On the Equivalence Between Different Descriptions of Molecules: Value for Computational Approaches

Romualdo Benigni,*,† Grazia Gallo,‡ Fabrizio Giorgi,‡ and Alessandro Giuliani†

Istituto Superiore di Sanita', Laboratory of Comparative Toxicology and Ecotoxicology, Viale Regina Elena 299, 00161 Roma, Italy, and Sigma-Tau Chemical Research Laboratories, Via Pontina Km. 30.400, Pomezia, Italy

Received November 24, 1998

The correlation between two different sets of chemical descriptors for the same heterogeneous set of molecules was investigated at both local and global levels. The two descriptions were well correlated at the global level, whereas they showed different characteristics on the detailed scale. This result has important consequences for both combinatorial chemistry strategies and QSAR analyses of congeneric series. It implies that the "maximal diversity" optimization programs are largely independent from the type of chemical descriptors only in large ensembles of molecules, but not in individual series. On the other hand, the general isomorphism between different chemical descriptions may give rise, for QSAR analyses of congeneric series, to a multiplicity of formally equivalent solutions for the same QSAR problem.

INTRODUCTION

As a consequence of the rapid explosion of combinatorial strategies in drug discovery in the past decade,[1−3] some totally new problems have emerged: maximizing the diversity of sets of molecules, optimizing subset selection with respect to some desired property, choice of the most efficient strategies to lead discovery out of huge collections of candidate structures, and so forth.[1−5] The nature of these problems is completely different from those classically faced by theoretical chemistry; the difference arises because these problems deal with the properties of heterogeneous populations of different molecules, whereas theoretical chemistry classically deals with single molecules or with "thermodynamic" ensembles made of only one (or few) molecular species. In between are the classical problems analyzed by QSAR, usually focusing on congeneric sets of molecules. This new perspective forces the chemist to make use of different conceptual tools.[6−9] (See also the growth of computational methods that can emphasize diversity or similarity.[10])

Despite the unique features of each sample of chemicals, the search for general theoretical insights into the problem has an obvious importance. Trying to summarize this complicated issue in only one basic question, we can ask to what extent the different vectorial representations of molecules (e.g., chemicophysical descriptors, topological indexes, or whatsoever) are isomorphic? To what extent can we map a particular representation into another? This question is at the basis of the legitimacy of all the efforts to maximize the internal diversity of sets of molecules or of exploring similarity spaces. The answer, to have practical utility, must be given not on a purely theoretical basis in the frame of basic quantum-molecular features (that are by definition population independent and thus valid only in the limit of infinite populations or single molecular species) but at the level of the actual finite populations with which the chemists deal. At this empirical, finite-scale level, the problem can be stated in terms of statistical correlations between different representations (choices of vectorial formalization) of the same set of molecules.[9] The degree of correlation must be investigated at two different levels: the global level of the entire population (relevant for the "factorial design" class of problems[11]) and the local level of single-molecule neighborhoods (relevant for the "similarity searching" or local "distance-maximizing strategies" class of problems[12]).

In this work we used this approach to study a relatively small (293 molecules) but highly heterogeneous population, for which we calculated two different sets of descriptors: (a) fingerprints (presence or absence of substructures); (b) one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) descriptors. The relationship between the two descriptions was investigated separately at the two levels by canonical analysis[13] for the global scale and cluster analysis[14] for the local scale.

MATERIAL AND METHODS

Data. The studied population consisted of 293 noncongeneric molecules including a large range of structural classes, such as, polycyclic aromatic rings, diazo aromatic derivatives, polyhalogenated hydrocarbons, dioxines, polyalcohols, cyano, nitro, and amino benzene derivatives. These molecules were selected from a carcinogenicity database that we had previously studied from the point of view of QSAR. The names of 280 of 293 compounds were reported previously.[15] In this work, we did not perform any QSAR, but we exploited the highly noncongeneric character of the database to compare two different sets of chemical descriptors.

The molecules were projected in two chemical spaces. The first one [Molkey (MK)], taken from the MACCS database

* Author to whom correspondence should be addressed. E-mail: rbenigni@iss.it.

† Istituto Superiore di Sanita.

‡ Sigma-Tau Chemical Research Laboratories.

