Molecular Similarity Matrices and Quantitative Structure–Activity Relationships: A Case Study with Methodological Implications

Romualdo Benigni,* Marina Cotta-Ramusino,[†] Fabrizio Giorgi,[‡] and Grazia Gallo[‡]

Laboratories of Toxicology and Ecotoxicology and of Pharmaceutical Chemistry, Istituto Superiore di Sanita', Rome, Italy, and Department of Chemical Research, Sigma-Tau, Pomezia, Italy

Received May 12, 1994[®]

Recently, statistical analysis of molecular similarity matrices has been applied to the quantitative structure–activity relationship (QSAR) analysis of a number of molecular series. This paper addresses a number of methodological issues relative to the similarity matrices. A series of halogenated aliphatic hydrocarbons, for which the mutation (aneuploidy) induction ability had previously been determined, was used as test bench. The chemical information carried by the similarity matrices was shown to overlap to a considerable extent the information carried by the classical descriptors (physical chemical and quantum mechanical parameters). A good QSAR was obtained on the basis of the similarity matrices, in analogy with that obtained with the classical descriptors; however, the similarity matrices neither complemented the classical descriptors nor were able to improve on their performance. The effect of the compound's spatial orientation on the similarity values was also investigated.

Introduction

The recognition of the critical roles that shape and, more generally, steric aspects have in molecule–receptor interaction has stimulated interest in methods suitable for comparing molecules in this respect. Whereas it is obvious that similar molecules are expected to exert similar activities, there is no rigorous or unambiguous method for defining and calculating their similarity. It may be expected that, in certain cases, the overall similarity will produce the similar activity, whereas, in other cases, only the similarity of certain (active) regions of the molecules will give rise to similar activities. Moreover, there is more than one similarity index, and there are different techniques with which to evaluate similarity and to model the properties whose spatial modulation is to be compared. These unanswered questions, together with the obvious importance of this subject, have stimulated much work in this field. Among other investigations, there has been the development of new methods for the estimation of the overall electrostatic and steric similarity of molecules.[1] More recently, $N \times N$ similarity matrices, by which each molecule is compared with all the other molecules under study, have been considered. The chemical information carried by the similarity matrices was compared with the biological activity through a partial least squares statistic, and good quantitative structure–activity relationships (QSARs) have been reported.[2,3]

In light of this, we started a study to further investigate the properties of the chemical similarity matrices. As test bench, we used a set of halogenated aliphatic hydrocarbons, for which we previously determined a range of physical chemical and quantum mechanical parameters, and defined a QSAR for the induction of aneuploidy in *Aspergillus nidulans*. Aneuploidy is a type of genetic mutation, which has severe effects on health; *A. nidulans* is a suitable test system for this genetic end point.[4,5] Additional molecular properties were determined specifically for this work. This study compares the chemical information carried by the similarity matrices with that carried by the physical chemical and quantum mechanical parameters and studies the usefulness of the chemical similarity information for the definition of the aneuploidy QSAR. Some methodological problems are also addressed.

Data and Methods

Table 1 reports the names, biological activity, and molecular properties of the chemicals studied. The following are the molecular descriptors and their codes: log P, logarithm of the octanol–water partition coefficient; MR, molar refractivity; bp, boiling point; d, density; I_x, I_y, and I_z, principal moments of inertia; R_x, R_y, and R_z, lengths of principal axes of inertia; EV, ellipsoidal volume, i.e., volume of inertial ellipsoid; dip., dipole moment; HOMO, energy of the highest occupied molecular orbital; LUMO, energy of the lowest unoccupied molecular orbital; dist, greatest bond length between a carbon and a halogen; charge, charge on the carbon relative to the longest carbon–halogen bond; varch, variance of the net charges of the atoms in the molecule.

The ab initio molecular orbital parameters were determined by a Gaussian92 program, with fully optimized geometries (STO-3G, Murtangh–Sargent option). Log P and MR were calculated according to ref 6. Bp and d were found in the literature. $I_x, I_y, I_z, R_x, R_y, R_z$, EV, and dip. were obtained with the program TSAR.[7] The biological activity data for compounds **1–41** were published in refs 4 and 5, where the experimental methods are also extensively presented. New biological data, for compounds **42–56**, were kindly provided by Dr. R. Crebelli (unpublished results).

For the similarity calculations, the molecules were built with the program INSIGHT II,[8] and optimized with CVFF force field. The molecular similarity indices were computed with the program ASP.[7] Both Carbo and Hodgkin methods were used for the calculation of similarities in terms of electrostatic potential, shape, lipophilicity, and refractivity. For an exhaustive description of the principles of the calculation methods, and many technical details, see ref 3. Briefly, the Carbo similarity index is the following:

$$S_{ab}^{\ C} = \frac{\int P_a P_b \, dV}{(\int P_a^{\ 2} \, dV)^{1/2} (P_b^{\ 2} \, dV)^{1/2}}$$

[†] Laboratory of Pharmaceutical Chemistry.
[‡] Department of Chemical Research.
[®] Abstract published in *Advance ACS Abstracts*, January 1, 1995.

