Looking for an Unambiguous Geometrical Definition of Organic Series from 3-D Molecular Similarity Indices†

Marina Cotta-Ramusino, Romualdo Benigni,* Laura Passerini, and Alessandro Giuliani

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

Received June 3, 2002

A mathematically consistent definition of series was derived by the application of Principal Component Analysis to 3-D similarity indices (ASP Software). For two data sets of aromatic molecules, a mono-dimensional numerical ordering was derived, corresponding to the consensus distance from the series lead. The series is unambiguously defined in terms of location along the mono-dimensional axis, along which shape and electrostatic features of the molecules vary in a coordinated way. Molecules along the axis are characterized by the same pattern of electrostatic potential and shape shared by the lead compound, of which they represent linearly "shifted" replicas. This approach can contribute to the exploration of the chemical spaces from a local "fine grain" perspective, thus complementing the "coarse grain" descriptions normally used in the analysis of large data sets.

INTRODUCTION

The generation of measures of molecular diversity is a crucial problem when handling chemical sets and databases. Depending on the problem under consideration, the molecular diversity issue can be viewed either as the generation of suitable and easy-to-use maps of the explored "chemical territories" (e.g. for combinatorial chemistry purposes) or as the design of chemical series suited for empirical Quantitative Structure−Activity Relationship (QSAR) modeling. In the former case, the goal is mainly the definition of coarse grain maps of the chemical space.[1] In the latter case, the focus is shifted toward the definition of self-consistent organic series as a surrogate for a common mechanism of action that supports any successful QSAR procedure. The investigation on this issue dates back to the foundation of QSAR by Corwin Hansch and has led to the commonly accepted definition of a series of organic molecules as a set of chemicals sharing a common skeleton with different substituents attached to it (and possibly acting through the same mechanism). The preponderance of the common moiety with respect to the variable sites guarantees that the molecules belong to a series.[2]

Both the combinatorial chemistry and the classic QSAR approaches imply the existence of a chemical space endowed with a specific intrinsic geometry. The difference is that the combinatorial chemistry approach involves a mere chemically based characterization of the space, referring only to the chemical physical or topological description of the molecules, whereas the QSAR approach implicitly assumes the existence of a latent biological phase. This biological phase is represented by the macromolecule that acts as the receptor in the biological activity and that eventually "chooses" among the molecules those belonging to the series. In a strictly pharmacological sense, a series is the set of ligands of a given receptor. Since usually an explicit model of the receptor is lacking, to minimize the risk of departure from the recognition space of the receptor only relatively small displacements from an already known ligand are considered. When the receptor comes to light as an explicitly computable structure, the QSAR analysis can shift to the direct description of the strength and character of the macromolecule/ligand interaction. This eliminates the need for a series in a strictly chemical sense and replaces the surrogate with the real boundary conditions of the model.[3]

The chemical physical concept of series that approximates the unknown receptor model mainly resides upon qualitative reasoning, as for the definition of "small departures" from the basic skeleton. In this paper we will try and give a quantitative basis to the concept of series by investigating the fine grain structure of the chemical space, re-visiting the concept of organic series from the point of view of quantum chemical similarities. The tool we used was the ASP software,[4] which implements a method based on the superposition of each molecule to a lead compound, and on the estimation of the 3-D based similarity of steric and electronic distribution.[5−9] In this way, the usual molecule/descriptor matrices employed in the coarse grain explorations of the chemical space was enriched with a further dimension represented by the spatial characteristics of the molecules, thus adding information related to the latent biological phase.

Together with the applicative side of the problem, we addressed also a more theoretical issue: we checked the definition of organic series against the original mathematical meaning of the word "series" as a sequence of consecutive elements differing by a common ratio. In chemistry, this definition strictly holds only for linear aliphatic hydrocarbons (whose consecutive elements differ by the (-CH2-) common ratio). We investigated the applicability of the mathematically rigorous definition of series to chemical spaces where molecular chemical similarities are computable. The success

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

† This paper is dedicated to the memory of our friend and colleague Marina Cotta-Ramusino.

