Quantum Molecular Similarity. 1. BCP Space

P. L. A. Popelier

Department of Chemistry, U.M.I.S.T., 88 Sackville Street, Manchester M60 1QD, Great Britain

Received: December 14, 1998; In Final Form: January 29, 1999

We propose a new similarity measure operating in an abstract space spanned by properties evaluated at bond critical points defined by the theory of *Atoms in Molecules*. Consequently, we represent molecules compactly and reliably, extracting the relevant information from their ab initio wave function. Typical problems of continuous quantum similarity measures are thereby avoided. The practical use of this novel method is adequately illustrated via the Hammett equation for *para* and *meta* substituted benzoic acids. On the basis of our definition of distances between molecules in BCP (Bond Critical Point) space, we are able to reproduce the experimental sequence of acidities determined by the well-known σ constant of a set of substituted congeners. Moreover, our approach points out where the common reactive center of the molecules is. Due to these promising results we embark on a research program systematically addressing further issues outlined in this work. The generality and feasibility of our approach will enable predictions in medically related QSARs.

1. Introduction

The design of a novel or improved drug is an extremely challenging but highly rewarding task, which explains the current plethora of approaches. One group of techniques resides under the heading of what is commonly referred to as "quantitative structure−activity relationships" (QSAR).[1] This approach is based on the simple idea that the chemical behavior of a molecule in a chemical environment (e.g., reactivity, ligand docking, acidity) is due to the very structure of that molecule. Put in almost trite terms this is equivalent to the statement that "a molecule acts like it acts because it is what it is". An important consequence of this apparently trivial assumption is that we need not understand the often extremely complex details of the molecule's action in the chemical environment. In other words, using the data of one molecule's action we can predict the action of another closely related molecule by merely comparing how *similar* the original molecule is to the other one. This is the basis of the *molecular similarity*[2] postulate, which we adopt in this and future work.

The new exciting field of combinatorial drug discovery[3] recently emerged from advances in high-throughput screening and solid-phase organic synthesis. Despite great enthusiasm for the concomitant concept of "molecular diversity", it has been argued that combinatorial drug design does not obviate "computer-assisted drug design" (CADD).[4] Indeed, the mass screening brought about by combinatorial chemistry should not imply an irrational brute force method but should be combined with "rational drug design", as advocated by Martin et al.[5] Encouraged by the continuing prominence of molecular similarity measures, we embarked on a program of molecular similarity research based on the theory of *Atoms in Molecules* (AIM).[6,7] This theory operates on the electron density obtained by ab initio calculations. In view of the spectacular hardware developments of recent years, all required information of a single thirty-atom molecule can be obtained in several hours. However, to perfectly compete with diversity screening techniques, millions of compounds would have to be processed by computers in less than a month. This is currently not feasible with the typical CPU power available, but then this is not the goal of our research. Our goal is to extract detailed but compact electronic structure information out of available wave functions and transfer it to a QSAR problem, in an attempt to obtain insight in the cause of a given molecule's activity. For example, we want to know where the active center of the molecule is.

This contribution is the first of a series of papers proposing a useful method to measure molecular similarity employing ab initio calculations. Here we discuss the methodology of our approach, and discuss the practicalities arising from full program implementation. To show that our approach works in real cases we present a successful example in connection with the oldest known QSAR: the Hammett relation for substituted benzoic acids.

2. Brief Review of "Atoms in Molecules"

Almost three decades ago the foundation of the theory of "Atoms in Molecules" (AIM) was laid by the realization that a molecule could be naturally partitioned into atoms[8] whose energy can be precisely defined using quantum mechanics. Indeed, each atom obeys the virial theorem in the same way the whole molecule does. The ability to define the energy of an atom inside a molecule is most remarkable and renders the atomic subspace into a unique portion of space dominated by the nucleus it contains. One of the consequences of this partitioning is that atoms exhibit a bewildering variety of shapes, which reflects the uniqueness of each "molecular atom" and the complexity of chemistry itself.

