Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities

Roberta Bursi,*,† Thuy Dao,‡ Theo van Wijk,‡ Marcel de Gooyer,§ Edwin Kellenbach,‡ and Paul Verwer⊥

Molecular Design & Informatics, Analytical Chemistry for Development, and Department of Pharmacology, N. V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands, and CAOS/CAMM Center, University of Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

Received April 15, 1999

A novel 3D QSAR approach, comparative spectra analysis (CoSA), in which molecular spectra are used as three-dimensional molecular descriptors for the prediction of biological activities, is presented and discussed. To this purpose, experimentally determined ^{1}H NMR, mass, and IR spectra, as well as simulated IR and ^{13}C NMR spectra, for a set of 45 diverse progestagens are converted by a program, SpecMat, into matrixes, which are subsequently employed in a multivariate regression analysis (PLS). The results are compared with those resulting from a comparative molecular field analysis (CoMFA). When used individually, spectral descriptors yield better correlations and predictions than molecular field descriptors. A combination of spectral descriptors with other descriptors, either spectral or molecular field in nature, leads in most cases to models that are statistically superior to the ones obtained by their corresponding individual spectral or molecular field descriptors.

INTRODUCTION

Since the pioneering work of Hammett,[1,2] considerable effort has been put into the development of appropriate molecular descriptors to predict physicochemical properties and/or biological activities of molecules. Comparative molecular field analysis (CoMFA)[3] paved the way for the use of three-dimensional (3D) molecular descriptors in quantitative structure−activity relationship (QSAR) studies.[4−6] Although this approach has in many cases proven to be useful, it suffers from a number of limitations such as the requirement of molecular superposition, a dependence of the statistical quantities on the grid-point distance, and the use of partial charges in representing electrostatic interactions. While ways have been found to deal with some of these limitations,[7−9] molecular superposition remains an *intrinsic* and problematic requirement of the technique, which can only be eliminated if the binding mode of a ligand with respect to the receptor, i.e., the alignment of the ligand with respect to the receptor and the conformation of the receptor-bound ligand, is experimentally known. This involves, in general, knowledge of the structure of the ligand−receptor complex, a condition that is unfortunately seldom fulfilled. In the majority of the cases, the theoretical derivation of alignments and ligand conformations is therefore a *conditio sine qua non* for QSAR studies. These derivations are generally time-consuming and inevitably affected by personal intuition.

The above considerations make clear that QSAR studies would benefit tremendously from the development of other types of molecular descriptors that are not subject to such constraints. Molecular spectra, such as for example NMR, IR, Raman, UV−vis, and mass spectra, would in this respect seem to be promising candidates: they are in principle a molecule-unique reflection of the structural and chemical parameters but are not subject to the restrictions inherent to CoMFA. More specifically, they are (1) *true fingerprints* of the 3D structure of molecules; (2) in contrast with molecular fields, they are *observables*, i.e., *measurable* properties; (3) in gas and solution phases, they are *invariant* to molecular orientation; and therefore, (4) they do *not* require *alignment rules*. Furthermore, in simulations, (5) *no* particular assumptions on a *scoring function* are needed and (6) knowledge of partial *charges* is not necessary. Recently reported (Q)SAR studies indeed show that such spectra can be used to ones advantage. In the work of Fesik et al.,[10,11] for example, experimentally determined 2D ^{1}H and ^{15}N heteronuclear single-quantum correlation (^{1}H and ^{15}N HSQC) spectra were employed to find and optimize the alignment and conformation of ligands toward various target proteins, while Ferguson et al.[12,13] performed QSAR studies on the basis of simulated IR spectra.

In the present work, we explore and take such an approach, which we will designate as Comparative Spectra Analysis (CoSA), several steps further. *In-house* experimentally determined mass, IR, and ^{1}H NMR spectra and simulated IR and ^{13}C NMR spectra have been employed as 3D spectral descriptors for the biological activities of an in-house synthesized congeneric data set of 45 structurally diverse progestagens (see Figure 1 and Table 1). To this purpose, a program, SpecMat, has been developed that transforms spectra into matrixes ready to be analyzed by multivariate regression analysis techniques (PLS)[14] in SYBYL.[15] Apart from using each of the various types of spectra separately

* To whom correspondence should be addressed. E-mail: r.buma@organon.oss.akzonobel.nl.
† Molecular Design & Informatics, N. V. Organon.
‡ Analytical Chemistry for Development, N. V. Organon.
§ Department of Pharmacology, N. V. Organon.
⊥ University of Nijmegen.