Table 1. Properties and Activity of Halogenated Aliphatic Hydrocarbons[a]

no.	compound	aneu	log *P*	MR	bP	*d*	HOMO	LUMO	dist	charge	var_{ch}	I_x	I_y	I_z	R_x	R_y	R_z	EV	Dip
1	dichloromethane	+	1.26	16.335	39.9	1.325	−0.400	0.289	1.803	0.117	0.016	2.78	24.58	26.81	15.62	1.77	1.62	187.80	1.66
2	chloroform	+	1.98	21.151	61.0	1.492	−0.418	0.231	1.811	0.077	0.012	25.40	25.40	48.68	3.10	3.10	1.62	65.24	1.17
3	carbon tetrachloride	+	2.83	25.967	77.0	1.594	−0.437	0.174	1.818	0.039	0.008	48.68	48.68	48.68	1.76	1.76	1.76	22.88	0.00
4	1,1-dichloroethane	+	1.80	20.960	57.0	1.776	−0.486	0.292	1.820	0.141	0.017	12.91	25.64	36.05	4.90	2.47	1.75	88.98	2.01
5	1,2-dichloroethane	+	1.48	20.982	83.0	1.256	−0.380	0.288	1.809	0.150	0.014	2.95	55.07	56.94	28.68	1.53	1.48	273.50	0.00
6	1,1,1-trichloroethane	−	2.50	25.758	75.0	1.338	−0.411	0.232	1.821	0.096	0.016	35.28	35.28	49.19	3.24	3.24	2.32	101.90	1.94
7	1,1,2-trichloroethane	+	2.07	25.798	112.5	1.435	−0.393	0.259	1.818	0.123	0.014	23.24	57.33	77.51	5.34	2.16	1.60	77.36	1.37
8	1,1,1,2-tetrachloroethane	+	3.05	30.614	138.0	1.598	−0.407	0.217	1.822	0.082	0.013	44.21	76.11	81.34	3.70	2.15	2.01	66.78	1.80
9	1,1,2,2-tetrachloroethane	+	2.66	30.614	147.0	1.586	−0.408	0.235	1.816	0.108	0.013	49.20	64.42	110.90	3.38	2.58	1.50	54.76	0.00
10	pentachloroethane	−	5.72	35.430	161.5	1.680	−0.406	0.206	1.822	0.073	0.010	71.22	88.43	113.30	2.88	2.32	1.81	50.89	1.26
11	hexachloroethane	−	4.62	40.246	186.0	2.091	−0.418	0.189	1.820	0.045	0.007	96.26	120.90	120.90	2.51	2.00	2.00	41.89	0.00
12	1,1,2-trichloroethylene	+	0.96	25.317	86.9	1.464	−0.344	0.225	1.775	0.070	0.008	21.71	55.50	77.21	5.28	2.07	1.49	67.94	1.31
13	tetrachloroethylene	−	1.27	30.133	121.0	1.623	−0.348	0.201	1.769	0.041	0.004	52.53	58.63	111.20	2.28	2.04	1.08	20.99	0.00
14	1,2-dichloroethylene (cis + trans)	+	0.65	20.501	54.0	1.265	−0.338	0.251	1.772	0.114	0.011	1.63	55.04	56.67	51.38	1.52	1.48	484.20	0.00
15	1,1-dichloroethylene	+	1.21	20.501	31.0	1.213	−0.334	0.256	1.776	0.094	0.012	10.59	26.76	37.35	5.92	2.34	1.68	97.49	1.36
16	1,2-dichloropropane	−	2.02	25.607	95.5	1.156	−0.380	0.291	1.826	0.030	0.016	12.29	57.03	66.15	10.88	2.34	2.02	216.10	0.58
17	2,2-dichloropropane	−	2.34	25.585	67.0	1.082	−0.381	0.291	1.828	0.137	0.017	22.72	33.87	38.04	3.73	2.50	2.23	87.36	2.24
18	1,3-dichloropropane	−	1.74	25.629	121.0	1.192	−0.379	0.343	1.816	−0.056	0.014	5.91	92.00	96.32	26.57	1.71	1.63	309.70	1.98
19	1,2,3-trichloropropane	−	2.01	30.445	156.0	1.387	−0.389	0.273	1.819	0.030	0.015	23.13	96.63	110.90	8.78	2.10	1.83	141.30	3.18
20	1-chlorobutane	−	2.39	25.438	77.5	0.886	−0.368	0.358	1.812	−0.056	0.010	5.29	62.70	65.88	19.85	1.67	1.59	221.80	1.96
21	2-chlorobutane	−	2.57	25.416	69.0	0.873	−0.363	0.353	1.826	0.030	0.012	17.65	27.54	41.50	4.22	2.70	1.79	85.83	1.98
22	1,3-dichlorobutane	−	2.28	30.254	134.0	1.115	−0.373	0.338	1.824	0.030	0.014	18.33	88.30	101.80	12.82	2.66	2.31	330.20	2.17
23	2,3-dichlorobutane	−	2.56	30.232	118.0	1.107	−0.373	0.317	1.823	0.030	0.014	30.78	49.61	55.47	4.68	2.90	2.60	148.10	2.86
24	1-chloro-2-methylpropane	−	2.44	25.416	68.5	0.883	−0.368	0.360	1.814	−0.056	0.009	14.77	33.41	37.35	5.13	2.27	2.03	99.01	1.92
25	2-chloro-2-methylpropane	−	2.44	25.394	51.5	0.851	−0.362	0.349	1.833	0.106	0.015	18.77	27.07	27.07	3.20	2.22	2.22	65.81	2.05
26	1-chloropentane	−	3.11	30.085	107.5	0.882	−0.369	0.360	1.814	−0.054	0.008	5.89	105.50	108.80	29.67	1.66	1.61	331.50	1.98
