Infrared Spectra as Chemical Descriptors for QSAR Models

Romualdo Benigni,* Alessandro Giuliani, and Laura Passerini

Istituto Superiore di Sanita', Laboratory of Comparative Toxicology and Ecotoxicology,
Viale Regina Elena 299, 00161 Rome, Italy

Received September 30, 2000

Infrared spectra (IR) were used as regressors for a number of QSARs and compared with both mechanistically oriented descriptors and heuristic "chemically neutral" descriptors (modified adjacency matrices eigenvalues). IR spectra usually gave results inferior to those obtained with the mechanistically driven descriptors, with one notable exception, and comparable to those obtained by adjacency matrices eigenvalues. So the IR spectra cannot be considered as an "a-priori" optimal description of molecules for QSAR. However the relation of IR with the chemicophysical bases of drug-receptor interaction suggests the use of IR spectra for elucidating mechanistic details.

INTRODUCTION

In a previous paper,[1] we compared the information carried—as a whole—by the fingerprint region of infrared (IR) spectra with the information carried by a number of families of chemical descriptors used in QSAR research. We thought that, in principle, the IR spectra may 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. molecular recognition),[2] and they are extremely specific fingerprints of the molecules. Our aim was to contribute to the exploration of one of the main components of QSAR research, i.e., the identification and development of chemical descriptors relevant to the chemical or biological activity of interest. As shown in a comprehensive, recent review by Livingstone on the characterization of chemical structures, "despite 130 years of research into the experimental and theoretical investigation of chemical structure and physicochemical properties, there is still no agreement on what constitutes the *best* set of descriptors for molecular design".[3] The problem of finding the absolute best set of descriptors, while patently absurd on mechanistic bases (each singular end-point is driven by different chemico-physical interactions) has an obvious practical appeal when dealing with very large combinatorial problems such as in the exploration of large databases or in high troughput screening.

Our previous investigation[1] consisted in the calculation of several sets of descriptors (classical physical chemical and quantum mechanical properties; molecular connectivities; 2-D molecular distances; the EVA infrared range vibration based theoretical descriptor) for a highly noncongeneric set of molecules; these descriptors were compared with the IR spectra of the same molecules. Much redundancy and overlapping was found among descriptors such as connectivities, 2-D distances, and theoretical EVA descriptors. On the contrary, the IR spectra appeared to carry quite specific information, markedly different from that of the other families of descriptors, consistently with the fact that they pertain to different ranges of energies with respect to other descriptors.

Based on these promising results, we performed a second investigation, to assess the usefulness of IR spectra as a basis for QSAR models of individual chemical classes. We selected from the literature a number of classes of congeneric chemicals, for which other authors had performed QSAR analyses. These sets were selected because of the following: (a) the IR spectra for the molecules were available in the literature and (b) the sets represented different biological activities/mechanisms of action/chemical descriptors. Moreover some of the reported original models were suboptimal with only a moderate predictive power so giving some room for a possible improvement by means of an alternative viewpoint as the IR spectra.

The IR spectra were used to perform QSAR analyses; the results were compared with those of the original investigations. To put these results in a larger perspective, we also derived QSAR models based on the eigenvalues from the modified adjacency matrices[4] (BME), which is a theoretical (as opposed to the experimental IR values) and relatively "unbiased" representation of the chemical structures. This strategy allowed us to compare three basic approaches: (1) classical, mechanistically oriented QSAR approach (original data); (2) use of an alternative viewpoint to QSAR by the use of a description based on a completely different range of energies with respect to classical QSAR (IR spectra); and (3) general purpose, not mechanistically oriented description (BME models).

DATA AND METHODS

Chemical Classes. Table 1 shows the chemical sets for which QSAR models were recalculated.

IR Spectra. The spectra, all derived in gas phase, in the absorption mode, were retrieved from the Aldrich compilation.[5] The printed spectra were scanned with an Epson GT-6500 scanner. With the program Spectrum (Delta Sistemi, Roma), written ad 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−

* Corresponding author phone: +39 06 49902579; e-mail: rbenigni@iss.it.

