# Systematic Study of the Quality of Various Quantum Similarity Descriptors. Use of the Autocorrelation Function and Principal Component Analysis

**Greet Boon,[†] Wilfried Langenaeker,[‡] Frank De Proft,[†] Hans De Winter,[‡] Jan P. Tollenaere,[‡] and Paul Geerlings*[,†]**

*Eenheid Algemene Chemie, Vrije Universiteit Brussel, Fakulteit Wetenschappen, Pleinlaan 2, B -1050 Brussels, Belgium, and Janssen Research Foundation, Theoretical Medicinal Chemistry, Turnhoutseweg 30, B-2340 Beerse, Belgium*

*Received: April 18, 2001; In Final Form: June 25, 2001*

The quality of various quantum similarity descriptors is studied systematically for a series of peptide isosteres important in pharmacology. To cope with the drawbacks of the Carbó and Hodgkin-Richards type indices, those being the time-consuming three-dimensional integration and the importance of the relative orientation, position, and conformation of the molecules, the use of the autocorrelation function is investigated. In combination with the application of principal component analysis, this approach is proven to be a more practical and faster tool for generating similarity sequences.

## 1. Introduction

The concept of similarity is of fundamental importance in chemistry and pharmacology. To characterize similarity, a variety of similarity indices has been proposed in the literature. Of particular importance are the so-called quantum similarity indices presented by Carbó and by Hodgkin and Richards, being purely quantum mechanical in their origin and being based on the electron density.[1,2]

Recently,[3] we proposed a reactivity-based quantum mechanical similarity index similar to those of Carbó and Hodgkin-Richards but based on the local softness[4] proven, among others, by us to be a valuable reactivity indicator within the context of Pearson's HSAB principle.[5−7] The major drawbacks of this approach were found to be the time-consuming integration over the three-dimensional space and the importance of the relative orientation, position, and conformation of the molecules under consideration. The aim of the study presented here is to investigate the use of the autocorrelation function (vide infra) as a means of coping with these drawbacks. Starting from the electron density, we can distinguish several different ways, schematically represented in Figure 1, of generating similarity sequences using this autocorrelation function. As shown in this figure, the calculated autocorrelation functions can directly be used as the basic quantity in the Carbó and Hodgkin-Richards type indices, replacing the density or the local softness, leading to a similarity value. As an alternative, they can also be subjected to a principal component analysis (PCA)[8] or the Euclidean distance $D$ can be calculated (vide infra).

In this paper, the quality of the results of these different approaches to molecular similarity is studied systematically for the series of peptide isosteres used in ref 3 and which was extended for this study. Isosteric replacement of the peptide bond is an attractive strategy for circumventing the well-known susceptibility of peptide bonds toward hydrolysis.[9]

* Corresponding author. E-mail: pgeerlin@vub.ac.be.
[†] Vrije Universiteit Brussel.
[‡] Jausseu Research Foundation, Theoretical Medicinal Chemistry.

## 2. Theory and Computational Details

**2.1. Similarity Indices.** In this work, molecular similarities were calculated by using the Carbó index:[1]

$$R_{AB} = \frac{\int \rho_A(\vec{r})\rho_B(\vec{r})\, d\vec{r}}{\{[\int \rho_A{}^2(\vec{r})\, d\vec{r}][\int \rho_B{}^2(\vec{r})\, d\vec{r}]\}^{1/2}} \quad (2.1.1)$$

and the Hodgkin-Richards index[2]:

$$H_{AB} = \frac{2\int \rho_A(\vec{r})\rho_B(\vec{r})\, d\vec{r}}{[\int \rho_A{}^2(\vec{r})\, d\vec{r} + \int \rho_B{}^2(\vec{r})\, d\vec{r}]} \quad (2.1.2)$$

where $\rho_A(\vec{r})$ and $\rho_B(\vec{r})$ are the electron densities of molecules A and B, respectively, the molecules being considered. These two, electron density-based, indices describe the similarity of shape and the extent of the electron distributions.

In our previous paper,[3] we introduced a similarity index, which is more directly related to reactivity and which is based on the local softness[4]

$$s(\vec{r}) = \left[\frac{\partial \rho(\vec{r})}{\partial \mu}\right]_{v(\vec{r})} \quad (2.1.3)$$

$$R_{AB}{}^s = \frac{\int s_A(\vec{r}) s_B(\vec{r})\, d\vec{r}}{\{[\int s_A{}^2(\vec{r})\, d\vec{r}][\int s_B{}^2(\vec{r})\, d\vec{r}]\}^{1/2}} \quad (2.1.4)$$

$$R_{AB}{}^s = \frac{2\int s_A(\vec{r}) s_B(\vec{r})\, d\vec{r}}{\int s_A{}^2(\vec{r})\, d\vec{r} + \int s_B{}^2(\vec{r})\, d\vec{r}} \quad (2.1.5)$$

where $s_A(\vec{r})$ and $s_B(\vec{r})$ the local softness of molecules A and B, respectively.

Expression 2.1.3 for the local softness can be rewritten as

$$s(\vec{r}) = \left[\frac{\partial \rho(\vec{r})}{\partial N}\right]_{v(\vec{r})}\left[\frac{\partial N}{\partial \mu}\right]_{v(\vec{r})} \quad (2.1.6)$$

yielding

10.1021/jp011441n CCC: $20.00 © 2001 American Chemical Society
Published on Web 08/18/2001

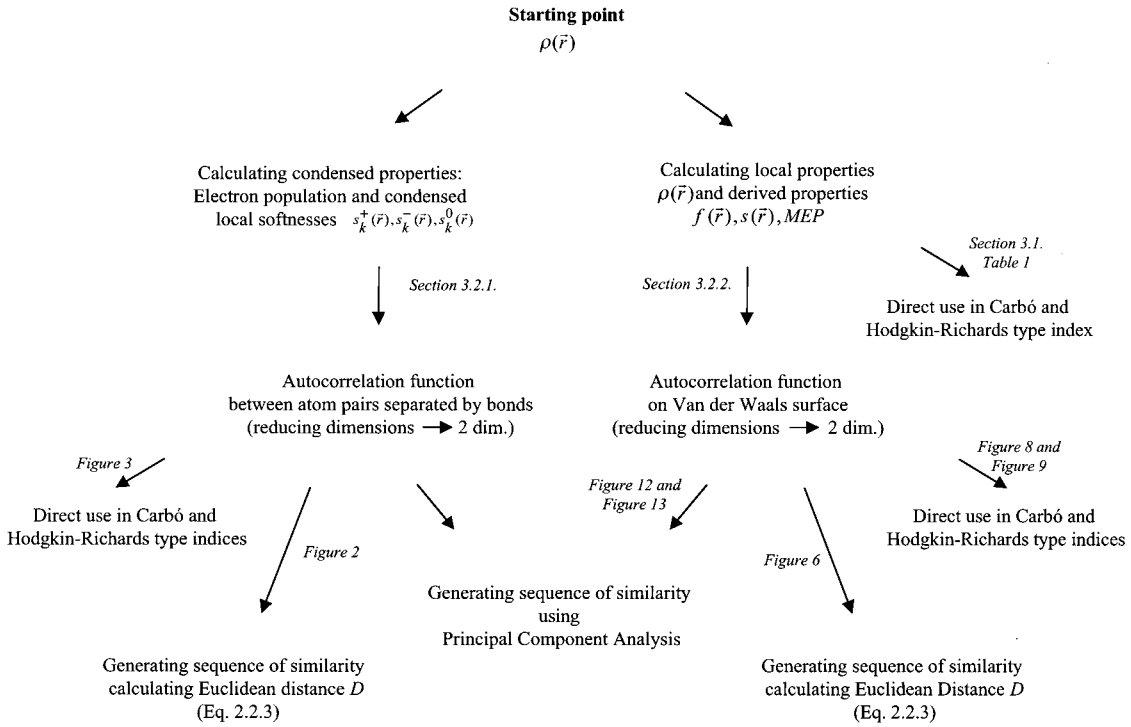


**Figure 1.** Schematic overview of the hierarchy of the calculations performed in this study.

$$s(\vec{r}) = f(\vec{r})S \tag{2.1.7}$$

where $S$ is the global softness and $f(\vec{r})$ is the Fukui function,[10] a frontier molecular orbital reactivity index, defined by Parr and Yang.

