Analyzing the Triple Density Molecular Quantum Similarity Measures with the INDSCAL Model

David Robert and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, 17071 Girona, Catalonia, Spain

Received December 19, 1997

This work proposes a statistical interpretation of the Triple Density quantum similarity measures, and studies their simultaneous visualization by means of metric multidimensional scaling methods. The individual differences INDSCAL model is presented as a good mathematical tool to evidence the role of the molecule-operators. Finally, a practical example is discussed using a small set of molecules, consisting of 12 cancer chemotherapeutic agents.

1. INTRODUCTION

Molecular Quantum Similarity[1] is a technique developed in order to obtain new insight into the molecular similarity problem from a quantum mechanical point of view. The model is general enough to encompass several structural forms. The most important of these is based on a positive definite operator defining the quantum similarity measures. A possible family of quantum similarity measures, the so-called Triple Density Measures, was presented in a previous paper.[2] They could be considered as representations of molecular density functions in the basis set of the same densities. This first work was the starting point for the present one, in which Triple Density Measures are studied by means of the multivariate statistical INDSCAL model as possible different viewpoints of the same underlying configuration, giving the proximity relationships between the elements of the studied molecular set.

The paper is organized as follows: First, a definition of quantum similarity measures is given; the definition and interpretation of the Triple Density Measures are then discussed; next, the INDSCAL model is briefly described; and finally, an application to a concrete case of 12 cancer chemotherapeutic agents is studied.

2. QUANTUM SIMILARITY MEASURES

2.1. Quantum Similarity Background. Molecular Quantum Similarity was developed in order to obtain a formal comparison between the elements of a molecular set in terms of a quantum mechanical descriptor: their density functions. The quantum mechanical postulates assume that the wave function, and in some degree, the density function, contain all the information of a system.[3] Thus, the descriptor chosen to establish the comparison between quantum objects has to be the best source of information associated with a quantum system. In this sense, Molecular Quantum Similarity can be considered as a theoretical body constructed within the quantum mechanical formalism, based in particular on Löwdin and McWeeny's density function framework.[4]

2.2. General Definition. Given the first-order density functions of two molecular systems: $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the molecular coordinates, the general form of a quantum similarity measure can be defined as the integral:

$$Z_{AB}(\Omega) = \int\int \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \qquad (1)$$

where $\Omega(\mathbf{r}_1,\mathbf{r}_2)$ is a positive definite operator.

3. TRIPLE DENSITY QUANTUM SIMILARITY MEASURES

One possible form of the quantum similarity measure (1) can be constructed using the operator $\omega(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_2)$. The measure (1) is then transformed into:

$$\begin{aligned} Z_{AB}(\Omega) &= \int\int \rho_A(\mathbf{r}_1)\omega(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_2)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \\ &= \int \rho_A(\mathbf{r})\omega(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r} \\ &= \langle \rho_A|\omega|\rho_B\rangle \qquad (2) \end{aligned}$$

The positive definite operator $\omega(\mathbf{r})$ can be substituted by another appropriate first-order density function $\rho_C(\mathbf{r})$, that is

$$\begin{aligned} Z_{AB;C} &= \int \rho_A(\mathbf{r})\rho_C(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r} \\ &= \langle \rho_A|\rho_C|\rho_B\rangle \qquad (3) \end{aligned}$$

This type of quantum similarity measure is known as Triple Density Measure.

4. INTERPRETATIONS OF THE TRIPLE DENSITY SIMILARITY MEASURES

Carbó et al.[2] considered the Triple Density Measures (3) as matrix element representations of the density function ρ_C, taken as an operator, in a basis set of the density functions where the couple $\{\rho_A, \rho_B\}$ belongs. Therefore, each element of a known set of molecules can be represented through the Triple Density Measures (3) using in turn all the elements of the set as a basis. In this way one can construct a symmetric matrix set, $\mathbf{Z}_I = \{Z_{RS;I}\}$, connected to each element belonging to a given density set. Every one of such matrixes constitutes the representation of the Ith density function of

S0095-2338(97)00121-2 CCC: $15.00
Published on Web 06/03/1998

the set with respect to all the possible density function pairs, including itself. From an algebraic point of view, the density representation matrixes can be viewed as projections of the density functions from the infinite dimensional space, where they belong to a finite dimensional Euclidean space.