10.1021/ci020375z CCC: $25.00
Published on Web 11/28/2002

Table 1. Computational Options (see Details in Methods)

run	similarity index	integration method	superposition option	exhaustive search
1	Carbò	Gaussian (3 Gaussians)	none	
2	Carbò	Grid	none	
3	Hodgkin	Gaussian (3 Gaussians)	none	
4	Hodgkin	Grid	none	
5	Carbò	Gaussian (3 Gaussians)	simple	full rigid (10° angle increment)
6	Carbò	Gaussian (3 Gaussians)	simple	full torsional (20° angle increment)
7	Hodgkin	Gaussian (3 Gaussians)	simple	full rigid (10° angle increment)
8	Hodgkin	Gaussian (3 Gaussians)	simple	full torsional (20° angle increment)

of this approach would correspond to the demonstration of a generalized, self-consistent definition of "series"; this would provide a tool for exploring chemical spaces from a local fine grain perspective, thus complementing the coarse grain descriptions.

The strategy was to derive a consensus metric by applying Principal Component Analysis (PCA) to a range of similarity indexes. The first PC, if endowed with all positive loadings with the original similarity indexes, corresponds to the mathematically defined series.

DATA AND METHODS

The molecules of two organic series were analyzed as for their relative similarity with the lead molecule of each series. The similarity indices were computed with the ASP[4] software.

The data set A included benzaldehyde (lead compound) and 26 benzaldehyde derivatives. Some of the considered compounds display conformational isomerism of the phenyl ring substituents with respect to the aldehydic group. In this instance the possible conformers have been explicitly considered as different molecular entities, giving rise to 42 molecular structures (Figure 1). The same line of reasoning was applied to the data set B which included nitrobenzene and 26 nitrobenzene derivatives. In this case 29 molecular structures were considered (Figure 2).

Since the ASP molecular similarity indices are sensitive to the relative orientation of the compounds as well as to the optimization of the particular structures, we adopted a multiplicity of optimization procedures to single out (if any) the consensus distance as the algorithm independent metric defining the series.

The molecular structures of the considered compounds were constructed with the molecular modeling package Sybyl[10] and fully optimized with the Tripos force field method;[11] the partial atomic charges were obtained with the Gasteiger-Marsili procedure.[12]

For some of the considered compounds, the location of the substituent groups on the phenyl ring can give rise to different conformers; in this case each possible conformer was taken into account as a different molecular entity.

For each data set of molecular structures, two databases were generated, one containing the considered compounds as they were built by Sybyl, while the second database was obtained with the "ALIGN DATABASE" command in Sybyl, by which an optimal rigid alignment of each molecular structure is sought with respect to a common template (the benzaldehyde ring and the nitrobenzene moiety for the first and the second set of compounds, respectively).

Two different approaches were used to calculate the similarity indices: 1) an analytical methodology in which three Gaussian functions were used to compare properties throughout space and 2) a grid based method in which a three-dimensional grid of points surrounded the molecules being compared. Moreover when the pre-aligned databases were not used, the computational procedure tried to maximize the similarity indices by performing a full orientation search using rigid rotations or a flexible search modifying the different torsion angles.

The selected computational options are reported in Table 1.

Computational runs 3, 4, 7 and 8 were performed using the pre-aligned databases.

At the end of the above-mentioned procedure each molecule is defined by three different similarity indices (based respectively on electrostatic potential and shape distribution, plus a combined similarity average of the first two) for each different computation procedure for a total of 24 variables named COMBi, CHAi, SHAi, i being the method label and COMB, CHA and SHA indicating combined, charge and shape similarity, respectively. The multivariate data matrix having as statistical units the molecules and as variables the 24 similarity indices was the raw material for the application of PCA on the correlation matrix. The choice of the correlation matrix, instead of the covariance matrix, has the consequence that the emergence of a first PC with all positive loadings is not due to the use of variables with positive signs but to the existence of a significant common axis of variation for all the similarity indices.[13]