Due to the discontinuity of the first derivative in eq 2.1.6 with respect to the number of electrons $N$, the following three functions can be defined in a finite difference approximation:

$$f^+(\vec{r}) \approx \rho_{N_0+1} - \rho_{N_0} \tag{2.1.8}$$

$$f^-(\vec{r}) \approx \rho_{N_0} - \rho_{N_0-1} \tag{2.1.9}$$

$$f^0(\vec{r}) \approx \frac{1}{2}(\rho_{N_0+1} - \rho_{N_0-1}) \tag{2.1.10}$$

where $\rho_{N_0}$, $\rho_{N_0+1}$, and $\rho_{N_0-1}$ are the electron densities of the systems with $N_0$, $N_0 + 1$, and $N_0 - 1$ electrons, respectively. For each Fukui function, a product for the local softness $s^+(\vec{r})$, $s^-(\vec{r})$, and $s^0(\vec{r})$ analogous to eq 2.1.7 can be written which can be used in eqs 2.1.4 and 2.1.5 to describe the similarity for a nucleophilic, an electrophilic, and a radical reaction, respectively.

In the calculations of the indices based on the local softness, the following expression for the global softness $S$ is used assuming a quadratic relationship between the energy $E$ and the number of electrons $N$ of the system under consideration and using a finite difference approximation:[11]

$$S = \frac{1}{\mathrm{IE} - \mathrm{EA}} \tag{2.1.11}$$

where IE and EA indicate the ionization energy and electron affinity, respectively.

For an explanation of the used method of numerical integration, we refer to ref 3.

In this work, we will also make use of the condensed forms[4,12] of the local softness based on a Mulliken population analysis:[13]

$$s_k^+ = S[q_k(N+1) - q_k(N)] = Sf_k^+ \tag{2.1.12}$$

$$s_k^- = S[q_k(N) - q_k(N-1)] = Sf_k^- \tag{2.1.13}$$

$$s_k^0 = S\frac{[q_k(N+1) - q_k(N-1)]}{2} = Sf_k^0 \tag{2.1.14}$$

where $q_k(N)$ represents the electronic population on atom $k$ obtained by subtracting the charge on atom $k$ from the atomic number $Z_k$.

Besides using this condensed form of the local softness $s_k(\vec{r})$, it is also recommended to use the electron population $q_k(N)$ on atom $k$ being the condensed form of the electron density $\rho(\vec{r})$.

**2.2. Autocorrelation Function.** As mentioned earlier, using the Carbó and Hodgkin-Richards type indices has two major drawbacks, those being the time-consuming three-dimensional integration and the search for the best relative orientation and position of the considered molecules to obtain the maximum similarity value. Furthermore, conformational flexibility has to be taken into account. A way of coping with these drawbacks is using the autocorrelation function. The consideration of the spatial distribution of a certain property in an area consisting of several distinct regions is often recommended. If interdependence exists between the presence of that property in one region and the presence of it in a neighboring region, the data exhibit spatial autocorrelation.[14,15]

Several applications of the use of autocorrelation functions in molecular modeling and quantitative structure−activity relationships (QSARs) have been published. Moreau and Broto[16] first applied an autocorrelation function to the topology of molecular structures:

$$A_p(d) = \sum_{ij} p_i p_j \tag{2.2.1}$$

where $A_p(d)$ is the autocorrelation coefficient referring to the $i$ and $j$ atom pair separated by $d$ bonds and $p_i$ is an atomic

**TABLE 1: Calculated Similarity Indices $R_{AB}$ and $H_{AB}$ for a Series of Peptide Isosteres with respect to *N*-Methylacetamide**

| | | $R_{AB}$ | | | $H_{AB}$ |
|---|---|---|---|---|---|
| 1 | *N*,*N*-ethylmethylamine | 0.653 | 1 | *N*,*N*-ethylmethylamine | 0.644 |
| 2 | propene | 0.629 | 2 | propene | 0.597 |
| 3 | ethene | 0.617 | 3 | 2-methyl-2-butene | 0.585 |
| 4 | 2-methyl-2-butene | 0.591 | 4 | *trans*-2-butene | 0.569 |
| 5 | *trans*-2-butene | 0.582 | 5 | ethene | 0.556 |
| 6 | dimethyl-2-butene | 0.538 | 6 | dimethyl-2-butene | 0.537 |
| 7 | butanone | 0.531 | 7 | butanone | 0.530 |
| 8 | *cis*-2-butene | 0.506 | 8 | *cis*-2-butene | 0.499 |
| 9 | butane | 0.466 | 9 | butane | 0.444 |
| 10 | (*Z*)-2-fluoro-2-butene | 0.418 | 10 | (*Z*)-2-fluoro-2-butene | 0.404 |
| 11 | (*Z*)-2-chloro-2-butene | 0.134 | 11 | ethyl methyl thioether | 0.028 |
| 12 | ethyl methyl thioether | 0.050 | 12 | (*Z*)-2-chloro-2-butene | 0.007 |
| | | $R^{s_0}_{AB}$ | | | $H^{s_0}_{AB}$ |
| 1 | *N*,*N*-ethylmethylamine | 0.769 | 1 | *N*,*N*-ethylmethylamine | 0.761 |
| 2 | (*Z*)-2-fluoro-2-butene | 0.635 | 2 | (*Z*)-2-fluoro-2-butene | 0.614 |
| 3 | (*Z*)-2-chloro-2-butene | 0.614 | 3 | (*Z*)-2-chloro-2-butene | 0.599 |
| 4 | *trans*-2-butene | 0.576 | 4 | *trans*-2-butene | 0.556 |
| 5 | propene | 0.563 | 5 | propene | 0.545 |
| 6 | 2-methyl-2-butene | 0.539 | 6 | 2-methyl-2-butene | 0.517 |
| 7 | dimethyl-2-butene | 0.525 | 7 | dimethyl-2-butene | 0.501 |
| 8 | *cis*-2-butene | 0.501 | 8 | *cis*-2-butene | 0.483 |
| 9 | ethene | 0.496 | 9 | ethene | 0.476 |
| 10 | butane | 0.495 | 10 | butane | 0.411 |
| 11 | ethyl methyl thioether | 0.405 | 11 | ethyl methyl thioether | 0.405 |
| 12 | butanone | 0.108 | 12 | butanone | 0.101 |

property. Thus, a so-called autocorrelation vector is obtained as a series of coefficients for different topological distances. These autocorrelation vectors have some useful properties. First and foremost and as already mentioned in the Introduction, they are independent of the orientation of the molecules. Second, the autocorrelation coefficients are canonical; i.e., they are independent of the original numbering of the atoms. And finally, the length of the vector is independent of the size of the molecule.[15] In QSAR studies, the topological autocorrelation vectors were used as molecular descriptors.[17,18]

Gasteiger and co-workers[15] extended this concept to the spatial autocorrelation of molecular surface properties by introducing a three-dimensional (3D) descriptor based on the autocorrelation of properties at distinct points on the molecular surface. Here, the distances between surface points are sorted into preset intervals $[d_{lower}, d_{upper}]$:

$$A_p(d_{lower}, d_{upper}) = \frac{1}{L}\sum_{ij} p_i p_j \; (d_{lower} < d_{ij} < d_{upper}) \qquad (2.2.2)$$

where $A_p(d_{lower}, d_{upper})$ is the autocorrelation coefficient and $p$ is the property value at points $i$ and $j$ having a distance $d_{ij}$ belonging to the distance interval $[d_{lower}, d_{upper}]$. The sum is weighted by the total number $L$ of distances in the interval yielding a vector of autocorrelation coefficients for a series of distance intervals with different lower and upper bounds $d_{lower}$ and $d_{upper}$, respectively.