Going further in this interpretation, and assuming the uniqueness in the proximity between the different molecules of the set, one could then consider that the matrix elements of the matrixes $\boldsymbol{Z}_I$ (giving the similarity relationships between all the elements of the set) are nothing but different perceptions of these proximities, *seen* in a different way by the different molecule-operators. Thus, the matrixes $\boldsymbol{Z}_I$ would be *different viewpoints of the same underlying situation*. All of the possible proximity information patterns designable by the usual methods[5] will then be distorted patterns, in the sense that they will have a subjective tendency toward the density functions that work as operators. The above considerations transform these matrixes (a particular case of *three-way, two-mode* statistical data) into variables of a well-known problem in multivariate statistics: the individual differences problem.[5a]

5. INDIVIDUAL DIFFERENCES INDSCAL MODEL

Usual three-way, two-mode data refer to the dissimilarities (or more general proximities) between a set of objects assigned by another set of individuals or *judges*. In our case objects and subjects become blurred because they are represented by analogous mathematical tools, the molecular density functions, as a consequence of their dual character: they can work as functions and as quantum operators. In the matrix element $Z_{RS;I}$, R and S are the molecules compared, and I is the molecule that plays the role of the judge. To obtain a perfect agreement in the application of the model, the Triple Density similarity measures will be transformed into Euclidean distances (a particular case of dissimilarities) using the expression:[1a]

$$\delta_{RS;I} = (Z_{RR;I} + Z_{SS;I} - 2Z_{RS;I})^{1/2} \qquad (4)$$

In the early work on individual differences there were two basic models to represent three-way, two-mode data. The first model was based on averaging over individuals, but a great amount of information was lost in this framework. The second method was based on the comparison of the results individual by individual. Obviously, this could become a particularly difficult task when the number of elements increases. Tucker and Messick[6] proposed the first method to give a pictorial representation of the overall relationships between the objects, and also to quantify the differences between the individuals.

We will follow the method developed by Carroll and Chang.[7] Their premise was that there exists an underlying space, the ‘group stimulus’ space (that we will call here the *objects space*), in which systematic differences arise between individuals because they perceive this space in different ways. In particular, there exist a fixed set of reference axes in the space such that all differences between individuals can be explained by assuming that different individuals attach different ‘weights’ to these axes. In other words, each individual identifies the same underlying sources of variation among the objects, but the individuals differ in the relative importance they attach to each of these sources. These axes constitute the ‘subjects’ space (called here the *operators space*). Both spaces will be chosen to have dimension p. Points in the objects space represent the compared molecular set, and constitute the underlying configuration. The molecule-operators are represented as points in the operators space, the *point-operators*. The coordinates of each point-operator are the weights required to give the weighted Euclidean distances between the points in the objects space, the values that best represent the corresponding dissimilarities for that molecule-operator.

Let the points in the objects space be given by x_{RT} ($R = 1,..., n$; $T = 1,..., p$), and the points in the operators space have coordinates w_{IT} ($I = 1,..., n$; $T = 1,..., p$). Then the weighted Euclidean distance between molecule R and S, for the Ith molecule-operator is defined by:

$$d_{RS;I} = \{\sum_{T=1}^{p} w_{IT}(x_{RT} - x_{ST})^2\}^{1/2} \qquad (5)$$

The molecule-operator weights $\{w_{IT}\}$ and molecule coordinates $\{x_{RT}\}$ are then sought as those that best match $\{d_{RS;I}\}$ to the original $\{\delta_{RS;I}\}$ values. Using metric classical scaling[5a−b] one can transform dissimilarities $\{\delta_{RS;I}\}$ into distance estimates $\{d_{RS;I}\}$, and estimation of weights and coordinates is then made by means of a least-squares algorithm. The distances associated with each individual are given by the matrixes $\mathbf{B}_I$, where

$$[\mathbf{B}_I]_{RS} = b_{RS;I} = \sum_{T=1}^{p} w_{IT} x_{RT} x_{ST} \qquad (6)$$