RESULTS AND DISCUSSION

The Definition of Series. In chemistry, a mathematically rigorous definition of organic series holds only for linear aliphatic chains. It should be recalled that a mathematical series is unambiguously defined by two parameters: the seed (the initial value of the series) and the operation to be performed to obtain the (n+1)th element of the series from the nth one. Setting the two parameters generates a mono-dimensional metric space, where the number of steps separating each element from the initial seed can be defined. A geometrical image of the series is a narrow path—starting from the seed—along which all the subsequent elements of the series are located, whereas the objects outside this path do not belong to the series. To translate this image into chemistry, the two basic concepts to retain are as follows: 1) the seed and 2) the mono-dimensional path (e.g. rule to be applied to the nth element of the series to obtain the (n+1)th). The seed corresponds to the leader of the series (the simplest structure, made only of the basic moiety), while the mono-dimensional path corresponds to a unique distance from the seed independent of the particular variable or algorithm used to compute the distance itself. This is

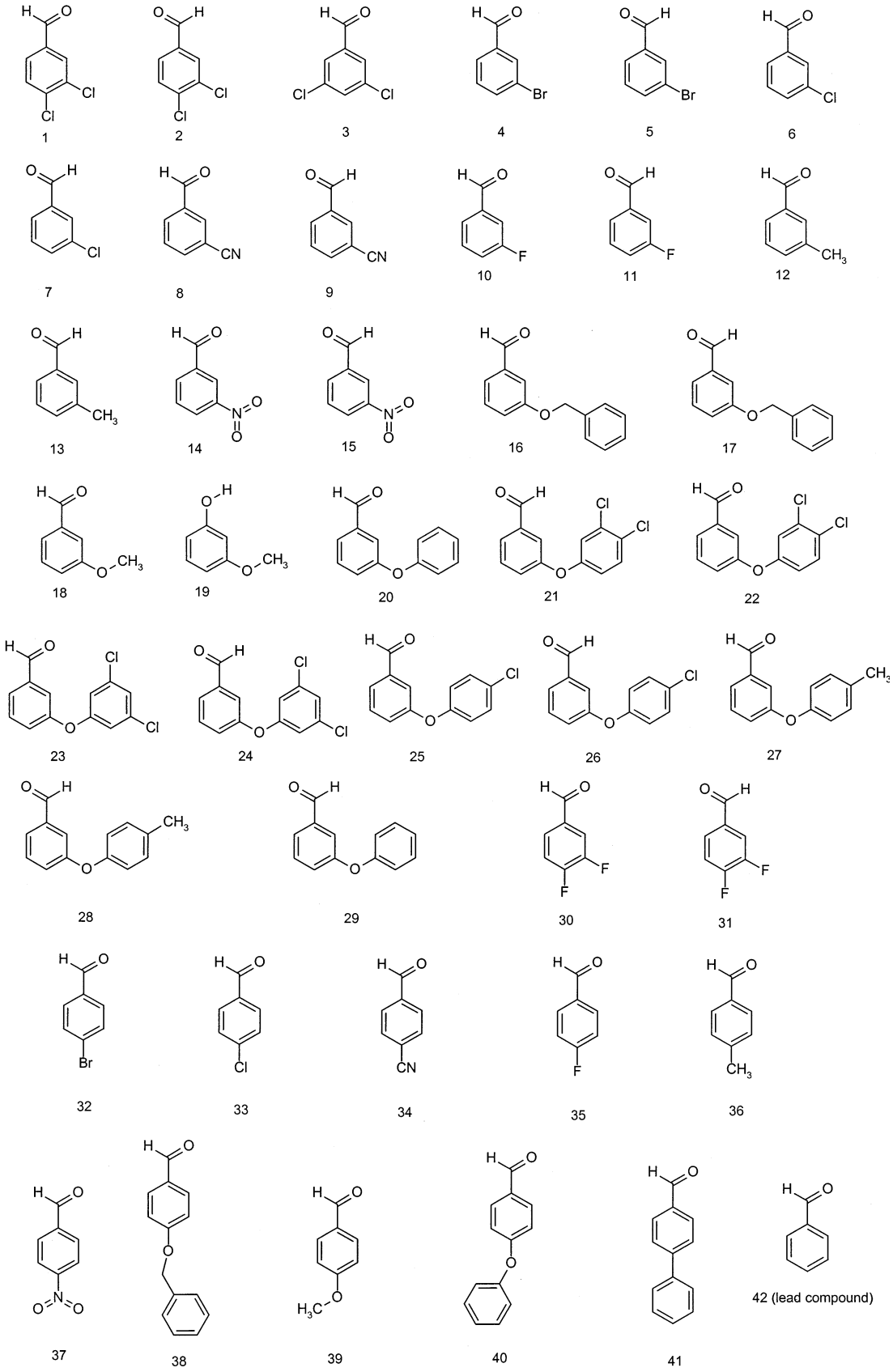

Figure 1. Molecules in set A.

particularly evident for the aliphatic linear hydrocarbons, where many chemical physical characteristics can be traced back to the number of carbon atoms that constitute a mono-dimensional path in whatsoever high dimensional chemical space.

To enlarge the reach of such a definition, we studied two sets of aromatic molecules by computing their three-dimensional molecular similarities with a lead compound and checked for the existence of a mono-dimensional path in the form of a consensus distance correlating all the used metrics.