27	1-chlorohexane	−	3.65	34.699	133.5	0.879	−0.368	0.360	1.813	−0.054	0.008	7.62	160.60	165.10	34.80	1.65	1.61	386.20	1.96
28	1-chlorooctane	−	4.85	43.979	183.0	0.875	−0.367	0.360	1.814	−0.054	0.008	9.86	323.90	329.50	53.34	1.62	1.60	579.60	1.96
29	1,2-dichloropropene	−	1.79	25.126	76.5	1.169	−0.327	0.262	1.780	0.070	0.011	21.24	35.74	56.45	4.62	2.75	1.74	92.41	1.49
30	2,3-dichloro-1-propene	+	1.75	25.148	94.0	1.204	−0.339	0.246	1.813	−0.058	0.013	14.54	49.19	57.33	8.46	2.50	2.14	189.90	1.88
31	1,3-dichloropropene (cis + trans)	−	1.19	25.148	112.0	1.181	−0.337	0.245	1.813	−0.058	0.012	6.73	74.55	80.74	22.01	1.98	1.83	335.50	1.32
32	1,1,3-trichloropropene	+	1.50	29.964	131.5	1.403	−0.345	0.220	1.811	−0.062	0.011	27.87	100.20	127.60	6.23	1.73	1.36	61.68	0.63
33	3-chloro-2-(chloromethyl)propene	−	2.01	29.795	138.0	1.080	−0.347	0.245	1.818	−0.063	0.013	21.79	69.04	85.21	9.16	2.89	2.34	259.60	2.54
34	1-chloro-2-methylpropene	−	1.89	24.935	68.0	0.920	−0.305	0.294	1.781	−0.040	0.013	9.92	36.68	45.56	8.32	2.25	1.81	142.30	1.65
35	3-chloro-2-methylpropene	−	1.89	23.929	75.0	0.917	−0.325	0.309	1.818	−0.061	0.012	9.67	37.14	44.03	7.79	2.03	1.71	113.20	2.00
36	chlorodibromofluoromethane	+	2.80	26.727	79.5	2.317	−0.346	0.215	1.939	0.060	0.012	44.39	76.46	98.36	3.95	2.29	1.78	67.75	0.68
37	bromoform	+	2.38	29.842	150.5	2.894	−0.323	0.255	1.929	0.021	0.011	66.36	66.36	130.10	3.36	3.36	1.71	80.98	1.01
38	bromochloromethane	+	1.40	19.232	68.0	1.991	−0.338	0.300	1.905	0.036	0.016	3.08	38.08	40.62	22.49	1.82	1.70	292.10	1.61
39	bromotrichloromethane	+	3.10	28.864	105.0	2.012	−0.377	0.185	1.923	0.132	0.009	48.68	72.40	72.40	2.68	1.80	1.80	36.52	0.30
40	bromodichloromethane	+	2.67	24.048	87.0	1.980	−0.359	0.237	1.911	0.089	0.013	25.44	45.29	68.49	4.47	2.51	1.66	77.85	1.14
41	chlorodibromomethane	+	2.48	25.917	119.5	2.451	−0.337	0.247	1.919	0.055	0.013	33.30	66.32	97.22	4.99	2.50	1.71	89.31	1.09
42	1-bromo-2-chloroethane	+	1.62	23.879	106.5	1.723	−0.327	0.304	1.914	−0.021	0.014	2.89	86.54	88.39	45.36	1.52	1.48	427.60	0.08
43	1-bromobutane	−	2.71	28.335	102.0	1.276	−0.302	0.368	1.918	−0.053	0.008	5.49	90.10	93.50	28.23	1.72	1.66	336.80	1.94
44	2-bromobutane	−	2.71	28.313	91.0	1.255	−0.300	0.364	1.930	−0.070	0.009	21.53	36.16	54.13	4.75	2.83	1.89	106.40	1.94
45	1-bromo-3-chloropropane	+	1.88	28.526	144.5	1.592	−0.316	0.323	1.914	−0.032	0.011	6.20	140.00	144.60	40.20	1.78	1.72	517.50	1.89
46	2-bromo-1-chloropropane	−	2.16	28.504	116.5	1.478	−0.315	0.328	1.922	−0.032	0.012	12.40	88.48	97.90	17.64	2.47	2.23	408.10	0.43
47	1-bromo-2-methylpropane	−	2.58	28.313	91.0	1.260	−0.302	0.368	1.921	−0.056	0.010	10.97	58.73	66.08	11.04	2.06	1.83	174.70	1.95
48	2-bromo-2-methylpropane	−	2.45	28.291	73.0	1.189	−0.299	0.364	1.924	−0.073	0.011	18.25	41.00	41.02	4.95	2.20	2.20	100.50	2.00
49	1-bromo-2-methylpropene	NT	1.80	27.832	92.0	1.318	−0.272	0.307	1.916	−0.031	0.000	9.89	55.91	64.76	12.47	2.21	1.90	219.70	1.54
50	1-bromopentane	NT	3.25	32.982	130.0	1.218	−0.302	0.368	1.981	−0.054	0.000	5.99	149.90	153.20	42.33	1.69	1.65	496.80	1.95
51	1-bromooctane	NT	4.87	46.923	201.0	1.118	−0.301	0.368	1.919	−0.054	0.000	10.34	444.50	450.70	74.20	1.72	1.70	912.50	1.94
52	1-bromo-4-chlorobutane	−	2.42	39.341	81.0	1.488	−0.312	0.340	1.916	0.043	0.010	5.65	225.20	228.80	67.00	1.68	1.65	781.20	0.08