10.1021/ci980228u CCC: $18.00
Published on Web 04/24/1999

software (Maccs II; Molecular Design Ltd, San Leandro, CA) and made up of 166 binary variables describing the structural key, relied on the use of a predefined fragment dictionary of small generic and specific fragments with the number of occurrences up to 3. For our data set we used a total of 148 binary variables, after deletion of ones having null variability or a correlation coefficient with another binary variable equal to one. The second chemical space (TSAR) was a 1D, 2D, and 3D description of the data set, including 37 variables calculated with the TSAR program (TSAR v3.1: Oxford Molecular Ltd., England): 1D, molecular mass, number of atoms, number of halogen atoms, number of heteroatoms, number of H-bond donors and acceptors, number of 3,4,5,6, and 7 rings; 2D, Kier ChiV (six descriptors), κ indexes (three descriptors), κ α indexes (three descriptors), shape flexibility index, Randic, Balaban, and Wiener topological indexes, sum of E-state index; 3D, molecular surface area, molecular volume, inertia moment size and length (six descriptors), and ellipsoidal volume.

Statistical Methods. A useful classification of the multivariate techniques[7,16] separates these techniques into "microstructure" and "macrostructure" methods.[17] Although the former are particularly suited to highlighting the local structure of a data field, the latter ones highlight the general trends and correlations present in the data. The common material of both the classes is the multivariate data matrix having the statistical units as rows (in our case, chemicals) and the variables as columns (in our case, MK and TSAR descriptors). The microstructural techniques are driven by "goal functions" based on the similarities (or equivalent distances) between pairs of rows (and thus on local properties of the data field).[16,17] On the contrary, the macrostructural methods are driven by ensemble statistics spanning all the data field (general features) like the correlation coefficients between variables. The most commonly used representatives of the micro and macro methods are cluster analysis and principal component analysis (PCA), respectively. Cluster analysis techniques allow for a thorough definition of the neighborhood of a particular unit but cannot give a global metric representation of the data field.[16] On the other hand, PCA (and in general factorial techniques) defines an optimal orthogonal low-dimension basis for multivariate sets[18] but provokes unescapable distortions on the local scale.[16,17]

We made use of this complementarity to appreciate the different correlation behaviors of the two chemical spaces at different levels of definition. The macroscale correlation was quantitated by canonical correlation analysis,[13] a widely used factorial technique driven by the search for the linear combinations of two sets of variables (X and Y; in our case, MK and TSAR) maximizing their mutual correlation (Pearson's r between X and $Y = \max$).

The concordance between the microscale features of the two spaces was measured by the relative degree of "clusterization" of the two spaces. In practice the R^2 relative to different applications of the k-means clustering procedure,[14,19] computed separately at various values of k for the two spaces, was computed.

RESULTS AND DISCUSSION

The aim of the present work was to compare two different types of chemical descriptors for the same set of chemicals (293 noncongeneric compounds). The first description (MK) consisted of 148 structural keys, whereas the second one (TSAR) consisted of 37 1D, 2D, and 3D variables. To rule out possible confounding effects caused by the different "information grain" of the two spaces (148 MK variables and 37 TSAR variables, with MK descriptors singularly carrying a very low information content with respect to TSAR), and by the existence of correlation structures within the individual spaces, both the canonical and cluster analyses were performed after the original spaces were filtered by PCA. The results of the PCA filtering of both MK and TSAR spaces are reported in Figure 1; the increase in percentage of explained variance at increasing numbers of eigenvectors [principal components (PCs)] is shown. Because MK had a lower information density than TSAR, we decided to perform the comparison of the two descriptions at equal levels of explained variance. Thus, the canonical and cluster analyses procedures were computed separately on three distinct pairs of PC spaces, each corresponding to a different level of explained variability: A, approximately 50% explained variance; B, approximately 75% explained variance; C, approximately 90% explained variance.