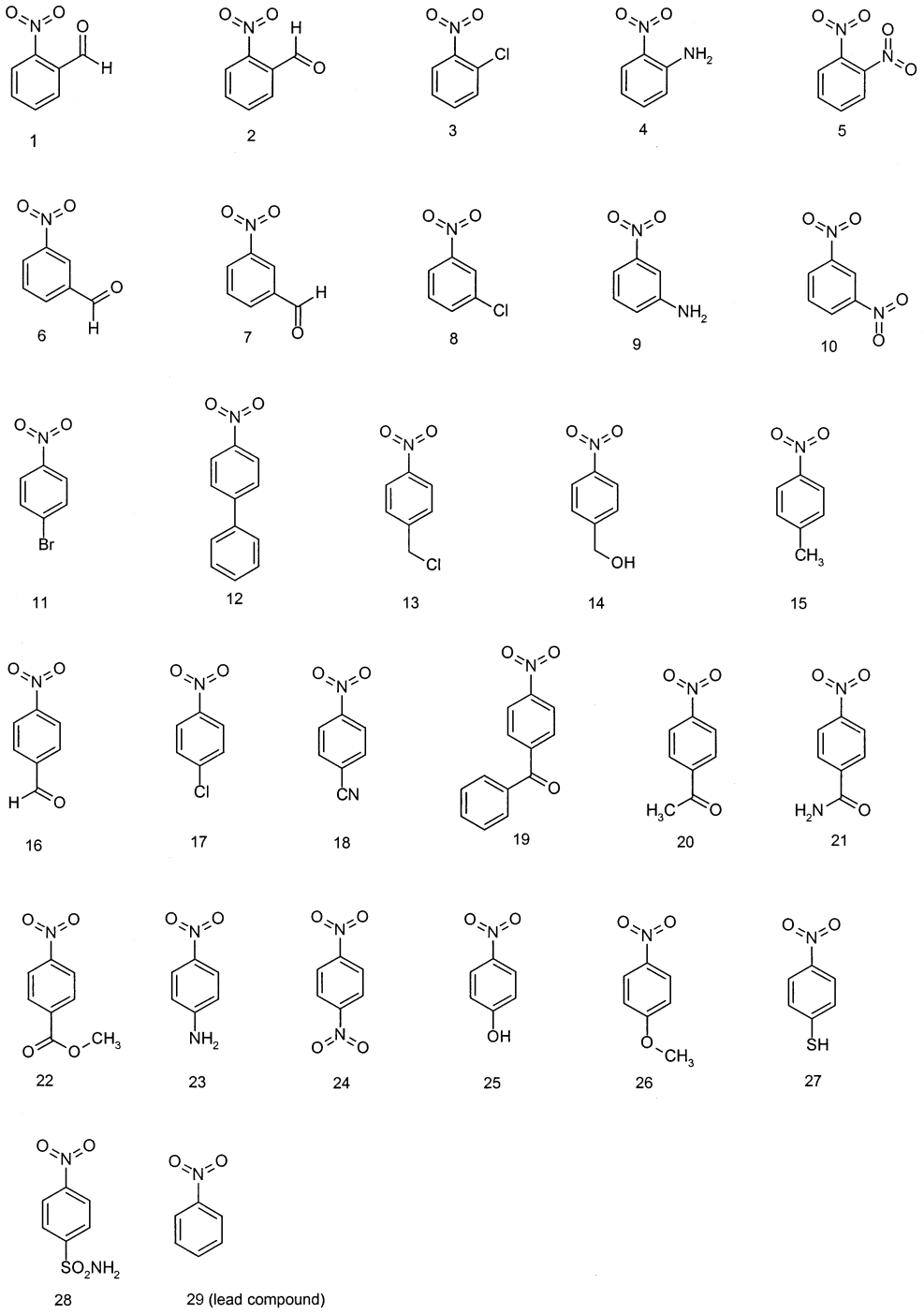

Figure 2. Molecules in set B.

Chemical Series and Molecular Similarity Patterns. The decision of using the similarity computation as the basic tool for the present investigation derives directly from the nature of the aim of this work; in fact analyzing a chemical series corresponds to measuring if and how much a number of individual chemicals fit a group of or is similar to a group of chemicals considered to be part of the series. In recent years, algorithms for the measure of chemical similarity have developed to a remarkable degree of sophistication.[5−9] They permit to explore the two basic aspects of mass and charge distribution of the molecules and allow the investigators to exploit different calculation options for both aspects. For example, since the treatment of the conformational freedom is critical, we decided to a) consider different conformers as different structures and b) use different alignment options (rigid; gradual exploration). All these options are described in detail in the Methods Section. It should be remarked that the availability of a wide range of calculation options (hence of similarity measures) is a bless and not a curse for the present work. Since our aim was to find an "absolute" definition of series independent from the individual similarity measures, the higher the number of available measures, the more refined the operational definition of series obtained through the PCA of the set of measures. This is equivalent to improving the estimation of a parameter by repeating the measures and then calculating its average value.

As stated above, the computation of similarity indices requires different choices: a) for the chemical physical