10.1021/ci000308+ CCC: $20.00 © 2001 American Chemical Society
Published on Web 03/31/2001

Table 1. Essential Information on the Original QSARs Studies, for Which We Recalculated the Models by Using IR and BME

ref	chemicals	end-point	original descriptors
12	nitriles	oxidative metabolism (ethanol accumulation) (in vitro)	polarizability
	nitriles	oxidative metabolism (inhibition) (in vitro)	polarizability, HOMO, LUMO
	nitriles	LD50 (mouse)	polarizability
13	organochlorine	LC50 (aquatic) (fish)	log K_{ow}, log K_{nbp}
	organochlorine	electrophilicity (log K_{NBP}) (in vitro)	ΔH, ΔS
14	*m*-anilines *p*-phenols	growth inhibition (*T. pyriformis*)	electronic, steric, partition
		biodegradability (bacteria)	electronic, steric
15	chlorobenzenes	LC50 (flounder)	LogP, connectivity
	chlorobenzenes	LC50 (sole)	LogP, connectivity
16	noncongeneric	allergic dermatitis (human and animal)	LogP, MR, hydrogen bond substructures
17	oestrogens	binding (α-receptor)	CoMFA
	oestrogens	binding (*β*-receptor)	CoMFA
18	aryl- and benzylhalides	LC50 (*Daphnia magna*)	LogP

Table 2. Comparison between the Original Results and the IR and BME Models[a]

			original		IR		BME	
chemicals	end-point	*n*	r^2	q^2	r^2	q^2	r^2	q^2
nitriles	EtOH accumulation	16	0.74	0.69	0.53	<0.5	0.68	0.63
	inhibition	11	0.51	<0.5	0.71	0.68	0.30	<0.5
	LD50	14	0.82	0.78	0.68	0.65	0.52	<0.5
organochlorine	LC50	8	0.77	<0.5	0.90	<0.5	0.90	<0.5
	log K_{NBP}	10	0.80	0.66	0.77	0.63	0.30	<0.5
m-anilines +	growth inhibition	15			0.77	0.71 (13)	0.29	<0.5
p-phenols	biodegradability	13	0.96	0.94	0.40	<0.5	0.66	0.61
chlorobenzene	LC50 flounder	6	0.90	0.69	0.81	0.62	0.93	0.71
	LC50 sole	6	0.88	0.58	0.81	0.57	0.93	0.62
noncongeners	allergy	71	0.68	0.55	0.44	<0.5 (25)	0.47	<0.5
oestrogens	α-receptor	31	0.95	0.70	0.49	<0.5 (29)	0.72	0.53
	β-receptor	31	0.95	0.60	0.56	<0.5 (29)	0.78	0.55
aryl- and benzylhalides	LC50	22	0.66	<0.5	0.21	<0.5	0.55	<0.5

[a] N = number of compounds (in parentheses the number of chemicals considered for IR models is given when different from the original data set). The value of q^2 not reaching the statistical significance was simply indicated as < 0.5. The paucity of some of the considered data sets makes only the pure descriptive (r^2 based instead of q^2) comparison possible. The growth inhibition induced by *m*-anilines and *p*-phenols was the only situation for which the authors[14] do not report any significant model.

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.

Eigenvalues of the Modified Adjacency Matrix (BME). The BME were calculated according to Burden.[4] In brief, each molecule was represented by an adjacency matrix. The off-diagonal elements, representing bonding connections, were assigned values 1, 2, and 3 for single, double, and triple bonds. To take into account the electronic environment of the atoms, the diagonal elements were roughly proportional to the electronegativity of the atoms. The value of carbon was set to zero; the values for hydrogen, nitrogen, and oxygen were set to 0.15, 0.9, and 0.9, respectively. In the original Burden's paper, all halogens were set to 2.3, since they represented minor structural modifications in the database studied. In our study, we differentiated the halogens by setting the diagonal values at 2.3, 0.9, 0.8, and 0.5 for fluorine, chlorine, bromine, and iodine, respectively. We also set the values for sulfur and phosphorus to 0.5. In our implementation, we followed both an electronegativity scale (based on the Burden's original setting), and we made some empirical tests (results not shown). For BME descriptors, see also our paper.[6]

RESULTS AND DISCUSSION

Table 1 shows the sets of chemicals selected for studying the suitability of IR spectra information as the basis for QSAR modeling. The chemical classes are quite diverse as well as the types of activity (biological and physicochemical). Most of the sets are congeneric classes in a strict sense; two sets are noncongeneric (oestrogens and fragrance allergens).