Table 1. (Continued)

no.	compound	aneu	log *P*	MR	bP	*d*	HOMO	LUMO	dist	charge	var_{ch}	I_x	I_y	I_z	R_x	R_y	R_z	EV	Dip
53	bromoethane	NT	1.63	19.041	38.5	1.460	−0.305	0.369	1.918	−0.051	0.013	2.83	21.80	23.59	12.56	1.63	1.50	129.00	1.92
54	1-bromo-2-methylbutane	NT	3.12	32.982	121.5	1.223	−0.299	0.366	1.922	−0.055	0.000	12.75	97.76	104.00	16.52	2.15	2.02	301.90	1.94
55	1-bromo-3-methylbutane	−	3.12	32.960	120.5	1.261	−0.300	0.368	1.920	−0.054	0.009	12.64	112.40	120.10	18.95	2.13	1.99	337.30	1.94
56	2-bromo-2-methylbutane	−	2.99	32.938	107.0	1.182	−0.297	0.365	1.926	−0.076	0.010	31.09	46.38	58.61	4.45	2.98	2.36	131.00	1.98
57	trichlorofluoromethane	NT	2.29	20.933	23.7	1.494	−0.429	0.192	1.827	0.285	0.026	34.67	34.67	48.68	2.29	2.29	1.63	35.69	0.44

[a] Aneu: aneuploidy. +: positive. −: negative. NT: not tested.

The Hodgkin similarity index is the following:

$$s_{ab}^{H} = \frac{2\int P_a P_b \, dV}{\int P_a^2 \, dV + \int P_b^2 \, dV}$$

where P_a and P_b are the structural properties of the two molecules being compared. For electrostatic potential, lipophilicity, and refractivity, the properties were calculated as follows:

$$P_{(r)} = \sum_{i=1}^{n} \frac{p_i}{|r - r_i|}$$

where $P_{(r)}$ is the potential at the *r* distance, p_i is the charge, lipophilicity, or refractivity on atom *i*, r_i is the position of atom *i*, and *n* is the total number of atoms in the molecule. A three-terms Gaussian approximation was used to fit the **1/*r*** curve.

In the shape calculations, the atomic electronic density functions were used as structural properties, instead of using the van der Waals radius. These functions were determined from the square of the STO-3G atomic orbital wave functions; three Gaussian functions were fitted to the resulting electron density of each atom type. The Carbo index is sensitive to the shape of a property's distribution, whereas the Hodgkin index is more sensitive to its magnitude.[3]

Results and Discussion

This paper investigates on three points relative to the similarity matrices. First, for a subset of molecules, different spatial orientations were considered, and the similarities with the other molecules were calculated for each spatial orientation. This permitted the study of the influence of the spatial orientation on the similarity values. Second, the similarity matrices, calculated according to both Hodgkin and Carbo, were globally compared to each other, and their information content was compared with that of the classical descriptors. Third, we studied the ability of the similarity matrices to (a) complement the classical descriptors in modeling the aneuploidy QSAR for the halogenated compounds and (b) to provide QSAR models alternative to those based on the classical descriptors.

Influence of the Spatial Orientation of the Molecules on the Similarity Index. In the standard procedure for the comparison of molecules in the program ASP, the similarity index of two compounds is obtained by initially superimposing their barycenters and then rotating and translating the molecules until the similarity index is maximized. The maximization of the index is carried out with an optimization procedure (Simplex). Since the optimization procedures are not "exact", and may lead to local minima, we studied how the initial conditions (spatial orientation of the molecules) affect the similarity indices.

We chose three molecules (**15**, **30**, and **50**), and we presented each of them to the ASP program in four different spatial orientations. In practice, a molecule was taken in an initial arbitrary orientation, and in three other orientations, which were obtained by 90° rotation around each of the three axes. Thus, different orientations of the same compound were considered as independent compounds, and a 12 × 12 shape similarity matrix (Hodgkin index) was computed (Table 2). In order to graphically display these results, the table was analyzed with principal component analysis (PCA)

Table 2. Shape (Hodgkin) Similarity Matrix: Effect of Different Spatial Orientations[a]

	compounds											
	15	**15**	**15**	**15**	**30**	**30**	**30**	**30**	**50**	**50**	**50**	**50**
15	1.00	1.00	0.99	1.00	0.94	0.94	0.94	0.94	0.74	0.74	0.74	0.74
15	0.99	1.00	0.99	0.99	0.95	0.94	0.94	0.94	0.74	0.74	0.73	0.74
15	0.99	0.74	1.00	0.99	0.94	0.94	0.94	0.94	0.74	0.74	0.74	0.74
15	1.00	0.75	0.77	1.00	0.94	0.95	0.94	0.94	0.74	0.74	0.74	0.74
30	0.94	0.94	0.94	0.94	1.00	1.00	0.92	0.95	0.80	0.81	0.81	0.81
30	0.94	0.81	0.77	0.68	1.00	1.00	0.92	1.00	0.81	0.81	0.81	0.81
30	0.94	0.80	0.81	0.80	0.92	0.73	1.00	0.95	0.80	0.81	0.81	0.81
30	0.94	0.94	0.94	0.94	0.99	1.00	0.95	1.00	0.80	0.81	0.81	0.81
50	0.74	0.74	0.74	0.74	0.80	0.81	0.81	0.80	1.00	1.00	0.99	0.95
50	0.74	0.74	0.74	0.74	0.81	0.81	0.81	0.70	0.95	1.00	0.95	1.00
50	0.74	0.74	0.74	0.74	0.81	0.81	0.81	0.60	0.99	0.48	1.00	0.94
50	0.74	0.74	0.74	0.74	0.81	0.80	0.81	0.72	0.95	0.47	0.49	1.00

[a] Compounds **15**, **30**, and **50** were taken in four different spatial orientations. Each orientation was considered as an individual chemical, and the 12 × 12 shape (Hodgkin) similarity matrix was calculated.