Both types of autocorrelation functions were considered in this study. For calculating the autocorrelation function between atom pairs separated by bonds, we used as atomic property $p$ in eq 2.2.1 the electron population on each atom obtained using the Mulliken charges[13] and, resulting from this electron population analysis, the condensed local softness $s_k$.[12]

As a molecular surface, we opted for using the van der Waals surface describing the valence region of the atoms, which essentially determines their reactivity. For calculating the autocorrelation function of molecular properties on the van der Waals surface of the considered molecules, we used as $p$ in eq 2.2.2 the molecular electrostatic potential (MEP)[19−21] already used in similarity studies by Hodgkin and Richards,[2] the electron density $\rho(\vec{r})$, and the local softness $s(\vec{r})$.

**TABLE 2: Sequence of Similarity and Calculated Euclidean Distances *D* Based on the Calculation of the Autocorrelation Function for Atom Pairs with the Electron Population Taken as the Atomic Property**

| | electron population | distance *D* |
|---|---|---|
| 1 | (*Z*)-2-fluoro-2-butene | 13.57 |
| 2 | butanone | 16.05 |
| 3 | 2-methyl-2-butene | 35.20 |
| 4 | ethyl methyl thioether | 100.42 |
| 5 | butane | 108.73 |
| 6 | *N*,*N*-ethylmethylamine (a) | 112.13 |
| 7 | *N*,*N*-ethylmethylamine (b) | 112.20 |
| 8 | (*Z*)-2-chloro-2-butene | 137.22 |
| 9 | *cis*-2-butene | 142.37 |
| 10 | *trans*-2-butene | 142.47 |
| 11 | dimethyl-2-butene (a) | 180.24 |
| 12 | dimethyl-2-butene (b) | 180.29 |
| 13 | dimethyl-2-butene (c) | 180.37 |
| 14 | dimethyl-2-butene (d) | 180.39 |
| 15 | propene | 239.80 |
| 16 | ethene | 317.62 |

To evaluate similarity using the autocorrelation function, three methods can be used. First, we can calculate for each distance $d$ the Euclidean distance $D$ between the autocorrelation function $A(d)$ of the reference molecule for a given property and the other molecule $j$ under consideration. Therefore, the following expression is used:

$$D_j = \sqrt{\{\sum_d [A_{ref}(d) - A_j(d)]^2\}} \qquad (2.2.3)$$

The smaller the Euclidean distance $D$, the higher the degree of similarity between the two molecules.

Second, the calculated autocorrelation functions can directly be used in Carbó and Hodgkin-Richards type indices replacing the density $\rho(\vec{r})$ or the local softness $s(\vec{r})$ and replacing the integrals with summations:

$$R_{AB} = \frac{\sum_d A_A(d) A_B(d)}{\sqrt{[\sum_d A_A^2(d)][\sum_d A_B^2(d)]}} \qquad (2.2.4)$$

$$H_{AB} = \frac{2\sum_d A_A(d) A_B(d)}{\sum_d A_A^2(d) + \sum_d A_B^2(d)} \qquad (2.2.5)$$

where $A_A(d)$ and $A_B(d)$ are the autocorrelation functions for each distance $d$ of the reference molecule e.g., A and the molecule B, under consideration.

Finally, the calculated autocorrelation functions will be subjected to principal component analysis (PCA) discussed in the following section to gain more insight into the structure of this data set.

The list of distinct points on the van der Waals surface was generated using an algorithm of Connolly.[22] To calculate the electron density on this generated list of points, we used the program Morphy[23] running on the COMPAQ-DIGITAL Alphaserver DS20 of our laboratory. The population analysis and

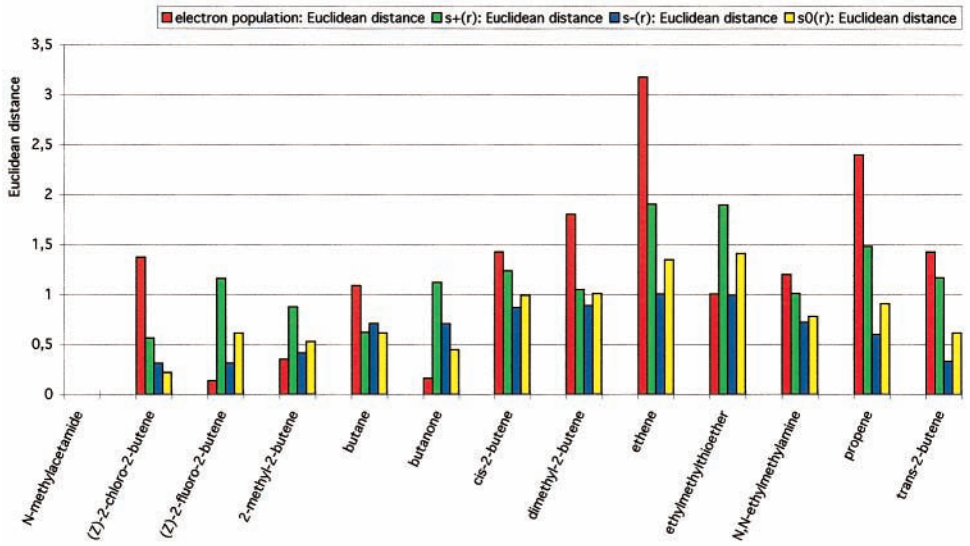


**Figure 2.** Representation of the results for the similarity calculations based on the Euclidean distance $D$ calculated between the autocorrelation functions between atom pairs separated by $d$ bonds. Both the electron population and the condensed local softnesses were used as atomic property $p$.

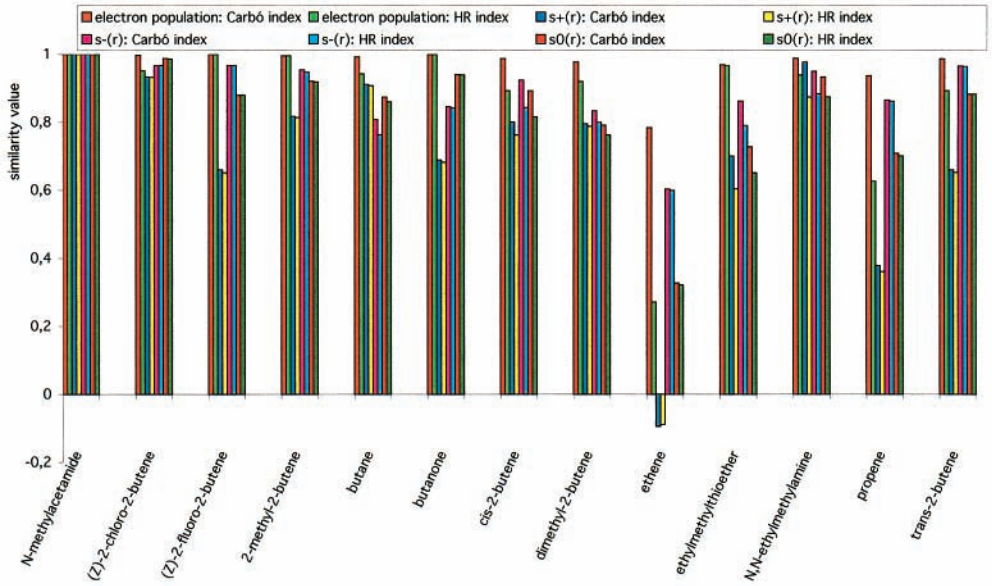


**Figure 3.** Representation of the results for the direct similarity calculations using the calculated autocorrelation functions between atom pairs separated by $d$ bonds in the expressions for the Carbó and Hodgkin-Richards type indices. Both the electron population and the condensed local softnesses were used as atomic property $p$.

the calculations of the MEP values were performed with the GAUSSIAN 98[24] program running on a COMPAQ-DIGITAL Alphaserver GS140 of the Brussels Free Universities Computer Centre. All these calculations were performed with density functional theory (DFT)[11] techniques involving the B3PW91 functional[25−27] in combination with a 6-311G* basis set which proved to be reliable in generating high-quality charge distributions.[28−30] To calculate the autocorrelation function of molecular surface properties, an in-house program was used.