Least-squares estimates of $\{w_{IT}\}$ and $\{x_{RT}\}$ are then found by minimizing the function

$$S = \sum_{R,S,I} (b_{RS;I} - \sum_{T=1}^{p} w_{IT} x_{RT} x_{ST})^2 \qquad (7)$$

Other individual differences models exist,[8] but they will not be used here.

6. CANCER CHEMOTHERAPEUTIC AGENTS: A PRACTICAL EXAMPLE

A small set of molecules (Table 1), consisting of 12 cancer chemotherapeutic agents of the family of camptothecin[9] (CPT) (Figure 1), is used to apply the INDSCAL model as described in the previous section. These camptothecin analogues have been chosen with 20*(S)* stereochemistry because of their high activity levels with regard to those possessing 20*(RS)* and 20*(R)* configurations.[10,11] Triple Density quantum similarity measures have been calculated using the Atomic Shell Approximation (ASA) fitted densities,[12] and full geometry has been optimized.[13] Identification numbers shown in Table 1 will be used when referring to any compound studied here.

6.1. The Objects Space. The representation of the molecules in the objects space (Figure 2) shows the underlying configuration of the set once the differences between the viewpoints have been removed. It can be seen that a clear clustering arises. The five groups found are the following: {**28**, **6**}, {**30**}, {**16**, **17**}, {**13**}, and {**18**, **3**, **9**, **10**,

Table 1. Cancer Chemotherapeutic Agents Studied and Their Topoisomerase 1 Inhibition Activities[9] [a]

compound	IC_{50} (μM)
10,11-methylenedioxy-20(*S*)-CPT **(28)**	0.027
9-methyl-20(*S*)-CPT **(18)**	0.038
9-amino-10,11-methylenedioxy-20(*S*)-CPT **(6)**	0.048
9-chloro-10,11-methylenedioxy-20(*S*)-CPT **(30)**	0.061
9-chloro-20(*S*)-CPT **(16)**	0.086
10-hydroxy-20(*S*)-CPT **(3)**	0.106
9-amino-20(*S*)-CPT **(9)**	0.111
10-amino-20(*S*)-CPT **(10)**	0.140
10-chloro-20(*S*)-CPT **(17)**	0.141
10-nitro-20(*S*)-CPT **(13)**	0.635
20(*S*)-CPT **(1)**	0.677
9-hydroxy-20(*S*)-CPT **(15)**	0.873

[a] The identification number for each compound appears in parentheses, following the notation of ref 9. IC_{50} is the minimum drug concentration (μM) that inhibited the cleavable complex formation by 50%.

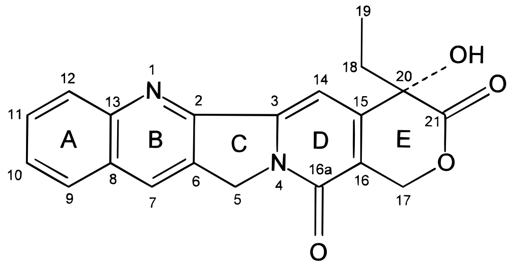

Figure 1. Camptothecin (CPT) structure.