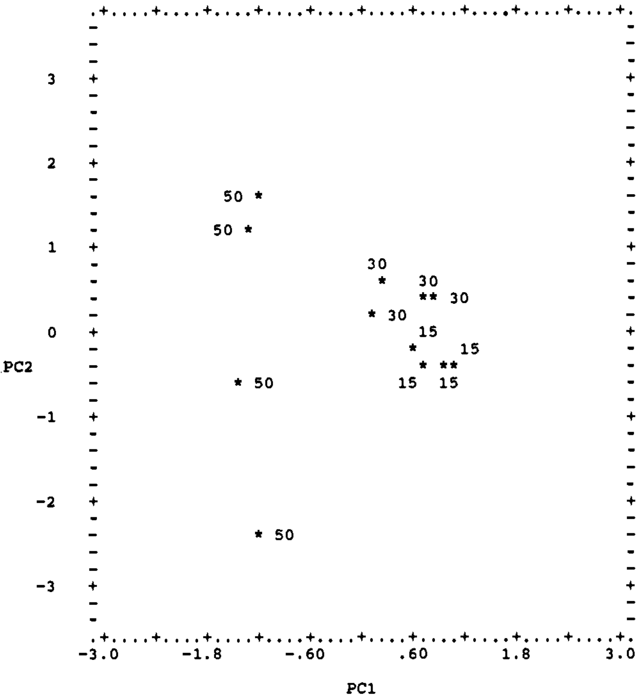

Figure 1. Graphical display of the relationships among compounds **15**, **30**, and **50** (each considered in four different spatial orientations) obtained by PCA of their shape (Hodgkin) similarity matrix (Table 2). PC1 and PC2 explain 78% of variance.

(Figure 1). The inspection of Table 2 and Figure 1 clearly shows that the similarity index values may be affected by the initial orientations of the compounds, in the sense that the same molecule, when taken in two different orientations and compared to itself, may give a similarity value lower than the expected 1.0, and the differently oriented versions may appear as really different molecules.

These results indicate that caution should be used in the analysis of similarity matrices data. When possible, the active intervention of the investigator in the initial orientation of the molecules may reduce the risk of occurrence of "false" distances.

Comparison between the Information Content of the Similarity Matrices and That of the Classical Chemical Descriptors. For the Table 1 compounds, the shape, electronic, refractivity, and lipophilicity similarity matrices were computed (both Carbo and Hodgkin indices). Each matrix was subjected to PCA. Table 3 lists the PCs with eigenvalues > 1.0 for each of the matrices. The Carbo index refractivity similarity matrix was not considered because all the similarity values were approximately 1.0. Also the Carbo index lipophilicity matrix showed a very limited variability, whereas both the refractivity and lipophilicity matrices calculated according to the Hodgkin index showed a wide distribution of values.

Table 3. PCs from the Similarity Matrices: Variance Explained[a]

	Hodgkin				Carbo		
	sha	ele	lip.	ref	sha	ele	lip.
PC1	.53	.44	.60	.60	.53	.45	.50
PC2	.19	.21	.40	.40	.19	.24	.34
PC3	.05	.11			.04	.09	.06
PC4	.02	.05			.03	.04	.03
PC5		.04				.03	
PC6		.02					

[a] The table lists the PCs obtained from each of the similarity matrices and reports their relative variance explained. Abbreviations: sha, shape; ele, electronic; lip., lipophilicity; ref, refractivity.

Table 4. Carbo Similarity PCs versus Classical Descriptors: Canonical Variable Loadings[a]

	CNVRF1	CNVRF2	CNVRF3	CNVRF4	CNVRF5
log *P*	−0.260	0.002	−0.410	−0.362	−0.404
MR	−0.625	0.040	−0.258	−0.149	−0.415
bP	−0.570	−0.187	−0.247	0.220	−0.467
d	0.297	−0.646	−0.296	0.031	0.185
HOMO	−0.023	0.504	0.306	0.106	0.403
LUMO	−0.278	0.827	0.167	0.011	0.227
dist	0.122	0.113	−0.008	−0.156	0.388
charge	0.345	−0.540	−0.103	−0.175	0.017
$\mathrm{var}_{\mathrm{ch}}$	0.385	−0.006	0.088	−0.007	0.079
I_x	0.210	−0.610	−0.240	−0.112	−0.525
I_y	−0.834	0.051	−0.363	−0.246	0.005
I_z	−0.779	−0.047	−0.398	−0.187	−0.012
R_x	−0.767	0.152	−0.009	−0.078	0.487
R_y	0.581	0.094	−0.020	0.037	−0.280
R_z	0.272	0.420	0.150	−0.139	−0.511
EV	−0.742	0.256	0.050	−0.030	0.416
prod	0.474	0.284	0.072	−0.041	−0.428
rat	0.588	−0.247	−0.141	−0.303	−0.474
dip.	0.152	0.818	−0.321	0.314	−0.238
	CNVRS1	CNVRS2	CNVRS3	CNVRS4	CNVRS5
C-S1	0.867	−0.303	0.330	0.179	−0.114
C-S2	−0.349	0.014	0.310	0.756	0.452
C-E1	0.405	0.849	−0.235	0.240	0.013
C-E2	0.534	−0.577	−0.531	0.242	−0.154
C-L1	0.626	−0.560	−0.408	−0.337	0.069
C-L2	0.365	0.079	0.485	0.417	−0.361

[a] CNVRF: canonical variable relative to the first set of variables. CNVRS: canonical variable relative to the second set of variables. Prod: R_yR_z. Rat: $(R_yR_z)/R_x$. C-S*i*: PC*i* from the shape similarity matrix (Carbo). C-E*i*: PC*i* from the electronic similarity matrix (Carbo). C-L*i*: PC*i* from the lipophilicity similarity matrix (Carbo).