**2.3. Principal Component Analysis.** Many books and papers are devoted to principal component analysis (PCA)[31−33] which was first described by Pearson in 1901[34] and by Hotelling in 1933.[35] In the past years, the use of PCA has increased, and now it is often applied in the field of chemometrics.[36] For a detailed overview of PCA, we refer to ref 8 and references given in this work.

PCA is designed to extract and highlight the systematic variation in a multivariate data matrix $X$ consisting of $N$ rows (observations, e.g., the series of molecules under investigation) and $K$ columns (variables, e.g., calculated autocorrelation functions). The structure of the investigated data set is visualized in score plots, which also provide information about the systematic deviations between the model and the data. The goodness of fit, how well one is able to mathematically reproduce the data, is given by the parameter $R^2$. The predictive power of a model is estimated by the goodness of prediction parameter $Q^2$. The most valid model is the one exhibiting the optimal balance between fit and prediction capability.

In our work, PCA[37] will be used to provide insight in the distribution of the calculated autocorrelation functions generating similarity sequences based on the calculation of the Euclidean distance $D$:

$$D_j = \sqrt{[\sum_i (t_{\mathrm{ref},i} - t_{j,i})^2]} \qquad (2.3.1)$$

where $t_{ref,i}$ and $t_{j,i}$ are the scores belonging to principal component $i$ of the reference molecule and the considered molecule $j$, respectively.

The smaller this Euclidean distance $D$, the higher the degree of similarity between the two molecules.

## 3. Results and Discussion

Peptide isosteres are often used to replace the R−CO−NH−R′ peptide bond in bioactive peptides. We considered the following series of isosteres, which is an extension of the series used in ref 3, where both R and R′ are methyl groups and where the R−CO−NH−R′ molecule is considered the reference molecule: $CH_2$═$CH_2$, R−CH═$CH_2$, R−CH═CH−R′, R−C($CH_3$)═CH−R′, R−C($CH_3$)═C($CH_3$)−R′, R−CF═CH−R′, R−CCl═CH−R′, R−$CH_2$−$CH_2$−R′, R−$CH_2$−S−R′, R−CO−$CH_2$−R′, and R−$CH_2$−NH−R′.

Figure 1 gives a brief overview of the calculations performed in this study indicating the sections where the results are discussed in detail. Starting from the electron density $\rho(\vec{r})$, we can distinguish seven different ways of studying similarity. When using the autocorrelation function, both the calculation of the Euclidean distance $D$ (eq 2.2.3) and the application of PCA result in a sequence of similarity. In the following sections, the merits of these different approaches to the study of molecular similarity will be evaluated.

**3.1. Similarity Indices Based on the Electron Density and the Local Softness.** Similarity indices depend on the relative orientation, position, and conformation of the molecules with respect to each other. To establish this relative position between the two molecules under consideration, different methods can be used. Side chains play an important role in the interactions of a peptide with its surroundings. To maintain this capability, we opted (as in our previous paper[3]) to put the centers of both molecules (taken as half of the CO−NH mimicking bond length) at the origin of a Cartesian coordinate system as this ensures a minimal movement of each of the side chains of the molecules.

Table 1 lists the calculated indices based on the electron density and the local softness for the series of peptide isosteres. These results show that the values obtained for the Carbó index are always higher than those obtained for the Hodgkin-Richards index, as could be expected because the latter takes into account both the shape and extent of the electron density.

The following sequence for the Carbó index based on the electron density $\rho(\vec{r})$ can be written in descending order of similarity: *N*,*N*-ethylmethylamine > propene > ethene > 2-methyl-2-butene > *trans*-2-butene > 2,3-dimethyl-2-butene > butanone > *cis*-2-butene > butane > (*Z*)-2-fluoro-2-butene > (*Z*)-2-chloro-2-butene > ethyl methyl thioether.

The sequence for the Hodgkin-Richards index based on the electron density is, except for some small differences, the same.

The sequence for both indices based on the local softness $s(\vec{r})$ can be written as follows: *N*,*N*-ethylmethylamine > (*Z*)-2-fluoro-2-butene > (*Z*)-2-chloro-2-butene > *trans*-2-butene > propene > 2-methyl-2-butene > 2,3-dimethyl-2-butene > *cis*-2-butene > ethene > butane > ethyl methyl thioether > butanone. This is different from the $\rho(\vec{r})$-based sequence. This could be expected taking into account the fact that the similarity indices based on the electron density and those based on the local softness contain different kinds of information. Depending on what kind of information for which we are exactly looking, similarity of shape and/or reactivity, we have to consider the $\rho(\vec{r})$-based and/or the $s(\vec{r})$-based similarity sequence, respectively.

In this study, we considered a different orientation for *N*,*N*-ethylmethylamine and ethyl methyl thioether compared to the one used in ref 3. We can see that for both indices based on the electron density and the local softness the present orientation of *N*,*N*-ethylmethylamine shows, contrary to the results in ref 3, the highest degree of similarity with the reference molecule. This again points out the importance of finding the best orientation of the considered molecules with respect to each other when calculating the Carbó and Hodgkin-Richards similarity indices.

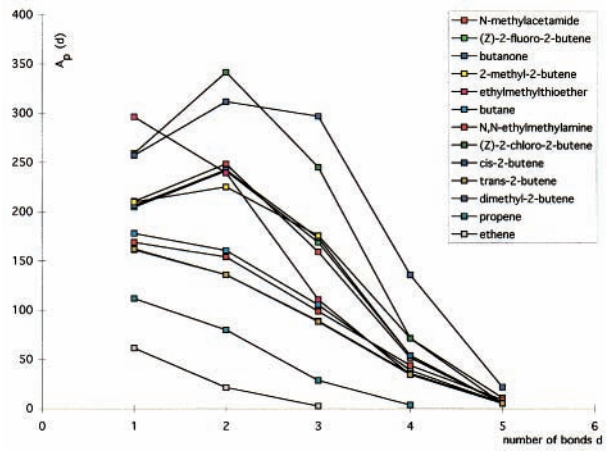


**Figure 4.** $A_p(d)$ vs the number of bonds $d$ for electron population taken as atomic property $p$.

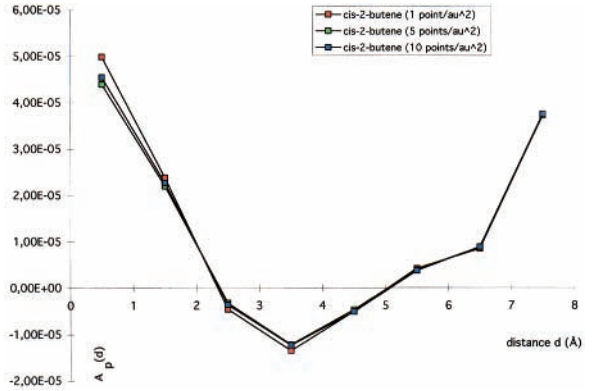


**Figure 5.** $A_p(d)$ vs the distance $d$ for the MEP taken as molecular property $p$ for the molecule *cis*-2-butene considering 1, 5, and 10 points per square atomic unit.

