29	0.31	0.00	0.00	0.00	0.00	0.00
2D2	0.00	**0.82**	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV2	0.00	**0.61**	−0.29	−0.47	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CO3	0.00	**0.60**	0.00	0.00	−0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CO4	0.00	**0.56**	0.00	**0.55**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV3	0.00	−0.53	0.00	−0.27	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2D3	0.00	−0.40	**0.70**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CO2	0.00	0.00	**0.68**	0.00	0.39	0.00	0.00	−0.27	0.00	0.00	0.00	0.00	0.00	0.00
IR1	0.00	0.00	**−0.56**	0.00	0.00	−0.46	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV4	0.00	0.00	0.00	**0.62**	0.00	−0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV5	0.00	0.00	0.30	0.00	0.00	−0.50	0.00	0.26	0.38	0.00	0.00	0.00	0.00	0.00
CO5	0.00	0.00	0.00	0.00	0.34	0.41	**0.62**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IR8	0.00	0.00	0.29	0.35	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00
IR13	0.00	−0.28	0.00	0.00	−0.30	0.00	0.00	0.00	0.00	0.00	0.00	**0.60**	0.00	0.00
IR5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	−0.26	0.00	0.00	0.00	**0.66**	−0.32
IR6	0.00	0.00	0.00	0.00	0.00	0.33	0.38	0.00	0.00	0.00	0.00	0.00	0.53	0.00
IR10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.53	0.00	0.00	**0.59**
IR11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	−0.37	0.47	−0.26	0.00	0.00	0.47
IR15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.33	0.41	0.00	−0.32
IR3	0.00	−0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.00	−0.26	0.00	0.00	0.25
IR4	0.29	0.00	0.00	0.35	0.00	0.00	0.00	−0.40	0.00	0.00	0.00	0.00	0.00	0.00
IR12	0.00	0.00	0.00	0.00	−0.43	0.00	0.31	0.00	0.00	0.38	0.00	0.00	0.00	0.00
IR14	0.00	0.00	0.00	−0.26	0.00	0.00	0.00	0.29	0.00	0.00	−0.41	0.49	0.00	0.00
IR2	−0.35	0.31	0.00	0.26	0.00	0.00	−0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EV6	0.00	0.26	0.34	0.00	0.36	0.33	−0.25	0.36	0.00	0.00	0.00	0.00	0.00	0.00
IR7	0.00	0.00	0.00	0.00	0.44	0.00	0.32	0.00	−0.29	0.00	0.00	0.00	0.00	0.00
IR9	0.00	0.00	−0.26	0.00	0.00	0.32	−0.33	0.00	0.28	0.00	0.00	−0.29	0.00	0.00
LUMO	−0.36	0.00	−0.46	0.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HOMO	0.35	0.00	0.00	−0.42	0.00	0.00	0.00	−0.46	0.00	0.00	0.00	0.00	0.00	0.00

[a] IR1 to IR15: principal components extracted from the IR spectra; EV1 to EV6: principal components extracted from the EVA descriptor; 2D1 to 2D3: principal components of the 2D-distance matrix; and CO1 to CO5: principal components obtained from 40 connectivity variables. Explained variability of the global principal components: 0.14, 0.09, 0.07, 0.06, 0.06, 0.05, 0.04, 0.04, 0.04, 0.03, 0.03, 0.03, 0.03, and 0.02.

Much redundancy and overlapping was found among descriptors such as connectivities, 2-D distances, and theoretical spectral EVA descriptors. Their correlation with MR showed that they are very sensitive to the size of the molecules. On the contrary, IR did not show this size component that seem to be shared by most of the other descriptors.

The collinearities among descriptors is a topic central to the QSAR research. As discussed by Franke,[18] such relations can be distinguished into three categories: (a) relations that can be attributed to well-known systematic features (as is the case between logP and CSA), i.e., where the variables involved are systematically interchangeable; (b) relations that express as yet unrecognized systematics or are valid only within a particular series—this category may be very problematic, since no conclusions about the relevant parameters can be drawn; and (c) relations that exist solely in certain samples of chemicals having a particular composition. If the latter is present without being recognized, a QSAR analysis can be entirely useless.