An analytical survey of the information content of the similarity matrices was performed by a canonical correlation analysis of the two groups of variables: (a) the PCs derived from the similarity matrices and (b) the classical chemical descriptors. Table 4 reports the results for the Carbo index calculations, and Table 5 reports the results for the Hodgkin index calculations. All the canonical variables were statistically significant ($p < 0.001$); this indicates that there is a remarkable communality between the information content of the two sets of variables.

The inspection of the canonical loadings in Table 4

Table 5. Hodgkin Similarity PCs versus Classical Descriptors: Canonical Variable Loadings[a]

	CNVRF1	CNVRF2	CNVRF3	CNVRF4	CNVRF5	CNVRF6	CNVRF7
Log *P*	−0.671	0.201	0.357	0.292	0.272	0.076	0.100
MR	−0.894	−0.087	0.391	0.055	−0.058	−0.092	−0.035
bP	−0.858	−0.098	0.069	0.098	−0.267	−0.157	0.253
d	−0.249	0.303	−0.640	0.117	−0.147	−0.310	−0.009
HOMO	0.173	−0.159	0.374	−0.253	−0.607	−0.076	−0.435
LUMO	0.250	−0.398	0.755	−0.016	−0.238	0.093	0.084
dist	−0.063	0.084	0.121	−0.141	−0.257	−0.243	−0.004
charge	0.121	0.221	−0.544	0.044	0.457	0.081	0.066
$\mathrm{var_{ch}}$	0.479	0.080	−0.192	−0.108	0.197	−0.096	0.372
I_x	−0.642	0.642	−0.323	0.117	0.088	0.109	−0.007
I_y	−0.663	−0.567	0.266	0.129	0.138	−0.281	−0.115
I_z	−0.726	−0.485	0.175	0.128	0.049	−0.356	−0.178
R_x	−0.105	−0.904	0.122	0.001	0.042	−0.078	−0.049
R_y	0.031	0.665	0.120	−0.120	−0.158	−0.121	0.075
R_z	−0.028	0.455	0.519	−0.276	0.055	0.063	0.428
EV	−0.128	−0.844	0.226	−0.091	−0.060	−0.135	0.016
prod	0.007	0.616	0.351	−0.217	−0.062	−0.034	0.277
rat	−0.150	0.826	−0.060	−0.041	0.304	0.046	0.057
dip.	0.298	0.148	0.713	0.366	−0.204	−0.174	0.276
	CNVRS1	CNVRS2	CNVRS3	CNVRS4	CNVRS5	CNVRS6	CNVRS7
H-S1	0.413	0.661	−0.463	−0.290	−0.243	0.144	0.083
H-S2	−0.112	−0.443	−0.035	−0.056	−0.708	0.176	0.129
H-S3	−0.567	0.598	0.375	−0.192	0.103	0.107	0.292
H-E1	0.512	0.127	0.708	0.276	−0.297	−0.121	0.123
H-E2	0.060	0.664	−0.489	0.480	0.051	−0.239	−0.064
H-E3	−0.014	−0.322	−0.261	0.058	−0.252	−0.306	0.666
H-L1	−0.235	0.139	0.198	−0.421	−0.119	0.185	−0.146
H-L2	0.694	−0.011	−0.152	−0.303	−0.417	−0.104	0.204
H-R1	0.048	0.307	0.120	−0.458	−0.348	−0.278	0.134
H-R2	0.913	−0.070	−0.374	−0.041	0.083	0.081	0.020

[a] H-S*i*: PC*i* from the shape similarity matrix (Hodgkin). H-E*i*: PC*i* from the electronic similarity matrix (Hodgkin). H-L*i*: PC*i* from the lipophilicity similarity matrix (Hodgkin). H-R*i*: PC*i* from the refractivity similarity matrix (Hodgkin).

Table 6. Carbo Similarity PCs versus Hodgkin Similarity PCs: Canonical Variable Loadings

	CNVRF1	CNVRF2	CNVRF3	CNVRF4	CNVRF5	CNVRF6
C-S1	0.948	0.157	0.272	0.047	0.008	−0.009
C-S2	−0.205	0.133	0.841	0.478	0.052	0.036
C-E1	0.020	0.806	−0.338	0.485	0.020	−0.012
C-E2	0.723	−0.479	−0.180	0.457	−0.063	−0.039
C-L1	0.686	−0.461	−0.404	−0.001	0.376	0.113
C-L2	0.481	0.440	0.394	0.044	−0.292	0.577
	CNVRS1	CNVRS2	CNVRS3	CNVRS4	CNVRS5	CNVRS6
H-S1	0.908	0.169	0.363	0.076	0.011	−0.063
H-S2	−0.279	0.143	0.810	0.392	−0.138	−0.047
H-S3	0.167	0.041	−0.293	−0.288	−0.688	0.040
H-E1	−0.099	0.838	−0.283	0.445	−0.061	−0.007
H-E2	0.749	−0.341	−0.303	0.460	0.077	0.036
H-E3	−0.055	−0.192	0.250	0.377	−0.489	−0.269
H-L1	0.036	0.192	0.369	−0.111	−0.131	0.072
H-L2	0.458	0.420	0.405	0.206	0.117	−0.477
H-R1	0.319	0.305	0.358	0.123	−0.121	−0.423
H-R2	0.503	0.237	0.269	0.113	0.533	−0.102