Table 2. Component Loadings for Data Set A[a]

metrics	PC1	PC2	PC3
Comb1	0.981	0.017	−0.122
Cha1	0.835	*0.496*	−0.181
Sha1	0.852	*−0.496*	−0.024
Comb2	0.596	0.167	0.782
Cha2	0.172	*0.492*	0.839
Sha2	0.831	*−0.462*	0.130
Comb3	0.981	0.131	−0.129
Cha3	0.753	*0.617*	−0.207
Sha3	0.825	*−0.562*	0.031
Comb4	0.973	0.150	−0.138
Cha4	0.712	*0.642*	−0.245
Sha4	0.802	*−0.580*	0.078
Comb5	0.980	−0.025	−0.111
Cha5	0.840	*0.475*	−0.174
Sha5	0.846	*−0.512*	−0.018
Comb6	0.690	0.109	0.711
Cha6	0.236	*0.530*	0.803
Sha6	0.840	*−0.477*	0.174
Comb7	0.989	0.075	−0.115
Cha7	0.767	*0.603*	−0.203
Sha7	0.824	*−0.561*	0.034
Comb8	0.983	0.110	−0.126
Cha8	0.713	*0.649*	−0.244
Sha8	0.807	*−0.576*	0.076
% expl. var.	65.7	20.5	12.0

[a] The variables are the Combined (COMB), Electrostatic Potential (CHA) and Shape (SHA) similarity indices relative to the lead molecule obtained in the different computational runs (see Table 1). The values in italics point to the electrostatic potential/shape opposition measured by Component 2.

features taken into consideration (shape, electrostatic potential) and b) for the structure optimization procedures. In this paper, a set of different options were considered, and several similarity matrices were computed accordingly. We generated a consensus similarity index in the form of the first principal component of the matrix having 1) as statistical units the molecules of the series and 2) as variables the different similarity values in respect to the seed, as computed by different choices of the alignment metrics. If (and only if) the first principal component had all positive loadings with the original variables, then it could be considered a consensus mono-dimensional metric.