At present, several methods (e.g., PCA, PLS) that deal with the collinearity problems are available. However, the use of a limited set of individual parameters with clear mechanistic significance is still the approach that ensures the optimal comprehension of the results and gives the possibility of performing nonformal validations much superior to those provided by statistics.[1,18] For the above reasons, when addressing the study of new descriptors, it is essential to analyze their relationship with other types of descriptors whose significance has been already assessed. Regarding the IR spectra, we found in the literature a limited number of studies focusing on individual IR peaks in certain chemical classes. In the present work, we did not consider individual IR peaks but focused on a fingerprint region of the whole IR spectrum as a global chemical descriptor. The unique information content of the IR spectra demonstrated in this paper makes the continuation of this research very attractive, in particular for possible use in QSAR analyses.

From a more fundamental standpoint, we are interested in knowing how and whether the unique chemical information provided by the IR spectra has meaning in relation to biological activity. The discovery of a relation between IR and a certain biological end-point, at odds with the purely formal descriptors like topological indices, would permit the sketching of mechanistic hypotheses based on the well-known physical chemical meaning of IR spectra. From a technical point of view, the reduction of the spectra to a limited number of quantitative variables is feasible: for example, in this paper we have used PCA to summarize the information. In our laboratory, we plan to continue this research by investigating if, and in which cases, the IR spectra-derived descriptors can be useful for QSAR analysis.

REFERENCES AND NOTES

(1) Hansch, C., Leo, A. *Exploring QSAR. 1. Fundamentals and applications in chemistry and biology*; American Chemical Society: Washington, DC, 1995.

(2) Livingstone, D. J. Quantitative Structure-Activity Relationships. In *Similarity models in organic chemistry, biochemistry and related fields*; Zalewski, R. I., Krygowski, T. M., Shorter, J. Eds.; Elsevier: Amsterdam, 1991; pp 557−627.

(3) Albrecht-Buehler, G. In defense of “Nonmolecular” cell biology. *Int Rev. Cytol.* **1990**, *120*, 191−241.

(4) Silverstein, R. M.; Bassler, G. C. *Spectrometric identification of organic compounds*, 2nd ed.; Wiley International Edition: New York, 1968.

(5) Kovac, S.; Kristian, P.; Antos, K. Studies of the vibrational frequencies ni(asym.NCS) of m- and p-substituted phenylisothiocyanates by various solvents. *Collect. Czechoslov. Chem. Commun.* **1965**, 3664−3671.

(6) Wayland, B, B.; Drago, R. S. Determination of the donor sites in Lewis acid adducts of anisole and thioanisole. *J. Am. Chem. Soc.* **1964**, *86*, 5240−5244.

(7) Sotomatsu, T.; Nakagawa, Y.; Fujita T. Quantitative structure-activity studies of Benzoylphenylurea larvicides. *Pestic. Biochem. Physiol.* **1987**, *27*, 156−164.

(8) O’Sullivan, D. G.; Sadler, P. W. An example of correlation between Taft’s sigma* parameters and Infrared spectral frequencies. *J. Chem. Soc.* **1957**, 4144−4146.

(9) Johnson, M. A. Structure-Activity maps for visualizing the graph variables arising in drug design. *J. Biopharm. Statist.* **1993**, *3*, 203−236.

(10) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA - a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Design* **1997**, *11*, 143−152.

(11) Anonymous *The Aldrich library of FT-IR spectra*, 1st ed.; Aldrich Chemical Company, Inc.: Milwaukee,WI, 1985.

(12) Benigni, R.; Andreoli, C. Rodent carcinogenicity and toxicity, in vitro mutagenicity, and their physical chemical determinants. *Mutat. Res.* **1993**, *297*, 281−292.

(13) Stewart, J. J. P. Optimization of parameters for semiempirical methods II Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.

(14) Benigni, R.; Andreoli, C.; Cotta-Ramusino, M.; Giorgi, G.; Gallo, G. The electronic properties of carcinogens, and their role in SAR studies of noncongeneric chemicals. *Toxicol. Modeling* **1995**, *1*, 157−167.

(15) Benigni, R.; Richard, A. M. QSARs of mutagens and carcinogens: two case studies illustrating problems in the construction of models for noncongeneric chemicals. *Mutat. Res.* **1996**, *371*, 29−46.

(16) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Design* **1997**, *11*, 409−422.

(17) Sokal, R. R. Clustering and classification: background and current directions. In *Classification and clustering*; Van Ryzin, J., Ed.; Academic Press: 1977; pp 1−44.

(18) Franke, R. *Theoretical drug design methods*; Elsevier: Amsterdam, 1984.

CI980223X