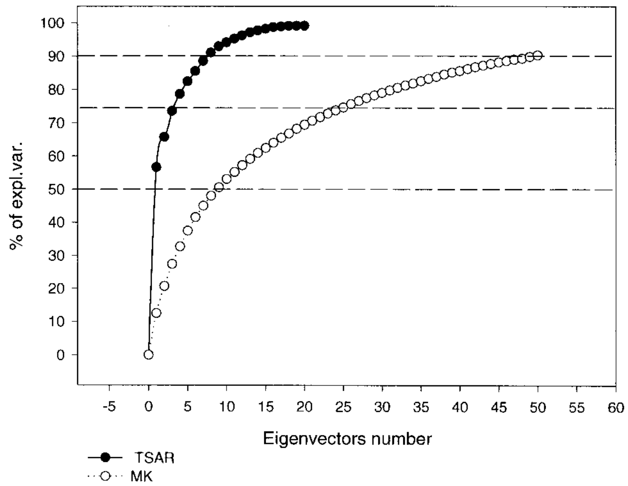

Figure 1. The progressive increase in percentage of explained variability of the principal component solution at an increasing number of components is reported for both TSAR (filled) and MK (open) spaces. The broken lines correspond to the three levels (A, B, C) of percentage of explained variability chosen for the analysis.

The chosen three levels of definition A, B, and C (50%, 75%, 90% explained variance) were reached at very different dimensionalities in the two spaces. Level A corresponds to only one TSAR component (TSAR1, 56.6% explained variance) versus 10 MK components (MK1−MKI10, 53.1% explained variance). Level B corresponds to three TSAR components (TSAR1−TSAR3, 73.5% explained variance) versus 25 MK components (MK1−MK25, 74.6% explained variance). Level C has eight TSAR components (TSAR1−TSAR8, 91.0% explained variance) and 50 MK components (MK1−MK50, 90.3% explained variance).

At each of the three levels of definition, MK and TSAR PCs were compared by canonical correlation analysis. The highest canonical correlation coefficients between the MK and TSAR component spaces were 0.77, 0.84, and 0.98 at levels A, B, and C, respectively. This result points to an almost perfect global mapping between MK and TSAR. There was a progressive increase in the concordance between the two spaces at increasing amounts of information, thus

Table 1

	k value (no. of clusters)			
chemical spaces	2	5	10	20
TSAR (A)	0.43	0.88	0.97	0.99
TSAR (B)	0.17	0.60	0.75	0.86
TSAR (C)	0.06	0.39	0.60	0.77
MK (A)	0.05	0.23	0.48	0.64
MK (B)	0.03	0.11	0.22	0.37
MK (C)	0.01	0.06	0.14	0.27

[a] The clustering propensity of the TSAR and MK spaces at a varying number of clusters (*k*) and percentage of explained variability (A = 50, B = 75, C = 90) of the relative principal component solution are reported. The clustering propensity is expressed as the R^2 of the particular *k*-means solution. Both the clustering propensity increase at increasing values of *k* and the clustering propensity decrease at increasing values of explained variability of the principal component solution are well-known structural consequences of PCA and *k*-means algorithms. The difference in clustering propensity between the TSAR and MK spaces points to a very different local strucure for the two spaces.

pointing to a substantial isomorphism between the total information content of the two spaces.

It is also important to stress that the observed isomorphism is a global characteristic of the two spaces: the correlations between single TSAR−MK component pairs were negligible, with the only exception of TSAR1,MK1 correlation ($r = 0.51$, $p < 0.0001$) indicating a common meaning (probably molecular size) of the first component of the two data fields.

The second type of comparison between MK and TSAR descriptors was performed by cluster analysis of the chemicals. In this case, the aim of the analysis was to investigate the “shape” of the two spaces in terms of relative “clusterization propensity”, i.e., the degree of departure from a uniform density of points. This departure was measured by the R^2 of different *k* means partitions applied to the data sets. R^2 is the ratio of interclusters to total variance.

Table 1 reports the clusterization propensity (R^2 of the partition) of the two spaces at the three different levels of definition and with four cluster number choices ($k = 2$, 5, 10, and 20). As explained above, TSAR(A) indicates that only the PCs that explain 50% of TSAR variance (level A) were used, and so on. The evidence provided by this analysis was 2-fold. The first evidence is obvious and derives from the nature of the algorithms used. The decrease in cluster propensity at an increasing number of components is a well-known result[16] coming from the sensitivity of PCA to each departure from symmetry in the data field (e.g., presence of clusters).[18] In fact the major components orient themselves in the direction of the main-order parameters (i.e., broken symmetries) present in the data, so exalting the clusterization of the data field.[20,21] On the other hand, the proportionality between R^2 and *k* is caused by the obvious fact that R^2 is the ratio of interclusters to total variance, and these two measures tend to coincide for increasing numbers of clusters. The second evidence is the most relevant for our analysis. The comparison of the results for MK and TSAR showed that TSAR had a much higher cluster propensity than MK. Probably this feature is a consequence of the different information density of the two spaces (at equal percentage of explained variability MK space has an higher dimensionality than TSAR). Whatsoever is the cause of this behavior, the important point is that the high correlation between MK and TSAR we observed at the global scale (seen by canonical correlation analysis), was paralleled by a totally different structure of the two spaces at the local level, as indicated by the completely different clusterization propensity.