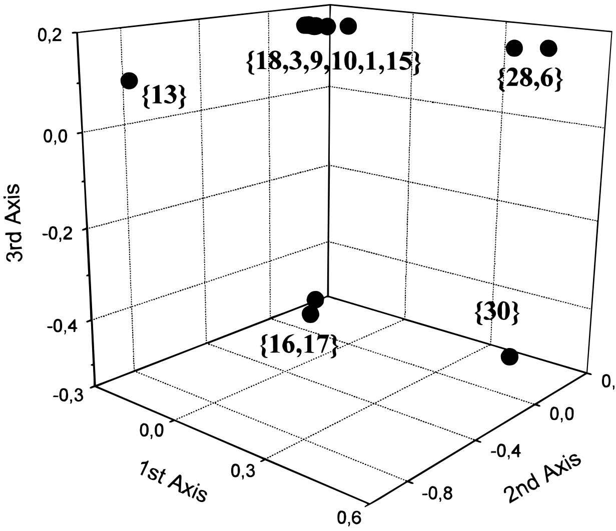

Figure 2. Molecule configuration in the objects space using the first three coordinates. The identification number of the molecules belonging to each cluster appears in brackets.

1, **15**}. The similarity between the different molecules is strongly dependent on their structure, and in particular, on the type of substituents they possess. By analyzing the configuration obtained it can be seen that cluster {**28**, **6**} corresponds to the compounds containing the 10,11-methylenedioxy ring, with high in vivo (life-prolongation activity in mice L1210 leukemia) and in vitro (topoisomerase 1, T-1, inhibition) activities.[10] Compound **30** also contains this substituent group (and is also a potent T-1 inhibitor[11a]), but the presence of a chlorine atom determines its location. The projections onto the first and second axes for these three compounds are very similar, and the third axis distinguishes between the two clusters. The distinctive character of the axes is confirmed by the existence of a cluster formed by compounds **16** and **17**, the chlorine substituted CPTs. Both molecules only differentiate from each other in the substituent location (**16** at C-9 and **17** at C-10), and they have a similar property: a high activity in the T-1 inhibition assay.[10] These two molecules, together with compound **30**, have a similar projection onto the third axis. The existence of the heavy NO_2 radical in the structure of agent **13** seems to determine its isolated location. Finally, the more numerous cluster, formed by the agents {**18**, **3**, **9**, **10**, **1**, **15**}, is constituted by the rest of the molecules, which have light substituents (NH_2, CH_3, OH). All of this information can be summarized by an *a posteriori* interpretation of the axes: the first axis distinguishes between the compounds that have the 10,11-methylenedioxy group and those that do not; the second axis separates the heavy and the light substituents, and the third axis discriminates between the agents with chlorine atoms. It also must be noted that activity differences between two molecules cannot be distinguished with this methodology when their structure is very similar. For instance, compounds **15** and **3** (hydroxyl substituted CPTs at C-9 and C-10 positions, respectively) have been grouped together even though they present considerable differences in in vivo activities.[10] It is possible that another axis would be necessary to distinguish the subdivisions in this last cluster. In any case, the detailed comparative discussion about the CPT analogues is not the main objective of this work.

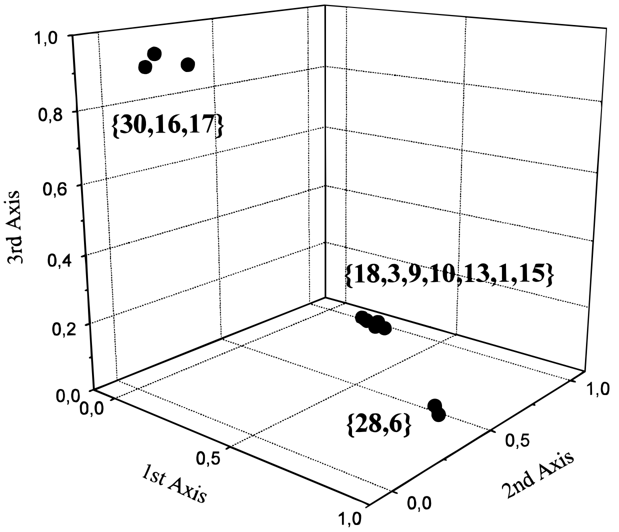

Figure 3. Molecule-operators configuration in the operators space using the first three coordinates. The identification number of the molecule-operators for each cluster appears in brackets.