10.1021/ci990038z CCC: $18.00
Published on Web 08/19/1999

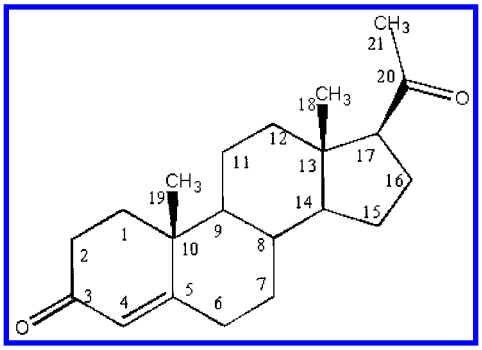


Figure 1. Two-dimensional structure of progesterone, **9**, and atom numbering.

as descriptors, analyses have been performed as well on the basis of more than one type of descriptor, be it a combination of spectral descriptors or a combination of spectral descriptors with molecular field descriptors. It will be shown that our results compare more than favorably with those obtained from CoMFA and argue for a further pursuit of CoSA.

COMPUTATIONAL DETAILS

SpecMat Program. In our studies, spectral data were obtained in several formats, e.g., as MOPAC[16] or Gaussian[17] output files or in the J-CAMP[18] format. A program, SpecMat, was developed to generate arrays from these spectral data. Each array line contains the name of a compound followed by its digitized spectrum, i.e., a list of intensities at given intervals. For each type of simulated and/or experimental spectrum (IR, NMR, ...), a separate array was generated. In all cases, data conversion was performed to present the data in a format that could be read in a SYBYL molecular spreadsheet and on which PLS analyses could be subsequently performed.

Simulated spectra are usually presented as a list of peak positions that may be accompanied by the corresponding intensities. At intervals of size L, the total intensity was calculated by adding the contributions from the individual peaks, assumed to have Gaussian shape and a width determined by σ (half-width at half-height). The total contribution of all peaks i, with intensity I_i centered at f_i, to a sample at position x was then taken as $\Sigma_i \exp\{(f_i - x)^2/2\sigma^2\}$. For simulated IR and ^{13}C NMR spectra, no intensities were provided. In those cases, all I_i's were set to 1. The sampling interval (L) determines the level of detail that is preserved in the digitized spectrum. In general, most information from the spectrum will be lost if L is taken to be larger than 2σ. Finally, the spectrum was normalized to obtain the same total intensity for each spectrum, independent of the size of the molecule.

Experimental mass spectra were obtained as a list of peak positions and corresponding intensities. In this case, the data can be treated in a way very similar to the treatment of simulated spectra. However, experimental ^{1}H NMR and IR data were supplied as sampled spectra, i.e., as a series of observed intensities at regular intervals. Data conversion for these two types of spectra is essentially a change of the sampling interval. To obtain the new samples at intervals L, the value of a sample at position x was taken to be the sum of all measured samples in the interval $x - L/2$ to $x + L/2$. If no samples were measured in this range, the value was obtained by linear interpolation between the nearest samples. Following this approach, both experimental IR spectra, which have relatively broad peaks originally sampled at intervals similar to those in the converted data, as well as ^{1}H NMR spectra, supplied as sharp peaks sampled at very small intervals, could be handled conveniently.

Simulation of the Spectra. IR spectra have been simulated using optimized geometries and force fields at the AM1 semiempirical level[19,20] employing the Gaussian94 suite of programs. The Gaussian outputs were used as input for SpecMat. ^{13}C NMR spectra were simulated using the ACD C NMR predictor software[21] version 3.0 and subsequently converted to the J-CAMP format. The simulation of the ^{13}C spectra was done by calculating the HOSE codes[22] of the carbons in the molecule and matching these against the HOSE codes of the ^{13}C nuclei in a database containing 50 000 compounds.

CoMFA and MIMIC. The AM1-optimized structures with AM1 partial charges obtained for the IR simulations were used for the analyses performed by the CoMFA technique. Standard CoMFA steric and electrostatic fields were used. Two types of alignments were considered for these CoMFA's. In the first alignment, a rigid superposition of the compounds, in which the C and D rings are optimally superimposed on each other, is made. The second is a field alignment in which superposition takes place by alignment of the steric and electrostatic molecular fields calculated on a 3D grid and described by Lennard-Jones and Coulomb potentials, respectively. The latter alignment was performed by means of the MIMIC[23] program. When molecular field and spectral descriptors or various spectral descriptors were combined, it was found that the standard CoMFA scaling option as opposed to the autoscale or no-scaling options could provide an optimal balance between them.

EXPERIMENTAL DETAILS

All compounds of the data set fulfilled the standard structural and purity requirements of Organon in view of subsequent meaningful pharmacological testing. Structures are routinely assessed by ^{1}H NMR (optionally ^{13}C and 2D NMR as well), IR, and mass data. Purities are routinely assessed by HPLC and ^{1}H NMR and are required to be higher than 95%. Binding affinities (Bind)[24] were used as biological activities all through this work.