CONCLUSION

Our results confirmed previous indications[9] about the existence of a different mapping behavior of the chemical spaces at different scales. The scaling behavior of the data set analyzed in this work pointed to an high degree of concordance between different representations at the global scale, coupled with a substantial lack of correlation at the local scale. This complex scaling, if confirmed, has two main consequences in combinatorial chemistry; although a factorial design procedure (that is based on gross scale characteristics) appears independent from the chosen formalization, this is not the case for subset selection and for the generation of homogeneous series of compounds (that are based on local features of the data field).

An important consequence of the correlation found between different chemical descriptors at the global level is that different QSAR models, with equivalent accuracy, generally can be expected for the same case. This correlation makes it impossible to derive purely automatic criteria of choice between competing models. It is important to note, however, that our results refer to a noncongeneric set of molecules and thus to a wide (even if sparse) sampling of the chemical space. This choice was dictated by the necessity to derive general information about chemical descriptions correlation; on the contrary QSAR analyses refer to very small but dense sampling of the chemical space (congeneric sets). At this fine detail, the observable correlations strictly depend on the particular data set analyzed and do not necessarily reflect the general trends observed in noncongeneric ensembles. Nevertheless the existence of a sort of “thermodynamic limit” in which all the descriptions converge is an important boundary condition to be taken into account when analyzing each particular “local” case.

REFERENCES AND NOTES

(1) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(2) Dean, P. M. *Molecular Similarity in Drug Design*; Chapman and Hall: London, 1995.

(3) Gallopp, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1385−1399.

(4) Ferguson, A. M.; Patterson, D. E.; Garr, C.; Underiner, T. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen* **1996**, *1*, 65−73.

(5) Moos, W. H.; Green, G. D.; Pavia, M. R. Recent Advances in the Generation of Molecular Diversity. *Annu. Rep. Med. Chem.* **1993**, *28*, 315−324.

(6) Lebart, L.; Morineau, A.; Warwick, K. M. *Multivariate Descriptive Statistical Analysis*; Wiley: New York, 1984.

(7) Benigni, R.; Giuliani, A. Quantitative Modeling and Biology: The Multivariate Approach. *Am. J. Physiol.* **1994**, *266(35)*, R1697−R1704.

(8) Klein, D. J.; Babic D. Partial Orderings in Chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 656−671.

(9) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. Four Association Coefficients for Relating Molecular Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909−915.

(10) Willett, P. *Computational Methods for the Analysis of Molecular Diversity*; Kluwer: The Netherlands, 1997.

(11) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nystrom, A.; Pettersen, J.; Bergman, R. Experimental Design and Optimization. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3−40.

(12) Potter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.

(13) Rencher, A. C. Interpretation of Canonical Discriminant Functions, Canonical Variates, and Principal Components. *The American Statistician* **1992**, *46*, 217−225.

(14) Everitt, B. *Cluster Analysis*; Halsted Press: New York, 1980.

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

(16) Sneath, P. H. A.; Sokal, R. *Numerical Taxonomy*; Freeman: San Francisco, 1973.

(17) Van Ryzin, J. *Classification and Clustering*; Academic Press: New York, 1977.

(18) Broomhead, D. S.; King, G. P. Extracting Qualitative Dynamics from Experimental Data. *Physica D (Amsterdam)* **1986**, *20*, 217−236.

(19) Hartigan, J. A. *Clustering Algorithms*; Wiley: New York, 1975.

(20) Giuliani, A.; Colosimo, R.; Benigni, R.; Zbilut, J. P. On the Constructive Role of Noise in Spatial Systems. *Phys. Lett. A* **1998**, *247*, 47−52.

(21) Watkin, T. L. H.; Rau A. The Statistical Mechanics of Learning a Rule. *Rev. Mod. Phys.* **1993**, *65*, 499−512.

CI980228U