Distortion in the configurations derived from Triple Density Measures is a real fact that can be evidenced by representing each Triple Density matrix by means of classical scaling. Thus, some compounds (**3**, **9**, **10**, and **15**) are located in an isolated position in the similarity space when they work as operators, although they have notable structural coincidences.

6.2. The Operators Space. On the other hand, the configuration of the molecule-operators in the operators space (Figure 3) has a different interpretation, namely the proximity between points indicates proximity in the viewpoints of the operators they represent. Thus, the groups found in Figure 3 are molecular clusters that attach the same importance to the underlying sources of variation between the molecules compared. Three clusters are clearly obtained in this representation: the first containing the operators corresponding to the molecules {**30**, **16**, **17**}; the second containing the operators {**18**, **3**, **9**, **10**, **13**, **1**, **15**}; and the third containing the operators {**28**, **6**}. The clustering of the operators is again related to the structural characteristics of the operator-molecules, but in a less appreciable way. Thus, the first group is formed by the molecules with chlorine atoms (note that the third axis again separates chlorine substitutions from the rest), the third group is formed by the molecules that have the 10,11-methylenedioxy ring (except for molecule **30**, which contains this substituent group but also a chlorine atom), and the numerous second group is formed by the rest of the molecules.

These results suggest a considerable reduction of the system's degrees of freedom, in the sense that the Triple Density Measures contain redundant information which can be eliminated. The information that the Triple Density Measures can produce because of the existence of different perceptions in the molecular similarity relationships is limited by the existence of groups in the operators space. Thus, from all the molecule-operators, only those which do not belong to a same cluster in the operators space produce new information. It is reasonable, therefore, to choose a *class representative* for each cluster, defined, for instance, as the closest operator to the centroid of the cluster where it belongs. In the present case, class representatives for the obtained clusters are compound **16** for cluster 1, compound **10** for cluster 2, and either compounds **28** or **6** for cluster 3.

7. CONCLUSIONS

A new interpretation for the Triple Density quantum similarity measures has been presented. The structure of the Triple Density Measures provides a natural way to obtain different perceptions of the same underlying molecular resemblance relationships. Thus, each Triple Density Measures matrix gives the finite representation of each molecule on a basis set formed by the overall molecular set (including itself), and the underlying configuration shows a more objective view of the similarity relationships between the molecular set. Also it has been demonstrated that the representation of the operators obtained with the individual differences model employed (Carroll and Chang's INDSCAL model) is a good mathematical tool to eliminate the redundant information contained in the Triple Density Measures, and to reduce the degrees of freedom of the system studied.

It must be noted that this interpretation does not end with Triple Density Measures. Any other type of quantum similarity measure[14] could also be thought of as a distortion of a hypothetic underlying molecular distribution. However, the Triple Density Measures, because of their evident symmetry, can be studied separately from other types of quantum similarity measures. For instance, relationships found between cluster members and the structure of the molecule-operators can no longer be maintained if other quantum similarity measures are included.

ACKNOWLEDGMENT

This work has been partially sponsored by the project SAF 96-0158 of the CICYT. The authors gratefully acknowledge Lluis Amat for providing the Triple Density matrixes used in this paper.

REFERENCES AND NOTES

(1) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189. (b) Carbó, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517−545. (c) Carbó, R.; Calabuig, B. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A. Maggiora, G. M., Eds.; Wiley: New York, 1990. (d) Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and N-Dimensional Representation of Quantum Objects. I. Theoretical Foundations. *Int. J. Quantum Chem.* **1992**, *42*, 1681−1693.

(2) Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: A General Connection Between Theoretical Calculations and Experimental Results. *Mol. Eng.* **1992**, *2*, 43−64.