Mass and gas-phase IR spectra were obtained on a Hewlett-Packard (HP) 5890 series II gas chromatograph equipped with an HP 5965A infrared detector and an HP5971A mass detector. A column and injector temperature of 300 °C was used, while the detector temperature was set at 280 °C. The resolution of the IR spectra, scanned from 550 to 3800 cm^{-1}, was 8 cm^{-1}. Mass and IR spectra were converted to the J-CAMP format using the library software and IR-processing package WINIR 3.0, respectively.

^{1}H NMR spectra were obtained on a BRUKER DRX400 NMR spectrometer operating at 400.13 MHz, using XWIN NMR software version 2.1. About 2 mg of material was dissolved in 0.7 mL of $CDCl_3$ containing a residual trace of $CHCl_3$ and tetramethylsilane. The NMR experiments were carried out at 25 °C. One hundred twenty-eight scans were obtained, acquiring 64K data points for a sweep width of 8000 Hz using a 60° flip angle. The data were Fourier transformed after applying an exponential multiplication of

Table 1. Two-Dimensional Structures and Binding Affinities

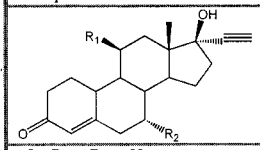
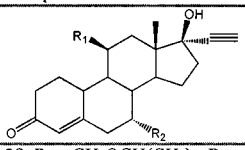
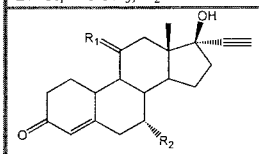
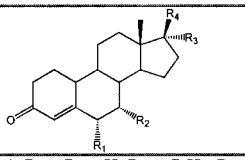
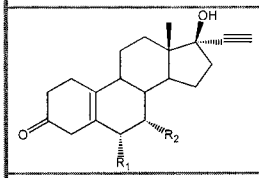
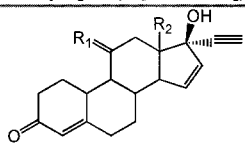
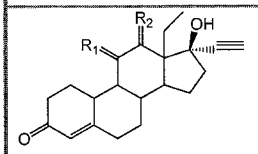
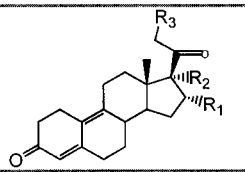
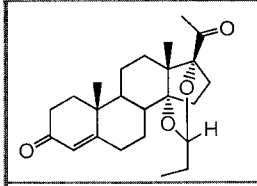
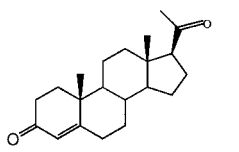
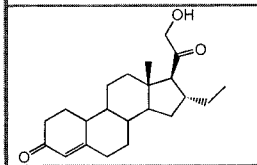
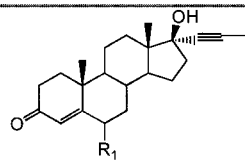
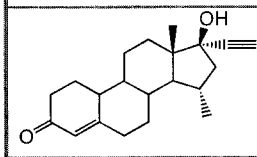
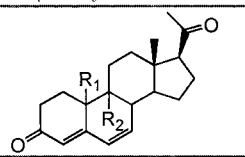
Compound	Log(Bind)	Compound	Log(Bind)
			
2 $R_1 = R_2 = H$	1.40	**28** $R_1 = CH_2OCH(CH_3)_2$, $R_2 = H$	0.00
11 $R_1 = CH_3$, $R_2 = H$	2.08	**37** $R_1 = CN$, $R_2 = H$	1.32
15 $R_1 = OCH_3$, $R_2 = H$	1.23	**38** $R_1 = CH{=}CH_2$, $R_2 = CH_3$	0.90
17 $R_1 = CH_2Cl$, $R_2 = H$	1.34	**39** $R_1 = CH_2OCH_3$, $R_2 = CH_3$	-0.52
18 $R_1 = C_2H_5$, $R_2 = H$	1.85	**40** $R_1 = CN$, $R_2 = CH_3$	0.70
19 $R_1 = CH{=}CH_2$, $R_2 = H$	1.60	**42** $R_1 = CHF_2$, $R_2 = CH_3$	1.11
20 $R_1 = CH{\equiv}CH$, $R_2 = H$	1.90	**44** $R_1 = C_2H_5$, $R_2 = CH_3$	1.30
22 $R_1 = CH_2OCH_3$, $R_2 = H$	0.0	**45** $R_1 = (E)CH{=}CHCl$, $R_2 = H$	1.48
24 Organon ownership	X	**49** $R_1 = H$, $R_2 = CH_3$	1.18
27 $R_1 = OCH_3$, $R_2 = H$	1.20		
			