To construct the QSAR models, we treated the IR spectra as in ref 1. The IR spectra were sampled between 1500 and 600 cm^{-1} at 10 cm^{-1} intervals. Each spectrum was normalized to its highest peak. The 91 sampled points (variables) of each set of chemicals (units) were analyzed with PCA. The extracted components were used as regressors in the derivation of the various QSAR models (stepwise multiple regression analysis). Table 2 shows the fitting of the QSAR models recalculated based on the IR spectra together with the fitting of the original models. Table 3 reports the number

Table 3. Number of Principal Components Extracted for Both IR and BME Together with the Number of Components Entered into the Models

		IR PCs		BME PCs	
chemicals	end-point	extracted	used	extracted	used
nitriles	EtOH accumulation	15	3	6	1
	inhibition	15	1	6	3
	LD50	15	3	6	3
organochlorine	LC50	7	3	7	−
	log K_{NBP}	7	3	7	−
m-anilines + *p*-phenols	growth inhibition	11	3	7	1
	biodegradability	11	1	7	1
chlorobenzene	LC50 flounder	12	3	12	4
	LC50 sole	12	3	12	4
noncongeners	allergy	24	3	14	−
oestrogens	α-receptor	12	4	8	3
	β-receptor	12	4	8	3
aryl- and benzylhalides	LC50	6	−	6	3

of extracted components with the number of components selected by the stepwise regression.

To widen the perspective of this work, we recalculated the QSAR models also based on a different type of descriptor (Burden's eigenvalues from the modified adjacency matrices, BME).[4] The choice of BME was dictated by the quite "opposite" character of this descriptor with respect to both IR and the original descriptors. BME is not an experimental determination, like IR. Moreover, BME is a descriptor of chemical structure with very little overimposed chemical theory, since it is an eigenvalue decomposition of the modified adjacency matrix: in this sense BME is different from the mechanistically oriented descriptors used in the original papers. The eigenvalues relative to each chemical in a set were subjected to PCA: the resulting PCs were used in a stepwise regression versus the activity. Table 2 reports the fitting of the various models recalculated based on BME; Table 3 gives further details for each analysis.

The main results of this work are summarized in Table 2. The comparison between the original QSAR models and those IR-based shows that, except for the growth inhibition in *T. pyriformis*, the IR-based models are generally not as good as the original models. The same happens with the BME-based models: even more, there were a number of cases in which BME did not produce statistically significant QSAR models. A further check was performed on the value of the IR spectra for QSAR modeling. Since the three nitriles models in the original paper had limited goodness of fit, we checked if the IR information was complementary to the original descriptors for generating a better model. Only in the case of ethanol inhibition, one IR-derived PC contributed to increase the r^2 of the original QSAR of 0.20 units, whereas the other two models were not improved by IR.

In another case (LC50 of organochlorines), the apparent increase in r^2 linked to the use of IR was demonstrated to be unstable by the parallel analysis of q^2 parameter.

Thus, the overall conclusion is as follows: (a) IR information provided a basis for generating QSAR models; (b) however, the selection of descriptors more related to the mechanism of action generally gave better QSAR models than IR.

Conclusion (a) is in agreement with the results of Bursi et al.[7] that showed that the IR spectra were a suitable basis for generating a QSAR model for a group of 45 progestagens (binding affinity) (however, with an efficiency lower than other descriptors, including the calculated IR). The present work shows that IR can replace other descriptors to a certain extent, despite our previous observation[1] of the general independence between IR and other families of descriptors. The presence of two opposite behaviors (like the lack of correlation on the large (noncongeneric set) scale and the substantial correlation on the local scale of congeneric series) is not unusual, as we previously demonstrated.[8] As a matter of fact, this is a central issue of the investigations on the molecular descriptors, where conceptual or statistically based relationships among descriptors may vary and have to be validated class by class.[3,9] In any case this result points to the fact that IR spectra are not general purpose descriptors but specific, mechanistically oriented ones, that can be important when the key factor of the activity modulation is driven by the range of interactions relevant for IR (molecular recognition, receptor dynamics.).