Two sets of aromatic molecules were analyzed in this work (Figures 1 and 2). Each molecule of the set was indexed by 24 similarity scores with the relative series lead: this corresponded to the use of 8 different superposition procedures (Table 1), for 3 different kind of distances (electrostatic potential-based, shape-based and combined).

The data set A ($n = 41$) (Figure 1) gave rise to three components which explained 65.7, 20.5 and 12.03% of total variance (98.2% cumulative), while the "noise floor" started at component 4 (only 0.9% of total variance) (Table 2, Figure 3).

All the variables were positively loaded on PC1, that can thus be considered a "consensus" component summarizing—in the least squares optimal way—all the information common to the 24 similarity indices. This interpretation was further strengthened by the prominent loading of COMB variables on PC1.

PC2 represents the shape/electrostatic potential opposition. The shape similarities have a loading on this component with an inverse sign with respect to the electrostatic potential similarities and, consistently, the loading of their average

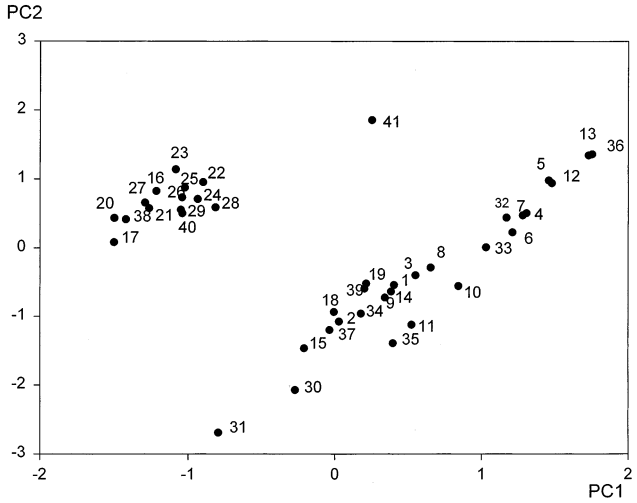

Figure 3. Space of the first two PCs computed over the 24 similarity indices for data set A.

(COMB) is very low in absolute value. Thus PC2 quantifies the lack of concordance between shape and electrostatic potential information (the COMB variable—average of shape and electrostatic potential—had no variance on this component, being "annihilated" by the two opposite signs of its shape and electrostatic potential components). In other words, PC2 quantifies the departure from the shape/electrostatic potential "consensus" axis represented by PC1.

PC3 has a direct, purely algorithmic explanation, due to the fact that the computational runs 2 and 6 were performed using a full torsional search to find the molecular superposition that gives the greatest similarity.

Based on the above considerations, the three components can be called consensus (PC1), distance from "series channel" (PC2) and singular methods (PC3) components.

More details are revealed by a closer inspection of the PC1/ PC2 space (Figure 3). For decreasing values of PC1 (right to left), the first molecules are single rings substituted with a methyl (Molecules 36 and 13), followed by a neat progressive increase in size and/or electrostatic potential asymmetry (with the possible exception of Molecule 41, phenyl-substituted). At the negative pole of PC1 are the substituted two-rings molecules. Thus, PC1 alone already scales the molecules in a way acceptable to a chemist's eye. However, Figure 2 also shows that PC2 adds a discontinuity to the chemical space and separates the series of the single ring molecules from the series of the two-rings molecules. Two parallel lines can be superimposed on the one-ring and two-rings molecules, respectively. Whereas the series of the single ring molecules is well drawn, the other series is quite "constricted", probably because the observation point (seed) chosen for the similarity analysis was the single ring seed. Thus, overall the ASP/PCA analysis was able to pick up chemically reasonable molecular series. In light of these observations, our interpretation of the consensus component PC1 can be better defined. Once the qualitative gap measured by PC2 is taken into account (separation of chemicals into two groups) and the best drawn series is considered (single rings), PC1 appears to be the quantitative counterpart of the common skeleton defining the series. In other words, it can be viewed as the mono-dimensional displacement of the elements of the set from the lead molecule along a common