14 $R_1 = CH_2$, $R_2 = H$	2.03	**1** $R_1 = R_2 = H$, $R_3 = C_2H_5$, $R_4 = OH$	1.51
26 $R_1 = CH_2$, $R_2 = CH_3$	1.48	**4** $R_1 = CH_3$, $R_2 = H$, $R_3 = C{\equiv}CH$, $R_4 = OH$	1.20
43 $R_1 = (E)CHF$, $R_2 = H$	1.53	**25** $R_1 = H$, $R_2 = CH_3$, $R_3 = CH_2CH{=}CH_2$, $R_4{=}OH$	1.43
		36 $R_1{=}R_2 = H$, $R_3 = CH_2CH(CH_3)_2$, $R_4{=}OH$	1.08
		41 $R_1{=}R_2 = H$, $R_3 = CH{=}CH_2$, $R_4{=}OH$	1.20
			
7 $R_1 = R_2 = H$	0.49	**29** $R_1 = H_2$, $R_2 = C_2H_5$	2.26
10 $R_1 = CH_3$, $R_2 = H$	-0.30	**30** $R_1 = CH_2$, $R_2 = CH_3$	2.15
47 $R_1 = H$, $R_2 = CH_3$	0.48	**31** $R_1 = H_2$, $R_2 = CH_3$	1.69
		32 $R_1 = CH_2$, $R_2 = C_2H_5$	2.30
			
13 $R_1{=}R_2{=}H_2$	1.90	**21** $R_1{=}R_3{=}H$, $R_2{=}OH$	1.56
16 $R_1{=}CH_2$, $R_2{=}H_2$	2.28	**23** $R_1{=}C_2H_5$, $R_2{=}H$, $R_3{=}OH$	2.11
34 $R_1{=}H_2$, $R_2{=}CH_2$	0.30		
			
35	1.61	**9**	1.15
			
8	2.0	**3** $R_1{=}\alpha CH_3$	0.78
			
33	1.47	**5** $R_1{=}\alpha CH_3$, $R_2{=}\beta H$	0.65

Table 2. Comparison of CoSA and CoMFA for a Set of 45 Progestagens

	q^2	s	no. of comp	r^2	s	F
CoMFA(rigid)	0.395	0.588	4	0.865	0.278	64.21
CoMFA (MIMIC)	0.391	0.582	3	0.841	0.299	72.01
IR SIM	0.533	0.516	4	0.987	0.086	762.71
IR EXP[a]	0.280	0.630	2	0.609	0.465	30.37
IR EXP[b]	0.437	0.587	5	0.827	0.326	33.44
mass EXP	0.484	0.536	3	0.924	0.204	166.96
^{1}H NMR EXP	0.548	0.515	5	0.969	0.135	243.60
^{13}C NMR SIM	0.395	0.613	5	0.987	0.088	587.45

[a] Forty-three molecules were considered here. The spectra of **32** and **25** could not be measured. [b] Forty-two molecules were considered here (**35** was treated as an outsider).

0.3 Hz and subsequently phase corrected to yield 32K real spectra. The 32K spectra turned out, however, to be too large to be handled efficiently in SpecMat. A comparison of the 32-, 16-, 8-, and 4K spectra led to the conclusion that 8K spectra still retain the most information, being at the same time much easier to handle. All analyses were therefore performed on 8K spectra. The peaks of the traces of $CHCl_3$ and tetramethylsilane were almost at the same positions in all spectra and were eliminated before the SpecMat conversion.

RESULTS

Individual Descriptors. For all CoSA analyses, an optimum was found between the spectral resolution and statistical performance. Only the final optimized CoSA results are discussed here. Overfitting was checked in all models by performing randomizations of the experimental activities. No "random" combination yielded statistics close to the correct one. The models obtained by means of CoSA and CoMFA on the whole set of 45 progestagens are given in Table 2. In this study, simulated (SIM) IR and ^{13}C NMR and experimental (EXP) IR, mass, and ^{1}H NMR spectra were considered, as well as CoMFA based on rigid and field alignments. Table 2 shows that the best statistics in terms of q^2 and s is provided by the SIM IR and the EXP ^{1}H NMR and mass spectra. In CoMFA, both alignments yield the same statistics, which are comparable to the SIM ^{13}C NMR spectra. A full analysis on 45 progestagens could not be performed on the EXP IR spectra because compounds **32** and **25** were unstable under GC conditions. The resulting analysis on 43 molecules was rather poor. This was caused in particular by compound **35**. No obvious experimental and/or computational reason could be found for this behavior.