On the basis of the reactivity index, (*Z*)-2-fluoro-2-butene and (*Z*)-2-chloro-2-butene also seem to be good peptide isosteres which can be rationalized by the fact that in both molecules the carbon−halogen bond is polarized, as is the case for the carbonyl group of a peptide bond, although in reality this polarization is not strong enough to provoke a nucleophilic attack of enzymes.

We will now compare the sequences for the similarities in this section with those we obtain using the autocorrelation function.

**3.2. Similarity Indices Based on the Autocorrelation Function.** *3.2.1. Autocorrelation Function Based on Molecular Topology*. In this section, both the electron population and the condensed local softnesses $s_k^+$, $s_k^-$, and $s_k^0$ were used as atomic property $p$ when calculating the autocorrelation functions using eq 2.2.1. These atomic properties are in turn calculated using electron populations on atoms (cf. eqs 2.1.12−14). The popular CHelpG[38] charges are obtained using a method in which the outer surface atoms of a molecule play a predominant role and which are therefore less sensitive to changes at the interior of the molecular framework.[39] Considering the drawback of

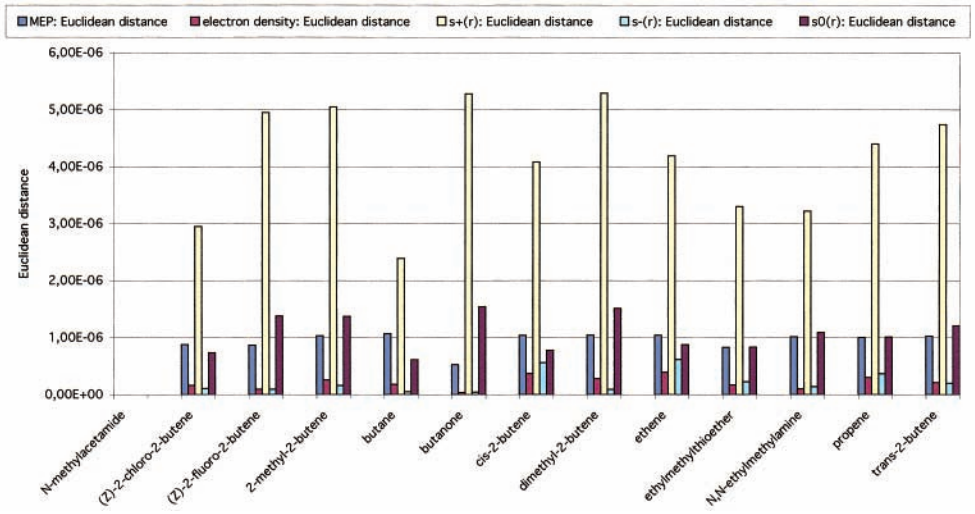


**Figure 6.** Representation of the results for the similarity calculations based on the Euclidean distance $D$ calculated between the autocorrelation functions between distinct points on the molecular surface. The MEP, the electron density, and the local softnesses were used as molecular property $p$.

these charges and the fact that in our study we want to include all the atoms of the molecules, we opted at this point to use only the Mulliken charges for the further calculations of the atomic electron population.

Table 2 shows the obtained sequence of similarity with respect to *N*-methylacetamide quantified by the Euclidean distance $D$ (eq 2.2.3). The electron population is taken as the atomic property $p$. For the isosteres dimethyl-2-butene and *N*,*N*-ethylmethylamine, different conformations, which are denoted by superscripts a−d, are considered. From these results, we notice that the sequence of similarity obtained using the autocorrelation function is indeed almost independent of the conformation of the molecules. This allows us to consider only one conformation for each molecule in the subsequent calculations.

Figures 2 and 3 depict all the results for the similarity calculations performed in this section. In Figure 2, the Euclidean distance $D$ (eq 2.2.3) between the autocorrelation functions is represented. It is seen from these results that the similarity sequences obtained using the electron population and the condensed local softness $s_k^0(\vec{r})$ are different in comparison with those obtained using the electron density- and local softness-based Carbó and Hodgkin-Richards type similarity indices (Table 1). Figure 3 shows the results obtained when the autocorrelation functions are directly used in the expressions for the Carbó and Hodgkin-Richards type indices (eqs 2.2.4 and 2.2.5).

Due to the fact that the electron density and the local softness are non-normalized functions, the Hodgkin-Richards index is a non-normalized similarity index analogous to using the Euclidean distance $D$. The sequences generated using both these approaches are comparable. Although the Hodgkin-Richards index is based on non-normalized functions such as $\rho(\vec{r})$ and $s(\vec{r})$, the sequence obtained for the Carbó index based on their normalized counterparts $\sigma(\vec{r})$ and $f(\vec{r})$ turns out to be almost the same as the one generated using the Hodgkin-Richards index.

Figure 4 is an example of a plot of the autocorrelation function $A_p(d)$ versus the number of bonds $d$ with the electron population taken as the atomic property $p$. The figure shows the large variation of the sequences of similarity as a function of the number of bonds. It is therefore important to consider the complete function and not just its value at one given distance.

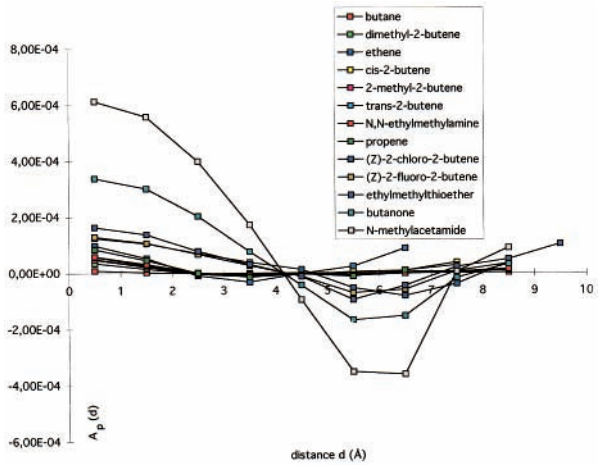


**Figure 7.** $A_p(d)$ vs the distance $d$ for the MEP taken as molecular property $p$.

Because performing the principal component analysis resulted in models where the aforementioned criteria were not satisfied (small $R^2$ and $Q^2$ values; see section 2.3), the sequences of similarity discussed in this section are limited to those obtained by calculating the Euclidean distance $D$ (eq 2.2.3) and by using the autocorrelation functions directly in the Carbó and Hodgkin-Richards type indices.

*3.2.2. Autocorrelation Function of Molecular Surface Properties.* In calculating the autocorrelation function of molecular properties on distinct points on the van der Waals surface, we first tested how the number of distinct points considered per unit surface influences the calculated autocorrelation function. Figure 5 is an example of a plot of the autocorrelation function $A_p(d)$ versus the distance $d$ for the molecule *cis*-2-butene with the MEP taken as molecular property $p$ considering 1, 5, and 10 distinct points per square atomic unit. The similarity of the three curves shows that one point per square atomic unit is sufficiently accurate for all further calculations.