Table 3. Component Loadings for Data Set B[a]

metrics	PC1	PC2	PC3
Comb1	0.969	−0.013	0.045
Cha1	0.866	*−0.424*	0.096
Sha1	0.741	*0.636*	−0.010
Comb2	0.910	−0.142	0.373
Cha2	0.828	*−0.394*	0.364
Sha2	0.778	*0.533*	0.261
Comb3	0.974	−0.154	−0.113
Cha3	0.904	*−0.386*	−0.135
Sha3	0.649	*0.749*	−0.030
Comb4	0.985	−0.075	−0.031
Cha4	0.892	*−0.409*	−0.005
Sha4	0.720	*0.661*	−0.080
Comb5	0.943	−0.111	0.278
Cha5	0.875	*−0.385*	0.224
Sha5	0.775	*0.531*	0.302
Comb6	0.980	−0.102	−0.155
Cha6	0.904	*−0.370*	−0.190
Sha6	0.637	*0.755*	0.038
Comb7	0.952	−0.118	−0.239
Cha7	0.899	*−0.346*	−0.228
Sha7	0.635	*0.737*	-0.150
Comb8	0.959	−0.158	−0.200
Cha8	0.906	*−0.363*	−0.178
Sha8	0.628	*0.727*	-0.174
% expl. var.	73.0	20.5	3.8

[a] The variables are the Combined (COMB), Electrostatic Potential (CHA) and Shape (SHA) similarity indices relative to the lead molecule obtained in the different computational runs (see Table 1). The values in italics point to the electrostatic potential/shape opposition measured by Component 2.

channel. The dimensional collapse along a single axis is due to the sharing of a common architecture that defines the series itself (in analogy with the -CH2- module of the aliphatic alkanes). We can easily define a series as a set of molecules that can be exhaustively described by only one dimension, independent of the numerosity of the original descriptors.

Thus, the analysis confirmed the value of our analogy, and pointed exactly to the elements we expected under our model: a "mono-dimensional series space" (PC1), a unique "displacement from the series" dimension (PC2) and an algorithmic singularity (PC3) axis. Another point of the analogy remained to be checked: the need for a specific seed to define the series. This corresponds to say that the above results should be strictly dependent on the specific seed (lead molecule) selected for the computation of similarities. This was actually the case: when we used Molecule 41 as seed, the above component structure was destroyed and no unambiguous consensus and "displacement from the series" components appeared (results not shown).

The analysis of series B ($n = 28$) (Figure 2) confirmed the previous results: PC1 and PC2 maintained exactly the same meaning of "consensus" and "displacement from the series", while PC3 pointed to algorithmic singularity, even though in this case this last element had a comparatively minor weight (Table 3, Figure 4).

CONCLUSIONS

The possibility of expanding the mathematical concept of series from the realm of aliphatic alkanes to sets of aromatic molecules for which a 3D similarity index can be computed was empirically demonstrated. This corresponds to the establishment of an unambiguous and operationally reliable

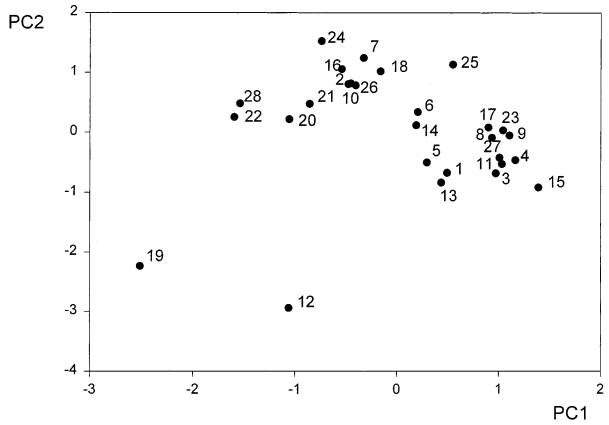

Figure 4. Space of the first two PCs computed over the 24 similarity indices for data set B.

concept of series in terms of mono-dimensional displacement from a lead molecule. This mono-dimensional displacement occurs along a "channel" defined by an invariant pattern of electrostatic potential and shape features of the molecules. The ASP/PCA procedure pointed to both the seed and the components of an aromatic series.