The predictive power of these models has been assessed by splitting the data set in two subsets, a training set and a test set. For the test set, the most diverse compounds, i.e., compounds with unique substituents and/or features, from the complete data set were taken. This led to training and test sets of 38 and 7 compounds, respectively, the 7 compounds in the test set being **17**, **33**, **34**, **35**, **36**, **42**, and **43**. Statistics was repeated on the training set and the resulting models used to predict the activities of the test set. The results are given in Tables 3 and 4. A measure of the quality of a model in fitting (training set) and in prediction (test set) can be assessed by calculating the root mean square error of calibration (RMSEC) and root mean square error of prediction (RMSEP), respectively,[25] the latter values being reported in Table 4 as well. For all descriptors, the statistics performed on the training set improve with the exception of EXP ^{1}H NMR and SIM ^{13}C NMR. While EXP ^{1}H NMR remains unchanged, SIM ^{13}C NMR yields a model consisting of three components instead of five but with a smaller r^2 value. A difference between the two alignment rules in CoMFA was also observed, and the best final model was provided by the rigid superposition of molecules. This model was used for predictions of the test set.

Table 3. Comparison of CoSA and CoMFA for the Training Set

	q^2	s	no. of comp	r^2	s	F
CoMFA(rigid)	0.519	0.561	4	0.914	0.238	87.56
CoMFA (MIMIC)	0.550	0.535	3	0.871	0.286	77.01
IR SIM	0.638	0.481	3	0.986	0.096	771.91
IR EXP[a]	0.323	0.645	2	0.693	0.435	29.28
mass EXP	0.624	0.489	3	0.939	0.198	173.03
^{1}H NMR EXP	0.524	0.567	5	0.973	0.134	233.18
^{13}C NMR SIM	0.397	0.619	3	0.957	0.165	255.13

[a] The spectra of **32** and **25** could not be measured. A set of 36 progestagens is considered here.

In Table 4, the RMSEP values show that with the exceptions to IR EXP CoSA give rise to better predictions than CoMFA. Compound **34** is badly predicted by all descriptors except EXP ^{1}H NMR. EXP IR also provides good predictions but clearly has difficulty in predicting the activity of compound **35**. Since IR EXP data were not available for the whole data set of 45 molecules, this descriptor was not included further in our analysis.

Combined Descriptors. An interesting question that arises from the analyses based upon single descriptors is whether a combination of them can lead to improved statistics and a better predictive power. To answer this question, models have been obtained with various combinations of descriptors. The results for the whole set (45 compounds) are given in Table 5, where the contributions of the individual descriptors to the analyses are given as well. They show that the best models are obtained with the combinations (CoMFA, ^{1}H NMR), (IR, ^{1}H NMR), and (^{1}H NMR, ^{13}C NMR). Importantly, it is found that these combinations of descriptors lead to models with better statistics than those based upon the individual descriptors. Table 5 shows that spectral contributions are larger than CoMFA contributions, while among spectral descriptors themselves mass and ^{13}C spectra receive more weight than ^{1}H NMR and IR spectra.

Models based on the three combinations (CoMFA, ^{1}H NMR), (IR, ^{1}H NMR), and (^{1}H NMR, ^{13}C NMR) have also been derived for the training set of 38 molecules (Table 6). With these models, predictions have been made for the test set (Table 7). It can be concluded from Table 7 that all three analyses produce comparable results and, more importantly, yield better RMSEP values than the corresponding ones obtained from the separate descriptors. It is noticed that when ^{1}H NMR is combined with CoMFA, almost all predictions produced by CoMFA are improved. In particular, the prediction of the activity of compound **34** now becomes significantly closer to the experimental value. A few attempts have been made to combine CoMFA with more than one type of spectral descriptor and more than two spectral

Table 4. Predictions of the Test Set

compd	**17**	**33**	**34**	**35**	**36**	**42**	**43**	
exp log(Bind)	1.34	1.47	0.30	1.61	1.08	1.11	1.53	
CoMFA (rigid)	1.62	1.68	1.82	1.34	1.88	1.69	1.85	
$\|\Delta_i\|$[a]	0.28	0.21	1.52	0.27	0.80	0.58	0.32	0.72 RMSEP[b]
IR SIM	1.54	1.44	1.52	1.00	1.49	1.16	1.79	
$\|\Delta_i\|$	0.20	0.03	1.23	0.61	0.41	0.05	0.26	0.56 RMSEP[b]
IR EXP	1.38	1.37	2.06	−0.95	1.13	0.88	0.25	
$\|\Delta_i\|$	0.04	0.10	1.76	2.56	0.05	0.23	1.28	1.27 RMSEP[b]
mass EXP	1.43	1.73	1.69	1.16	1.20	1.47	1.58	
$\|\Delta_i\|$	0.09	0.26	1.39	0.45	0.12	0.36	0.05	0.58 RMSEP[b]
^{1}H NMR EXP	1.98	1.28	0.75	1.02	1.44	1.72	0.89	
$\|\Delta_i\|$	0.65	0.19	0.45	0.59	0.36	0.61	0.64	0.52 RMSEP[b]
^{13}C NMR SIM	1.10	1.74	1.59	1.25	1.47	1.43	2.24	
$\|\Delta_i\|$	0.24	0.27	1.29	0.36	0.39	0.32	0.71	0.62 RMSEP[b]

[a] $|\Delta_i|$ = residual absolute value. [b] RMSEP = $\sqrt{\sum_i^n \Delta_i^2/n}$ (root mean square error of prediction); $n = 7$.