Table 7. Global Comparison of Molecular Descriptions: Correlation Coefficients

	classical	Hodgkin1	Hodgkin2	Carbo
classical	1.			
Hodgkin1	0.535	1.		
Hodgkin2	0.559	0.998	1.	
Carbo	0.502	0.946	0.937	1.

[a] Euclidian distance matrices among compounds were calculated on the basis of (a) the classical molecular descriptors of Table 1 (classical); (b) the combination of the shape, electronic, and lipophilicity similarity matrices according to Hodgkin (Hodgkin1); (c) the combination of the shape, electronic, lipophilicity, and refractivity similarity matrices according to Hodgkin (Hodgkin2); and (d) the combination of the shape, electronic, and lipophilicity similarity matrices according to Carbo (Carbo). The distance matrices were then compared to each other by calculating their correlation coefficients (see details in the text).

shows that the first canonical variable is essentially a shape−size descriptor, with shape similarity PC1 on one side and MR, R_x, and EV on the other side. The second canonical variable can be easily interpreted as an electronic descriptor, with LUMO on one side and electronic similarity PC1 on the other side. In Table 5, the first canonical variable shows a correspondence between MR and log *P* on one side and refractivity similarity PC2 and lipophilicity similarity PC2 on the other side. The second canonical variable is the shape−size descriptor, and the third canonical variable summarizes the electronic aspects exemplified by LUMO.

Table 6 reports the results of the canonical correlation analysis of the Carbo index PCs and Hodgkin index PCs. The canonical loadings indicate very clearly the correspondence between the shape and electronic PCs relative to the two indices.

Together with this, we performed a more global comparison, according to the following procedure. The Euclidian distances between the compounds were calculated on the basis of the classical molecular descriptors, after previous normalization of the variables. Moreover, we calculated three further distance matrices, based on (**a**) a combination of shape, electronic, and lipophilic similarity (Carbo index), (**b**) a combination of shape, electronic, and lipophilic similarity (Hodgkin index) and (**c**) a combination of shape, electronic, lipophilic, and refractive similarity (Hodgkin index). The matrix **a** was computed as follows. The three Carbo similarity matrices were computed, each consisting of *N* variables (similarity with *N* compounds), thus producing a total number of 3 × *N* similarity variables. The Euclidian distance matrix **a** was then computed on the basis of these 3 × *N* variables. In a similar way, the

Table 8. Discrimination between Mutagenic and Nonmutagenic Compounds

a		*b*		*c*		*d*		*e*		*f*	
var	*F*	var	*F*	var	*F*	var	*F*	var	*F*	var	*F*
LUMO	19.6	LUMO	19.6	LUMO	19.6	C-E2	18.8	LUMO	19.6	H-E2	13.8
MR	12.5	MR	12.5	MR	12.5	C-E1	4.5	H-S3	9.0	H-S3	9.0
dist	0.2	I_z	1.0	C-S1	6.2			dist	0.2	H-E1	7.2
		dist	0.2	dist	0.2			H-R2	12.0	H-L2	4.5
*F**	16.7		21.9		17.7		12.4		17.7		14.9
acc	88.2		92.2		94.1		74.5		88.2		92.2
	(84.3)		(90.2)		(92.2)		(74.5)		(88.2)		(88.2)

[a] In order to separate the mutagens from the nonmutagens, stepwise linear discriminant analysis was applied to different sets of descriptors. For each analysis, the table reports the variables entered into the final equation (var), the *F* statistics valuee at step 0 (*F*), the global *F* statistics of the separation (*F**), and the accuracy of the separation (acc). The accuracy obtained by jackknifed cross-validation is reported within brackets.

distance matrices **b** and **c** were computed from 3 × *N* and 4 × *N* variables, respectively.

To make a global comparison of the different types of information, one-half of each symmetrical distance matrix was considered and the correlation coefficients among these mathematical objects were calculated (see refs 9 and 10). Table 7 reports the correlation coefficients. It appears that the information carried by the Carbo and Hodgkin similarity matrices is largely overlapping. On the contrary, a weaker correlation exists between the similarity indices and the classical molecular descriptors.

Similarity Matrices and the QSAR for Aneuploidy. We previously found that LUMO and MR determine the distinction between the halogenated aliphatic hydrocarbons, which are able to induce aneuploidy in *A. nidulans*, and those which are inactive. The presence of LUMO in the discriminant equation was interpreted in terms of a role for the reductive metabolism.[4,5] For this work, we recalculated the discriminant equation, by also including the biological results relative to a number of additional compounds. The results confirmed the importance of LUMO and MR. The discrimination was improved by the inclusion of dist (the length of the longest carbon–halogen bond), which can be related to the ease of bond breaking (Table 8a). A further small improvement was obtained by also considering the variable I_z (Table 8b). The inertia momentum parametrizes the mass distribution along an inertia axis of the molecule: the inclusion of I_z in the QSAR may point to a role for 3-D characteristics.