In this section, the MEP, the electron density $\rho(\vec{r})$, and the local softnesses $s^+(\vec{r})$, $s^-(\vec{r})$, and $s^0(\vec{r})$ are taken as molecular properties $p$ in the calculation of the autocorrelation functions (eq 2.2.2). Although it is obvious that the electron density calculated on distinct points of the van der Waals surface is showing only small variations, we considered the calculation

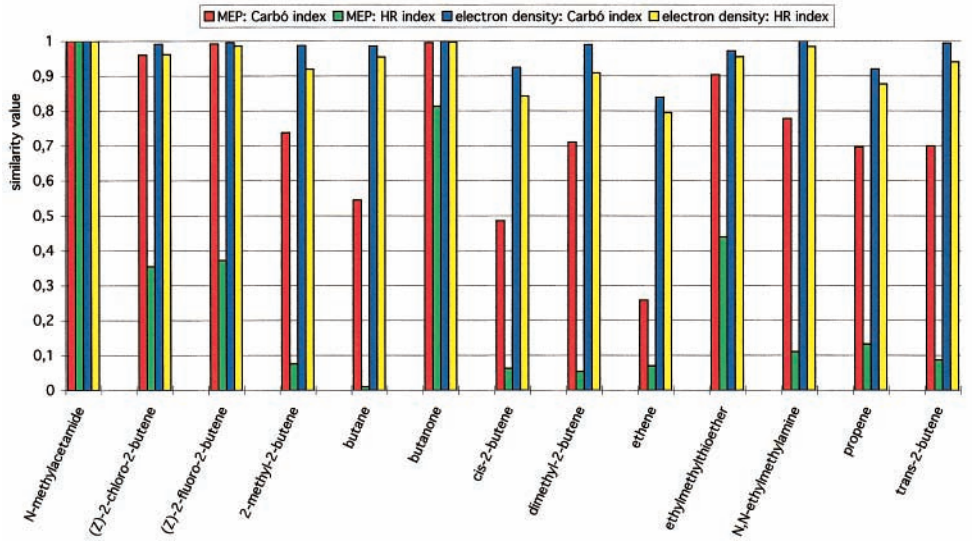


**Figure 8.** Representation of the results for the direct similarity calculations using the autocorrelation functions between distinct points on the molecular surface in the expressions for the Carbó and Hodgkin-Richards type indices. Both the MEP and the electron density were used as molecular property *p*.

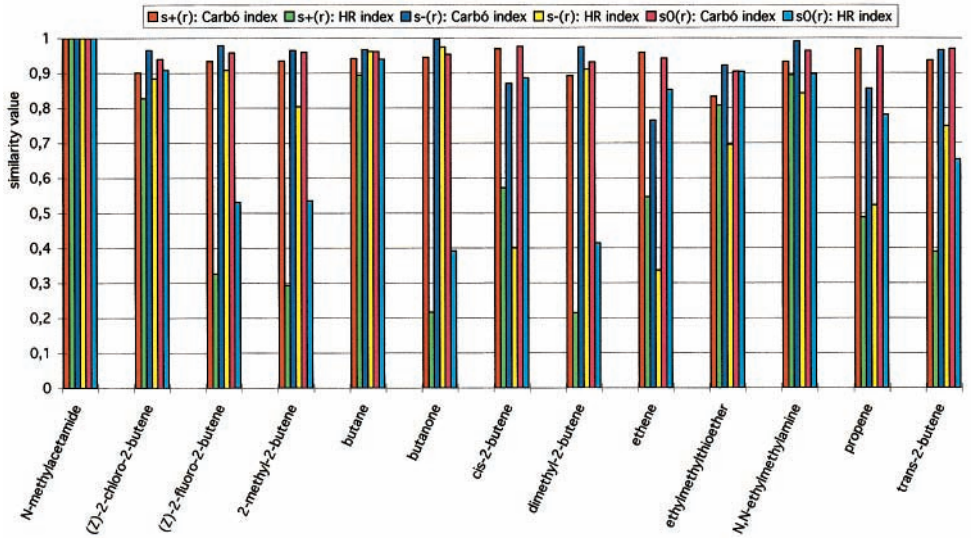


**Figure 9.** Representation of the results for the direct similarity calculations using the autocorrelation functions between distinct points on the molecular surface in the expressions for the Carbó and Hodgkin-Richards type indices. The local softnesses $s^+(\vec{r})$, $s^-(\vec{r})$, and $s^0(\vec{r})$ were used as molecular property *p*.

of the autocorrelation function using the electron density for the sake of completeness of our study.

Figure 6 represents the results for the similarity calculations using the Euclidean distance *D* (eq 2.2.3). The results obtained using the electron density compared with those obtained by calculating the autocorrelation function between atom pairs using the electron population (Figure 2) give the same trends in similarity sequences with butanone and (*Z*)-2-fluoro-2-butene showing the best similarity with the reference molecule. These comparable sequences could be expected due to the fact that, as mentioned earlier, the electron population $q_k(N)$ on atom *k* can be considered as the condensed form of the electron density $\rho(\vec{r})$.

It will be of interest to use more advanced techniques to generate charges, and consequently electron populations, as for example the Stockholder charges[40] or the charges obtained from Baders' atoms in molecules[41] as an alternative to the highly popular Mulliken population analysis.

Comparing the results using the local softnesses $s^+(\vec{r})$, $s^-(\vec{r})$, and $s^0(\vec{r})$ (Figure 6) with those obtained in section 3.2.1 (Figure 2) shows there are different similarity sequences for each type of local softness. Furthermore, butane instead of butanone apparently shows a high degree of similarity with the reference molecule. As the HOMO and the LUMO are especially located in the carbon−hydrogen $\sigma$-bonds of butane, which normally do not show any reactivity at all, this high degree of similarity is rather unexpected, and we cannot account for this on an intuitive chemical basis.

The similarity sequences based on the calculation of the Euclidean distance *D* obtained using both the electron density $\rho(\vec{r})$ and the local softness $s^0(\vec{r})$ (Figure 6) compared with the results for the corresponding Carbó and Hodgkin-Richards type similarity indices calculated in section 3.1 (Table 1) yield different similarity sequences.

Figure 7 shows a plot of the autocorrelation function $A_p(d)$ versus the distance *d* (in angstroms) with the MEP taken as the molecular property *p*. This figure shows that the similarity sequence is a function of the distance *d*, again indicating that the complete function should be considered when establishing similarity sequences.

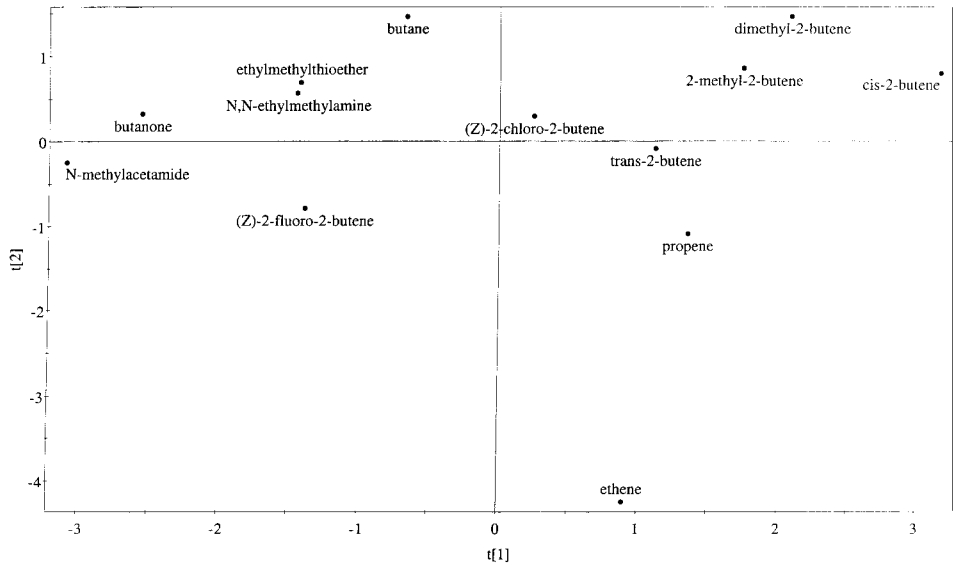


**Figure 10.** Score plot obtained by performing the PCA on the calculated autocorrelation functions using the electron density $\rho(\vec{r})$ as molecular property $p$.