Table 5. Descriptors Combination Analyses: Complete Set

$Descr_1$	$Descr_2$	q^2	s	no. of comp	r^2	s	$Descr_1$, %	$Descr_2$, %
CoMFA	mass	0.529	0.518	4	0.969	0.133	0.19/0.17[a]	0.64
CoMFA	IR	0.511	0.518	4	0.975	0.119	0.21/0.19	0.60
CoMFA	^{1}H NMR	0.640	0.453	4	0.974	0.122	0.22/0.18	0.60
CoMFA	^{13}C NMR	0.400	0.586	4	0.986	0.090	0.17/0.15	0.68
mass	IR	0.509	0.517	2	0.936	0.187	0.57	0.43
mass	^{1}H NMR	0.514	0.520	3	0.968	0.134	0.53	0.47
mass	^{13}C NMR	0.465	0.553	4	0.994	0.061	0.48	0.52
IR	^{1}H NMR	0.602	0.471	3	0.970	0.129	0.42	0.58
IR	^{13}C NMR	0.458	0.556	4	0.993	0.065	0.40	0.60
^{1}H NMR	^{13}C NMR	0.598	0.479	4	0.993	0.063	0.44	0.56

[a] Steric/electrostatic contributions.

Table 6. Descriptor Combination: Training Set

$Descr_1$	$Descr_2$	q^2	s	no. of comp	r^2	s	F
CoMFA	^{1}H NMR EXP	0.686	0.454	4	0.982	0.110	440.87
IR SIM	^{1}H NMR EXP	0.662	0.464	3	0.981	0.100	587.00
^{1}H NMR EXP	^{13}C NMR SIM	0.618	0.493	3	0.987	0.090	877.75

Table 7. Descriptor Combination: Test Set

compd	**17**	**33**	**34**	**35**	**36**	**42**	**43**	
Exp log(Bind)	1.34	1.47	0.30	1.61	1.08	1.11	1.53	
CoMFA + ^{1}H NMR	1.70	1.45	1.08	1.01	1.62	1.34	1.20	
$\|\Delta_i\|$[a]	0.36	0.02	0.78	0.60	0.54	0.23	0.33	0.47 RMSEP[b]
IR + ^{1}H NMR	1.84	1.35	1.12	1.06	1.46	1.36	1.08	
$\|\Delta_i\|$	0.50	0.12	0.82	0.55	0.38	0.25	0.45	0.49 RMSEP[b]
^{1}H NMR + ^{13}C NMR	1.54	1.61	1.31	1.12	1.42	1.40	1.37	
$\|\Delta_i\|$	0.20	0.14	1.01	0.49	0.34	0.29	0.16	0.47 RMSEP[b]

[a] $|\Delta_i|$ = residual absolute value. [b] RMSEP = $\sqrt{\sum_i^n \Delta_i^2/n}$, (root mean square error of prediction); $n = 7$.

descriptors with each other, but these combinations did not lead to statistically improved models.

DISCUSSION

The results presented above show that CoSA can be successfully used to describe the structure−activity relationships of molecules and that its performance is in many cases better than CoMFA: individually used, the ^{1}H NMR, mass, IR, and ^{13}C NMR spectra descriptors yield good correlations of the training sets and good predictions of the test sets. The performance of experimental IR spectra for this particular data set has been observed to be poor. In principle, experimental IR spectra are more informative than the simulated ones because they are characterized not only by fundamental transitions but also by overtones and combination bands. In this study, however, the measured spectral range was limited to 3800−550 cm^{-1}, and this might have caused some loss of information relevant for the analyses. Another possible explanation for the poor performance of these descriptors could be the experimental noise. In our opinion, however, these results do not preclude the application of experimental IR spectra in other studies.

Every time a new descriptor is found, designed, or simply used in an area other than the one of its origin—which justifies the adjective "new"—several questions need to be answered regarding its advantages and/or disadvantages, its performance, and its applicability range. The present study has been limited to a small, congeneric data set of steroids whose conformational flexibility is known to be limited, if not absent at all. Clearly, more studies need to be performed

in order to assess the generality and applicability range of these spectral descriptors. Although we therefore do not wish to claim that on the mere basis of the present study all of these issues can be addressed extensively, we do think that its results enable us to make a start in answering them.