To study the contribution of the similarity index information, we performed linear discriminant analysis by considering together the variables LUMO, MR, and dist and subsets of the PCs derived from the similarity matrices. In particular, the results of Table 8c were obtained by analyzing together the classical descriptors and the Carbo index similarity PCs. In Table 8e, the classical descriptors were analyzed together with the Hodgkin index similarity PCs. The results of Table 8c,e show that the information carried by the similarity matrices complement the classical variables to some extent (see the improvement of the accuracy in Table 8c compared to that in Table 8a). However, if the *F* values (*F**) are considered together with the accuracies, it is clear that the improvement of the discrimination is negligible and lower than that obtained by adding the I_z variable (Table 8b).

Table 8d,f show the discriminations based only on either the Carbo index PCs or the Hodgkin index PCs. It appears that these PCs are suitable for discriminating between actives and inactives. In particular, the discrimination based on the Hodgkin index PCs (Table 8f) showed an accuracy higher than that of LUMO, MR, and dist (Table 8a) and equivalent to that of LUMO, MR, I_z, and dist (Table 8b), even though with a lower total *F*. It should also be noted that the PCs involved in the discrimination are those related to the shape and electronic characteristics of the molecules, in analogy with MR and LUMO.

As a conclusion, in this specific QSAR study the similarity matrices, per se, were almost as good as the classical variables but were not able to improve on their performance. Two hypotheses can be drawn. One hypothesis is that the spatial modulation of the molecular properties is important but that the similarity matrices are not suitable for quantifying this information. The other, alternative hypothesis is that the information on the spatial modulation of the molecular properties is not essential, whereas the mass (average) properties (such as LUMO or MR) play major roles. Concerning this latter hypothesis, it could be hypothesized that the cellular systems, which metabolize the xenobiotics, do not have very strict spatial requirements (such as the pharmacological receptors) but have broader specificities, in order to cope with large spectra of chemical structures. The results of this investigation do not allow us to choose either of the two hypotheses.

Conclusions

The main conclusion of this work is that a satisfactory QSAR for aneuploidy in *A. nidulans* can be obtained on the basis of the similarity matrices, analogous to that obtained with the global descriptors, even though the QSAR based on similarity indices did not improve on the performance of the QSAR based on the classical properties. This should be considered together with the parallelism between PCs from similarity matrices and classical descriptors. It should be stressed that Tables 4 and 5 analyses point to scientifically sound correlations: for example, the Carbo shape similarity PC1 is correlated with MR, R_x, and EV, and the Carbo electronic similarity PC1 is correlated with LUMO (Table 4). This overall evidence (satisfactory aneuploidy QSAR based on similarities; relationships between similarity matrices and classical descriptors) indicates that the sum of local information (similarities between pairs of compounds), which is contained in an *N* × *N* similarity matrix, permits the construction of other, more global information on the average properties of the compounds. These average properties are those exemplified by the classical molecular descriptors. However, the results of this work cannot answer whether the spatial modulation of molecular properties is important for the aneuploidy QSAR or if the method of the similarity matrices simply failed to capture this aspect. In our opinion, in this case, a QSAR based on the classical descriptors may be preferred because it is more easily rationalized in scientific terms and is more suitable to comparisons with other QSARs (lateral validation).[11] More generally, it is advisable to further compare QSARs based on similarity indices and QSARs based on classical descriptors. This may provide interesting clues as to the suitability of this novel approach to QSAR studies and may answer specific questions, such as to which cases it should be applied and in which specific form (e.g., the difference between Carbo and Hodgkin indices).

Acknowledgment. Dr. Riccardo Crebelli is gratefully thanked for kindly providing new aneuploidy data. Dr. Alessandro Giuliani and Cristina Andreoli are thanked for their continued interest and helpful discussion. Ms. Eve Silvester is aknowledged for her patient editing of the text.

References

(1) Burt, C.; Huxley, P.; Richards, W. G. The application of molecular similarity calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.

(2) Good, A. C.; So, S.; Richards, W. G. Structure-Activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433–438.

(3) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.

(4) Crebelli, R.; Andreoli, C.; Carere, A.; Conti, G.; Conti, L.; Cotta-Ramusino, M.; Benigni, R. The induction of mitotic chromosome malsegregation in Aspergillus nidulans. Quantitative structure-activity relationship (QSAR) analysis with chlorinated aliphatic hydrocarbons. *Mutat. Res.* **1992**, *266*, 117–134.

(5) Benigni, R.; Andreoli, C.; Conti, L.; Tafani, P.; Cotta-Ramusino, M.; Carere, A.; Crebelli, R. Quantitative structure-activity relationship models correctly predict the toxic and aneuploidizing properties of six halogenated methanes in Aspergillus nidulans. *Mutagenesis* **1993**, *8*, 301–305.

(6) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of chemical property estimation methods;* McGraw-Hill: New York, 1982.

(7) Oxford Molecular Ltd., The Magdalen Center, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.

(8) Biosym Technologies, 9685 Scranton Rd., San Diego, CA 92121-2777.

(9) Benigni, R. Relationships between in vitro mutagenicity assays. *Mutagenesis* **1992**, *7*, 335–341.

(10) Richtsmeier, J. T.; Lele, S. A coordinate-free approach to the analysis of growth patterns: models and theoretical considerations. *Biol. Rev.* **1993**, *68*, 381–411.

(11) Hansch, C. QSAR and the Unnamed Science. *Acc. Chem. Res.* **1993**, *26*, 147–153.

JM940306E