(3) See for example: (a) Von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Princeton University Press: Princeton, 1955. (b) Bohm, D. *Quantum Theory*; Dover Publications Inc.: Mineola, NY, 1989.

(4) (a) Löwdin, P. O. *Phys. Rev.* **1955**, *97*, 1474. (b) McWeeny, R. *Proc. R. Soc.* **1955**, *A232*, 114.

(5) (a) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall: London, 1994. (b) Krzanowski, W. J.; Marriott, F. H. C. *Multivariate Analysis*; Edward Arnold: London, 1994; Vols. 1 and 2. (c) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990. (d) Tou, J. T.; González, R. C. *Pattern Recognition Principles*; Addison-Wesley: Reading, MA, 1974.

(6) Tucker, L. R.; Messick, S. An individual differences model for multidimensional scaling. *Psychometrika* **1963**, *28*, 333−367.

(7) Carroll, J. D.; Chang, J. J. Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart-Young" decomposition. *Psychometrika* **1970**, *35*, 283−319.

(8) (a) Carroll, J. D.; Chang, J. J. IDIOSCAL (Individual Differences in Orientation Scaling): a generalization of INDSCAL allowing idiosyncratic references systems. Psychometric Meeting, Princeton, NJ, 1972. (b) Borg, I. Geometric representation of individual differences. In *Geometric Representations of Relational Data*; Lingoes, J. C., Ed.; Mathesis: Ann Arbor: MI, 1977. (c) Winsberg, S.; De Soete, G. A latent class approach to fitting the weighted Euclidean model. *Psychometrika* **1993**, *54*, 217−229. (d) Cuadras, C. M.; Fortiana, J. Visualizing categorical data with related metric scaling. In *Visualization of Categorical Data*; Blasius, J., Greenacre, M., Eds.; Academic Press: New York, 1997, in press.

(9) Wall, M. E.; Wani, M. C.; Cook, C. E.; Palmer, K. H.; McPhail, A. T.; Sim, G. A. Plant antitumor agents. I. The isolation and structure of camptothecin, a novel alkaloidal leukemia and tumor inhibitor from camptotheca acuminata. *J. Am. Chem. Soc.* **1966**, *88*, 3888−3890.

(10) Wall, M. E.; Wani, M. C. Camptothecin and Analogues. In *Cancer Chemotherapeutic Agents*; Faye, W. O., Ed.; ACS Professional Reference Book: Washington, DC, 1995.

(11) (a) Wall, M. E.; Wani, M. E.; Wani, M. C.; Nicholas, A. W.; Manikumar, G.; Moore, L.; Truesdale, A.; Leitner, P.; Besterman, J. M. *J. Med. Chem.* **1993**, *36*, 2689−2700. (b) Wani, M. C.; Nicholas, A. W.; Wall, M. E. *J. Med. Chem.* **1987**, *30*, 2317−2319. (c) Jaxel, C.; Kohn, K. W.; Wani, M. C.; Wall, M. E.; Pommier, Y. Structure-activity study of the actions of camptothecin derivatives on mammalian topoisomerase I: evidence for a specific receptor site and a relation to antitumor activity. *Cancer Res.* **1989**, *49*, 1465−1469.

(12) (a) Constans, P.; Amat, Ll.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT; Vol. 1, pp. 187−211. (b) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053. (c) Amat, Ll.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting Using Elementary Jacobi Rotations. *J. Comput. Chem.*, in press.

(13) Amat, Ll.; Besalú, E.; Carbó-Dorca, R. A validation study of Tuned QSAR models. Institute of Computational Chemistry Technical Report, IT-IQC 01/97(P).

(14) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M.; Foundations and Recent Developments on Molecular Quantum Similarity. In *Topics in Current Chemistry*; Sen, K., Ed.; Springer-Verlag: Berlin, 1995.

CI970121R