For years now, great effort has been put in trying to develop molecular fingerprints that can be calculated very quickly for a large amount of compounds and that therefore enable fast database analyses. Some fingerprints have turned out to be better than others, but in general, they are all affected by limitations in terms of the 3D representation of molecules and, like molecular fields for example, are in general *artificial* representations of chemical entities—artificial in the sense that they cannot be measured and thus must be verified experimentally. A priori, one might say that CoSA can be considered as the opposite approach: spectra are *observables* and reflect directly intrinsic properties of molecules, but the 3D information included in these vectors—the extent determined by the type of spectrum considered—is not as transparent as in molecular fields. It is also important to notice that in many ways the various types of spectra contain complementary information: Mass spectra reflect the reactivity of a molecule in terms of molecular fragmentation, IR spectra vibrational motions and the spatial arrangements of molecular chemical groups, and ^{1}H and ^{13}C NMR the shape/spatial arrangement of the molecule. While current 3D QSAR methods (CoMFA and CoMFA-like approaches) require alignment rules because descriptors are calculated on a 3D grid around the molecules, CoSA does not require any alignment and therefore does not involve any assumption of this kind. Because of all these reasons, the amount of intrinsic bias present in CoSA is rather limited.

It might be argued that the use of experimental spectra has its drawbacks as well—though mainly from a practical point of view—considering that (i) measuring the spectrum requires the availability of a sample, (ii) experimental conditions such as the molecular stability in the gas phase and the presence of impurities and reference compounds in solution (to mention a few) may play a dominant role in a given project, and (iii) although the presently considered experiments are nowadays routinely performed, considerable human and technical resources are still needed. On the other hand, if measured spectra could be available on-line, data mining could easily be performed on this enormous amount of information.

The use of experimental spectra being in some aspects disadvantageous, it is reasonable to say that simulations might provide a worthwhile alternative. Simulations are not only time- and cost-effective, but they are also widely applicable because they do not require the physical presence of the molecules themselves. Moreover, simulated spectra are not subject to either noise or other experimental effects. IR[26−28] and ^{1}H NMR[29,30] spectra can nowadays be simulated with large accuracy by means of ab initio techniques and, in the case of IR, also by semiempirical approaches.[19,20] Their applicability to large databases is, however, prohibitively time-consuming. On the other hand, an empirical program has recently been developed that can simulate accurately IR spectra.[31] These simulations are very fast and can easily be performed on large data sets. Empirical packages to simulate ^{1}H NMR spectra are most of the time considered to be too inaccurate to be useful at all. Empirical ^{13}C NMR simulation packages are, on the contrary, successfully used: less than a second is needed to simulate a ^{13}C NMR spectrum with good accuracy. Finally, packages for simulating mass spectra have recently become available commercially as well.

A final point that should be addressed is the *display* of CoSA studies. In CoMFA, the 3D grid of points around the molecule provides a very simple, direct tool for visualization. At present, analyses based on CoSA can also be displayed in terms of favorable and unfavorable regions or peaks in the spectra, but this visualization is by far less transparent than the one available from CoMFA: one needs to perform spectral analyses, which implies spectroscopic notions or in any case notions about conformational analysis. It would be by far much clearer and of greater impact if CoSA could be translated automatically into the molecules instead of the spectra. Work along such lines is presently in progress.

CONCLUSIONS

Although considerable steps have been taken forward, describing and predicting the biological activity of molecules remains a hard challenge. Over the years, many and different descriptors have been developed to attempt to represent chemical structures and to establish their relationships with biological responses. In the present study, experimentally determined (^{1}H NMR, mass, and IR) and simulated (IR and ^{13}C NMR) molecular spectra have been employed as observable molecular fingerprints to predict the binding affinities of a diverse data set of steroids in a comparative spectra analysis (CoSA).

A program, SpecMat, was developed that converts spectra into matrixes, ready to be analyzed by multivariate regression techniques. The performance of CoSA was tested against the well-established 3D QSAR technique, CoMFA. Statistically, the CoSA results were good and compared favorably with CoMFA. The combination of various spectral descriptors and spectral descriptors with CoMFA turned out to be clearly advantageous, not only for fitting biological activities but also for predicting them.

In the present form, the first area of application of CoSA would seem to be QSAR and lead optimization. Possibilities for lead finding are present as well, although exploiting this direction to its fullest potential will require further and more extensive studies. Presently, further validation of CoSA is being performed by analyses on noncongeneric, more flexible data sets.

ACKNOWLEDGMENT

We are grateful to Dr. V. van Geerestein and Dr. J. Mestres (Molecular Design & Informatics, Organon) for helpful discussions and for providing MIMIC, respectively.

REFERENCES AND NOTES

(1) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17*, 7, 125−136.