Obviously, this definition of series has a pure chemical physical character and in principle does not relate to any specific biological activity of the studied molecules. The biological phase is expected to superimpose its ordering rules on the underlying chemical physical features of the studied molecule, complex macromolecular receptors being the ultimate designers of the biological activity series. The advantage of the proposed method has to be seen in its ability to unambiguously decide about the assignment of molecules to a pre-defined chemical physical pattern represented by a lead chemical. If belonging to the same pattern (e.g. series measured according to the present ASP-based method) translates into similar biological and/or physical chemical properties has to be investigated on a case-by-case basis. Moreover, this method permits the exploration of the chemical spaces from a local fine grain perspective, thus complementing the coarse grain descriptions.

On a more general ground, a variety of approaches has been developed to estimate molecular similarity and correspondingly to characterize chemical diversity. The ASP approach used here consists of the superposition of pairs of molecules, and similarities in terms of steric and electronic distribution are estimated all over the two molecules. As remarked by Duca and Hopfinger,[14] the molecular similarity can be measured on a relative or an absolute basis. Relative similarity is dependent upon an external reference frame (e.g. an alignment constraint), while absolute similarity is constraint independent. In this sense, ASP is—by design—alignment dependent. However, the method can be used in different ways, and both relative and absolute similarity estimates can be derived. In the present case, we focused on the similarity of a group of molecules in respect to one selected pattern (molecule). The relative similarity estimates were suitable to highlight major or minor departures from the selected pattern, thus suggesting that the method can point to chemical series at a fine grain detail. However, the approach can be used also to derive absolute similarity measures. This can be done by calculating an N x N similarity matrix:[8,15] an appropriate analysis of the matrix

(e.g. through PCA) extracts the correlated information from the sum of local information, and can measure molecular properties, such as those classically used in QSAR,[15] and even more elusive characteristics, such as chirality.[16] In this case the passage from a relative to an absolute metric is accomplished by computing the average (principal components) features of the alignment over a multiplicity of patterns, thus virtually eliminating the relative character of ASP approach.

Overall this evidence points to the flexibility and value of the approach and encourages further investigations.

Andrea Rodomonte is thanked for helpful discussions and comments.

REFERENCES AND NOTES

(1) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.

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

(3) Kulkarni, A.; Hopfinger, A. J.; Osborne, R.; Bruner, L. H.; Thompson, E. D. Prediction of eye irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol. Sci.* **2001**, *59*, 335−345.

(4) ASP, Automated Similarity Package, previously Oxford Molecular Ltd., now Accelrys (www.accelrys.com).

(5) Fradera, X.; Amat, L.; Besalu, E.; Carbo-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct. −Act. Relat.* **1997**, *16*, 25−32.

(6) Carbo-Dorca, A. L.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169−5180.

(7) Bowen-Jenkins, P. E.; Richards, W. G. Quantitative measures of similarity between pharmacologically active compounds. *Intl. J. Quantum Chem.* **1986**, XXX, 763−768.

(8) 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.

(9) Robert, D.; Carbo-Dorca, R. A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 469−475.

(10) SYBYL 6.6 Tripos Associate Inc., 1699 South Hanley Road, Suite 303, St. Louis, Missouri, 63144-2913.

(11) Clark, M.; Cramer III, R. D.; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(12) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1979**, *36*, 3219−3288.

(13) Darroch, J.; Mosimann, J. E.; Canonical and principal components of shape. *Biometrika* **1985**, *72*, 241−252.

(14) Duca, J. S.; Hopfinger, A. J.; Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(15) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrixes and quantitative structure−activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629−635.

(16) Benigni, R.; Cotta-Ramusino, M.; Gallo, G.; Giorgi, F.; Giuliani, A.; Vari′, M. R. Deriving a quantitative chirality measure from molecular similarity matrices. *J. Med. Chem.* **2000**, *43*, 3699−3703.

CI020375Z