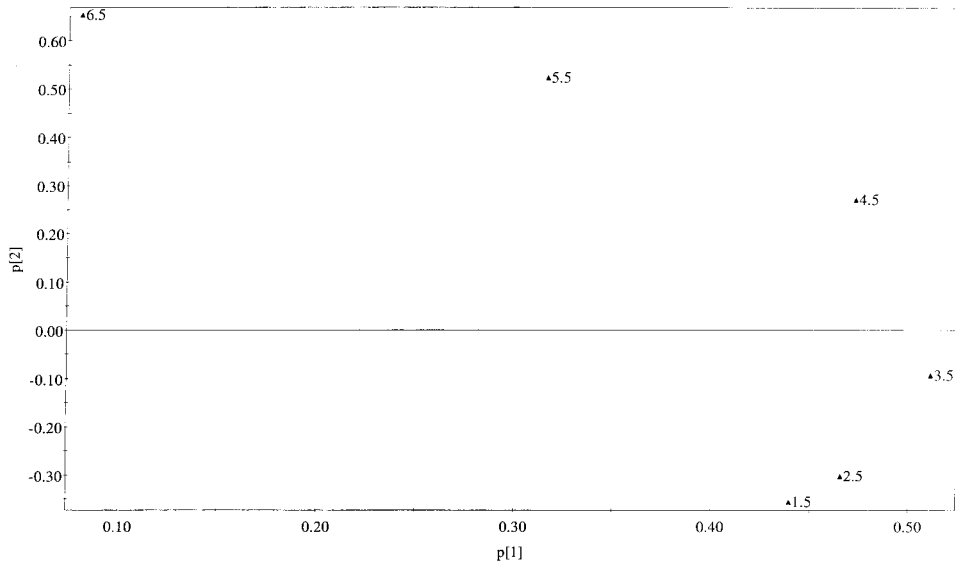


**Figure 11.** Plot of loadings obtained by performing the PCA on the calculated autocorrelation functions using the electron density $\rho(\vec{r})$ as molecular property $p$.

Figures 8 and 9 show the results obtained using the calculated autocorrelation function directly in the expressions for the Carbó and Hodgkin-Richards indices (eqs 2.2.4 and 2.2.5). As in section 3.2.1, the sequences using the Euclidean distance $D$ and the non-normalized Hodgkin-Richards index are the same.

Furthermore, no resemblance can be found when comparing the sequences in Figures 8 and 9, using the electron density $\rho(\vec{r})$ and the local softness $s^0(\vec{r})$ as the molecular property, with those given in Table 1.

At this point, two important remarks have to be made. First, and as already mentioned in section 3.1, molecular similarity of shape and/or reactivity can be evaluated by considering the $\rho(\vec{r})$-based and/or the $s(\vec{r})$-based similarity sequences, respectively. Furthermore, it is seen from the results, considering the three local softness functions $s^+(\vec{r})$, $s^-(\vec{r})$, and $s^0(\vec{r})$, that most of the molecules are more similar according to $s^-(\vec{r})$ than to $s^+(\vec{r})$. When mimicking the polarization of the carbonyl group of the reference molecule, the partial negative charge of this bond is more "expressed" than its partial positive charge, probably due to the fact that the latter is more embedded in the molecule as a whole. Therefore, most of the molecules show a higher similarity for an electrophilic attack. Second, when the quality of the autocorrelation function in studies of molecular similarity is being evaluated, an absolute reference point should be defined. This reference should be the maximum similarity value resulting from the calculation of the Carbó and Hodgkin-Richards type indices based on the electron density $\rho(\vec{r})$ and the local softness $s(\vec{r})$ and taking into account full optimization of the conformation, the relative orientation, and the position of the molecules. This optimization is computationally extremely hard and tedious because it needs to be carried out continuously and flexibly and a number of conformational degrees of freedom have to be incorporated. Besides this difficult optimization, the time-consuming calculation of the three-dimensional integration also needs to be performed. Consequently, the results given in Table 1 cannot be considered the real maximum similarity values, because the full optimization has not been carried out and because the integration is calculated numerically; albeit,

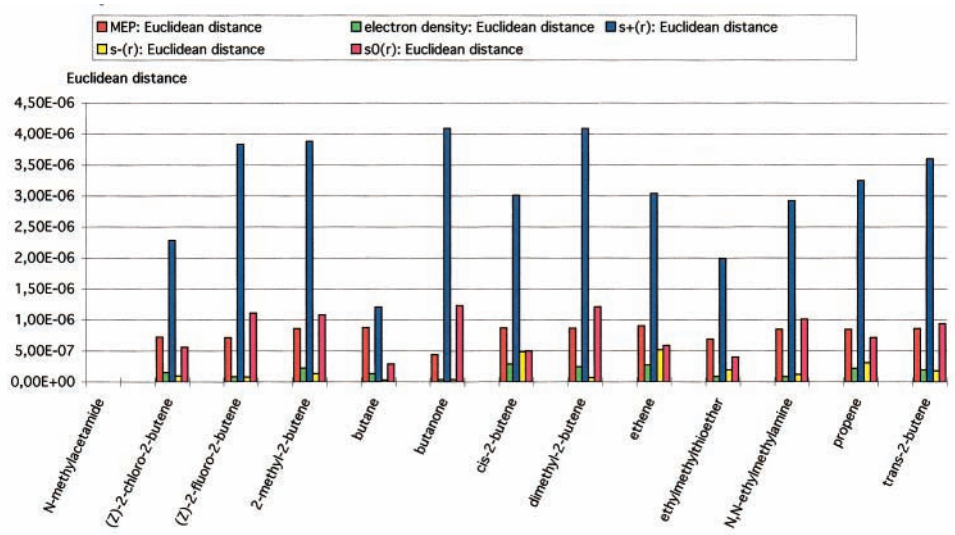


**Figure 12.** Overview of the results for the similarity calculations based on the calculation of the Euclidean distance $D$ (eq 2.2.3) between the autocorrelation functions between distinct points on the molecular surface. The considered distance interval is restricted between 1 and 7 Å.

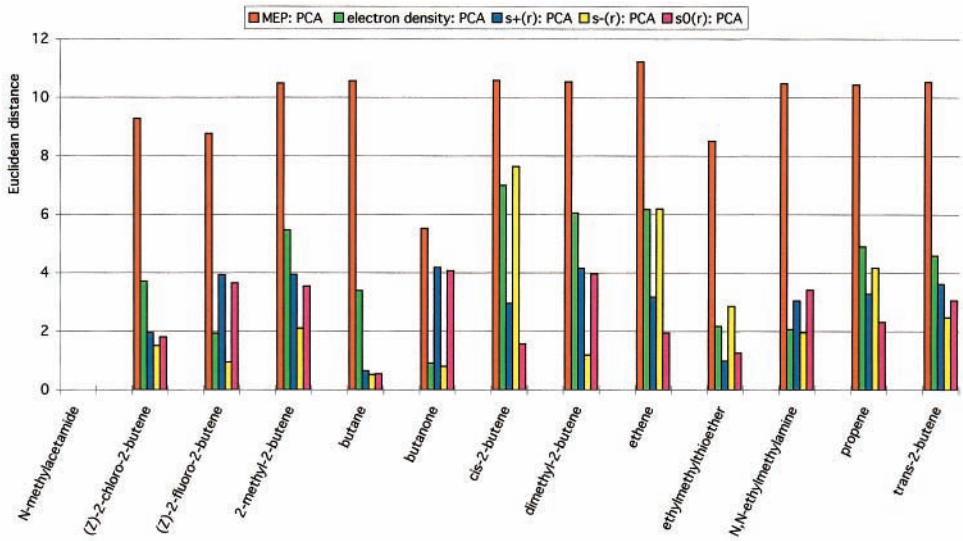


**Figure 13.** Overview of the results for the similarity calculations when performing a PCA on the autocorrelation functions and calculating the Euclidean distance $D$ (eq 2.2.3) between the scores of the obtained principal components. The considered distance interval is restricted between 1 and 7 Å.

they may serve as a first indication (cf. the arbitrary choice of the position and orientation of the molecules described in section 3.1).