(2) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(3) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(4) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology;* American Chemical Society*:* Washington, DC, 1995; Vol. 1.
(5) Kubinyi, H. *3D QSAR in Drug Design: Theory, Methods and Applications;* ESCOM: Leiden, 1993.
(6) Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in Drug Design: Recent Advances*; ESCOM (Kluwer): Leiden, 1997.
(7) Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.*−*Act. Relat.* **1993**, *12*, 9−20.
(8) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(9) Cho, S. J.; Tropsha, A. Crossvalidated r^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem*. **1995**, *38*, 1060−1066.
(10) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.
(11) Hajduk, P. J.; Sheppard, G.; Nettesheim, D. G.; Olejniczak, E. T.; Shuker, S. B.; Meadows, R. P.; Steinman, D. H.; Carrera, G. M., Jr.; Marcotte, P. A.; Severin, J.; Walter, K.; Smith, H.; Gubbins, E.; Simmer, R.; Holzman, T. F.; Morgan, D. W.; Davidsen, S. K.; Summers, J. B.; Fesik, S. W. Discovery of Potent Nonpeptide Inhibitors of Stromelysin Using SAR by NMR. *J. Am. Chem. Soc*. **1997**, *119*, 5818−5827.
(12) 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. Des*. **1997**, *11*, 143−152.
(13) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a Novel Range Vibration-Based Descriptor (EVA) for QSAR studies. 1. General Application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.
(14) Wold, S.; Albano, C.; Dunn, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. In *Chemometrics: Mathematics and Statistics in Chemistry*; Kowalski, B., Ed.; Dordrecht, The Netherlands, 1984.
(15) SYBYL 6.4; Tripos Assoc., 1699 S. Hanley Rd, St. Louis, MO 63144.
(16) Stewart, J. J. P. Mopac 6.0, Quantum Chemical Program Exchange 455, 1990.
(17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T. A.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Repogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. GAUSSIAN 94; Gaussian, Inc.: Pittsburgh, PA, 1995.
(18) McDonald, R. S.; Wilks, P. A., Jr. JCAMP-DX: a Standard Form for Exchange of Infrared Spectra in Computer Readable Form. *Appl. Spec.* **1988**, *42*, 151−162.
(19) Seeger, D. M.; Korzeniewski, C.; Kowalchyk, W. Evaluation of Vibrational Force Fields Derived by Using Semiempirical and ab Initio Methods. *J. Phys. Chem.* **1991**, *95*, 6871−6879.
(20) Coolidge, M. B.; Marlin, J. E.; Stewart; J. P. Calculations of Molecular Vibrations Frequencies Using Semiempirical Methods. *J. Comput. Chem*. **1991**, *12*, 948−952.
(21) ACD/CNMR Predictor. www.acdlabs.com/products/cnmr/cnmr_pred._ftr.html.
(22) Bremser, W. HOSE−a Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(23) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: a Molecular-Field Matching Program. Exploiting Applicability of Molecular Similarity Approaches. *J. Comput. Chem*. **1997**, *18*, 934−954.
(24) Berking, E. W.; van Meel, F.; Turpijn, E. W.; van der Vies, J. Binding of Progestagens to Receptor Proteins in MCF-7 Cells. *J. Steroid Biochem*. **1983**, *19*, 1563−1570.
(25) Esbensen, K. Multivariate Analysis−in practice. Wennbergs Trykkeri AS, Trondheim, 1996.
(26) Fogarasi, G.; Pulay, P. In *Vibrational Spectra and Structure*; Durig, J. R., Ed.; Elsevier: New York, 1985; Vol. 14, p 125.
(27) (a) Bursi, R.; Devlin, F. J.; Stephens, P. J. Vibrationally Induced Ring Currents? The Vibrational Circular Dichroism of Methyl Lactate. *J. Am. Chem. Soc*. **1990**, *112*, 9430−9432. (b) Bursi, R. Ab initio calculations of Vibrational Absorption and Vibrational Circular Dichroism. Ph.D. Thesis. University of Southern California, Los Angeles, 1991.
(28) Rauhut, G.; Pulay, P. Transferable Scaling Factors for Density Functional Derived Vibrational Force Fields. *J. Phys. Chem.* **1995**, *99*, 3093−3100.
(29) Kutzelnigg, W. Theory of Magnetic Susceptibilities and NMR Chemical Shifts in Terms of Localized Quantities. *Isr. J. Chem*. **1980**, *19*, 193−200.
(30) Schindler, M.; Kutzelnigg, W. Theory of Magnetic Susceptibilities and NMR Chemical Shifts in Terms of Localised Quantities. II. Application to Some Simple Molecules. *J. Chem. Phys.* **1982**, *76*, 1919−1933.
(31) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3D Structure of a Molecule in its IR Spectrum. *Fresenius J. Anal. Chem.* **1997**, *359*, 50−55.

CI990038Z