Regarding the Conclusion (b), the superiority of the QSAR models based on specific hypotheses on the mechanisms of action of the various classes in respect to the "automatic" approaches (see also the BME results in Table 2) further stresses the contribution that sound scientific reasoning gives to QSAR modeling.[10] This result can be considered obvious, but nevertheless it is important to stress the added value of accurate, mechanistically oriented QSAR when the research efforts in structure−activity relations are more and more oriented to pure chemoinformatics approaches (e.g. fingerprints).[3] As a final consideration, the issue of the specificity of IR for QSAR modeling (i.e. are there specific situations in which the IR spectra are the "best" descriptors?) was not solved by this work. We were able to find only one case (growth inhibition in *T. pyriformis*) in which the IR spectra permitted the construction of a QSAR model, where other descriptors failed, and two other cases (Inhibition of oxidative metabolism in nitriles, and LC50 in organochlorines) (Table 2) in which the IR-based QSARs were marginally superior to the original ones. However, this is too little evidence for generalizing. A better strategy would be of considering cases in which other families of descriptors failed to generate QSAR models. Unfortunately these unsuccessful stories are buried in the archives of companies and research institutes, and we were not able to retrieve them.

On a more theoretical ground, IR is expected to find its role as QSAR descriptor in the cases where the vibrational and oscillatory states are crucial for the drug-receptor interaction. This kind of situation is difficult to imagine, since the general consistency between vibrational (and oscillatory) states of a drug and its receptor is general assumed to be an a priori requirement for a QSAR investigation. Other determinants (partition, electronic characteristics, etc.) are more likely to be the rate-limiting factors in modulating the efficiency of the drug/receptor interaction. However if adequate data were available, the discovery of a specific IR-based QSAR could contribute to the elucidation of the chemical-physical bases of the drug action[11] and not only of its modulation. On the other hand, the present results indicate that the idea of considering IR as an optimal all-purpose molecular description for QSAR is not viable.

ACKNOWLEDGMENT

The skillful technical assistance of Mrs. Celestina D'Ascoli is gratefully acknowledged.

REFERENCES AND NOTES

(1) Benigni, R.; Passerini, L.; Livingstone, D. J.; Johnson, M. A.; Giuliani, A. Infrared spectra information and their correlation with QSAR descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 558−562.

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

(3) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(4) Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.−Act. Relat.* **1997**, *16*, 309−314.

(5) Anonymous. *The Aldrich library of FT-IR spectra*; Aldrich Chemical Co., Inc.: Milwaukee, WI, 1985.

(6) Benigni, R.; Passerini, L.; Pino, A.; Giuliani, A. The information content of the eigenvalues from modified adjacency matrices: large scale and small scale correlations. *Quant. Struct.−Act. Relat.* **1999**, *18*, 449−455.

(7) Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−867.

(8) Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani A. On the equivalence between different descriptions of molecules: value for computational approaches. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 575−578.

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

(10) Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: Toward a deeper understanding of chemico-biological interactions. *Chem. Rev.* **1996**, *96*, 1045−1075.

(11) Kohen, A.; Klinman, J. P. Hydrogen tunneling in biology. *Chem. Biol.* **1999**, *6*, R191−R198.

(12) Lewis, D. F. V.; Ioannides, C.; Parke, D. V. Interaction of a series of nitriles with the alcohol-inducible isophorm of P-450: computer analysis of structure−activity relationships. *Xenobiotica* **1994**, *24*, 401−408.

(13) Verhaar, H. J. M.; Rorije, E.; Borkent, H.; Seinen, W.; Hermens, J. L. M. Modeling the nucleophilic reactivity of small organochlorine electrophiles: a mechanistically based quantitative structure−activity relationship. *Environ. Toxicol. Chem.* **1996**, *15*, 1011−1018.

(14) Damborsky, J.; Shultz, T. W. Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. *Chemosphere* **1997**, *34*, 429−446.

(15) Furay, V.; Smith, S. Toxicity and QSAR of chlorobenzenes in two species of benthic flatfish, flounder (Platichtys flesus L.) and sole (Solea solea L.). *Bull. Environ. Contam. Toxicol.* **1995**, *54*, 36−42.

(16) Magee, P. S.; Hostynek, J. J.; Maibach, H. I. A classification model for allergic contact dermatitis. *Quant. Struct.−Act. Relat.* **1994**, *13*, 22−33.

(17) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor Alfa and Beta subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(18) Marchini, S.; Passerini, L.; Hoglund, M. D.; Pino, A.; Nendza, M. The toxicity of Aryl- and Benzylhalides to *Daphnia magna* and classification of their mode of action based on QSARs. *Environ. Toxicol. Chem.* **1999**, *18*, 2759−2766.

CI000308+