An illustration of this general problem of optimization is the significant difference in the value for the similarity index $R_{AB}$ and $H_{AB}$ (Table 1) between (*Z*)-2-fluoro-2-butene and (*Z*)-2-chloro-2-butene, which can be caused by not performing the full optimization of their relative orientation and position (translation and rotation) with respect to the reference molecule. This effect is even more likely due to the presence of the chlorine atom in (*Z*)-2-chloro-2-butene. It is known that molecular similarity calculations using electron densities are very dependent on the presence of heavier atoms in the molecules that are being compared.[42] The methods proposed in this paper could solve this serious drawback. Furthermore, the difference in similarity values between the cis and trans isomers from 2-butene stresses the importance of the orientation and conformation of the molecules.

As an alternative to the direct calculation of the Euclidean distance $D$ (eq 2.2.3), PCA is used. The best models were obtained when considering only the autocorrelation functions calculated for distance intervals between 1 and 7 Å. Performing the PCA on the autocorrelation functions using, for example, the electron density $\rho(\vec{r})$ as molecular property $p$ in eq 2.2.2 yields an acceptable model consisting of two principal components where the cumulative $R^2$ and $Q^2$ values are 0.977 and 0.913, respectively. Figure 10 is the score plot for these two principal components where butanone shows the highest similarity with the reference molecule. A plot of the loadings for the two principal components is given in Figure 11, showing that both the small (between 1 and 3 Å) and large (between 6 and 7 Å) distance intervals are contributing to PC1 and PC2.

Figure 12 is an overview of the results for the similarity calculations based on the calculation of the Euclidean distance $D$ (eq 2.2.3). Figure 13 represents the results for the similarity calculations when performing a PCA on the autocorrelation functions and calculating the Euclidean distance $D$ (eq 2.3.1) between the scores of the principal components. For both these figures, the considered distance interval is restricted between 1 and 7 Å. These two approaches lead to similar results. This is comparable to the cases shown in Figures 8 and 9. For example, for the electron density $\rho(\vec{r})$, the MEP, and the local softness

$s^{-}(\vec{r})$, butanone shows the highest similarity with the reference molecule, which again can be rationalized in terms of the polarization of the carbon−oxygen bond. This indicates that the autocorrelation function calculated for molecular surface properties is a practical and handy tool for generating molecular similarity sequences.

## 4. Conclusions

To evaluate molecular similarity of shape and/or reactivity, $\rho(\vec{r})$-based and/or $s(\vec{r})$- and MEP-based similarity sequences can be considered, depending on the kind of information, electron distribution or reactivity versus soft or hard partners, for which one is looking.

In contrast to the Carbó and Hodgkin-Richards type indices, the use of the autocorrelation function in evaluating molecular similarity has as an important practical advantage in the absence of the calculation of the time-consuming integration over the three-dimensional space. Furthermore, the optimization process of the similarity involving the conformation, orientation, and position of the molecules can also be omitted. Further research should reveal whether the inevitable loss of information associated with autocorrelation functions (transformation of three- into two-dimensional information) is a matter of concern in similarity calculations. This necessitates the calculation of the optimized similarity values, which should be used as reference points. Nevertheless, on the basis of the study presented here, the use of the autocorrelation function can be justified in studies of molecular similarity of large series of structurally highly related systems where the positioning and the conformational degrees of freedom are less important.

**Acknowledgment.** We thank Prof. P. Popelier (University of Science and Technology, Manchester, England) for allowing us to use the program Morphy. P.G. thanks the Free University of Brussels for a generous computer grant and the FWO for continuous support. He and G.B. thank Prof. R. Carbó for many stimulating discussions on different aspects of molecular similarity in recent years.

## References and Notes

(1) Carbó, M.; Arnau, M.; Leyda, L. *Int. J. Quantum Chem.* **1980**, *17*, 1185.

(2) Hodgkin, E. E.; Richards, W. G. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105.

(3) Boon, G.; De Proft, F.; Langenaeker, W.; Geerlings, P. *Chem. Phys. Lett.* **1998**, *295*, 122.

(4) Yang, W.; Parr, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 6723.

(5) Geerlings, P.; De Proft, F.; Langenaeker, W. *Adv. Quantum Chem.* **1999**, *33*, 303.

(6) Geerlings, P.; De Proft, F. *Int. J. Quantum Chem.* **2000**, *80*, 227.

(7) Roy, R. K.; De Proft, F.; Geerlings, P. *J. Phys. Chem. A* **1998**, *102*, 7035.

(8) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS)*; Umetrics AB: Umeå, Sweden, 1999.

(9) Dughan, L.; Burt, C.; Richards, W. G. *THEOCHEM* **1991**, *235*, 481.

(10) Parr, R. G.; Yang, W. *J. Am. Chem. Soc.* **1984**, *106*, 4049.

(11) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(12) Yang, W.; Mortier, W. J. *J. Am. Chem. Soc.* **1986**, *108*, 5708.

(13) Mulliken, R. S. J. *J. Chem. Phys.* **1955**, *23*, 1833.

(14) Cliff, A. D.; Ord, J. K. *Spatial Autocorrelation*; Pion Limited: London, 1973.

(15) Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769.

(16) Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.

(17) Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757.

(18) Zakarya, D.; Tiyal, F.; Chastrette, M. *J. Phys. Org. Chem.* **1993**, *6*, 574.

(19) Bonaccorsi, R.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1970**, *52*, 5270.

(20) Scrocco, E.; Tomasi, J. *Top. Curr. Chem.* **1973**, *42*, 95.

(21) Scrocco, E.; Tomasi, J. *Adv. Quantum Chem.* **1978**, *11*, 115.

(22) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.

(23) Popelier, P. L. A. *Comput. Phys. Commun.* **1996**, *93*, 212.

(24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.6; Gaussian, Inc.: Pittsburgh, PA, 1998.

(25) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(26) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 609.

(27) Stevens, P. J.; Delvin, F. J.; Chablaoski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(28) De Proft, F.; Martin, J. M. L.; Geerlings, P. *Chem. Phys. Lett.* **1996**, *250*, 393.

(29) Geerlings, P.; De Proft, F.; Martin, J. M. L. In *Recent Developments and Applications of Modern Density Functional Theory, Theoretical and Computational Chemistry*; Seminario, J., Ed.; Elsevier: Amsterdam, 1996; Vol. 4, p 773.

(30) De Proft, F.; Tielens, F.; Geerlings, P. *THEOCHEM* **2000**, *506*, 1.

(31) Jackson, J. E. *A user's guide to principal components*; John Wiley: New York, 1991.

(32) Malinowski, E. R. *Factor analysis in chemistry*, 2nd ed.; John Wiley: New York, 1991.

(33) Wold, S.; Esbensen, K.; Geladi, P. *Principal Component Analysis*; Chemometrics Intelligence Laboratory: Amsterdam, 1987; System 2, p 37.

(34) Pearson, K. *Mathematical contributions to the theory of evolution XIII. On the theory of contingency and its relation to association and normal correlation*; Drapers Co. Res. Mem. Biometric series I; Cambridge University Press: London, 1901.

(35) Hotelling, H. *J. Educ. Psychol.* **1933**, *24*, 417, 498.

(36) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; Elsevier Science: Amsterdam, 1998; Parts A and B.

(37) *SIMCA-P*, version 8.0; Umetrics AB, Inc.: Umeå, Sweden, 2000.

(38) Breneman, C. M.; Widberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(39) Jensen, F. *Introduction to Computational Chemistry*; John Wiley: Chichester, U.K., 1999; p 221.

(40) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129.

(41) Bader, R. F. W. *Atoms in Molecules, A Quantum Theory*; Clarendon Press: Oxford, 1990; p 182.

(42) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25.