Over the years AIM has evolved into a program to bridge the gap between modern ab initio wave functions and chemical insight. One of the main lines of thought in the development of this program is the use of the electron distribution ρ and related quantities such as the Laplacian of ρ (or $\nabla^2\rho$) as a starting point. The electron distribution is a common platform to a host of computational schemes such as the standard SCF-LCAO-MO technique, grid-based methods and even experiment (X-ray diffraction). In other words, many different levels of theory (including basis set types) ultimately lead to ρ, which can also be experimentally observed. If we take ρ as the source of information to study a considerable subset of a molecule's

10.1021/jp984735q CCC: $18.00 © 1999 American Chemical Society
Published on Web 03/20/1999

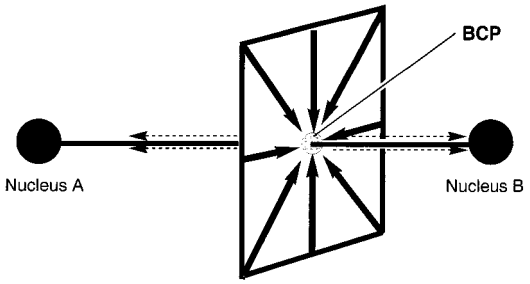

Figure 1. Schematic representation of a *bond critical point* (BCP) between two nuclei A and B. The curve linking A and B (the bond path) is not necessarily a straight line in general. The electron distribution ρ increases toward the BCP in a plane locally perpendicular to the bond path. Note that at the BCP ρ is a minimum along the bond path.

chemical characteristics, our ρ-based concepts will hold independent of how ρ has been obtained. In contradistinction, several population analyses for example only exist within a specific theoretical or computational framework.

An extensive review of AIM would encompass a full discussion of the topology of ρ including the powerful yet simple concept of the *gradient vector field*. For a comprehensive account, the reader is referred to ref 9. For the purpose of this paper we only focus on the so-called *bond critical points* (BCP). These are points in real 3D space where the gradient of the electron distribution vanishes (or $\nabla\rho = \mathbf{0}$) and where the Hessian of ρ (or $\nabla\nabla\rho$) has two negative eigenvalues and one positive one. These remarkable points occur roughly between two bonded nuclei and are part of a more complex topology of ρ, which we cannot review within the present scope. The Hessian $\nabla\nabla\rho$ is basically a matrix describing all possible second derivatives of ρ with respect to position coordinates x, y, and z. The reason we look at the eigenvalues λ_i is because they express the local curvature of ρ in a point independent of the choice of molecular coordinate system. By convention they are ordered as follows: $\lambda_1 < \lambda_2 < \lambda_3$. Consequently, at a BCP, $\lambda_1 < \lambda_2 < 0$ and $\lambda_3 > 0$. The latter positive curvature is associated with an eigenvector that is tangent to the *bond path* (BP). The BP is a curve in real space linking two bonded nuclei along which ρ is a maximum with respect to any neighboring line. The BCP lies on the BP and therefore it adopts the property that defines the BP; i.e., at the BCP ρ attains a maximum value for any displacement toward it in a plane perpendicular to the BP. This is made clear in Figure 1.

3. Quantum Molecular Similarity Measures

Many techniques to measure similarity have been proposed entirely outside the realm of quantum similarity. Examples encompass algorithms for clustering 2D structures, similarity searching through 3D databases, molecular surface matching, neural networks, shape-group methods to describe the topology of molecular shape, shape-graph descriptions and CoMFA (Comparative Molecular Field Analysis). Ultimately, any molecular similarity method must incorporate conformational flexibility requiring special techniques that avoid typical combinatorial explosions. A recent discussion of a fair cross-section of methods can be found in a book by Dean.[2] In this paper we focus on similarity measures operating on quantum mechanical information.

In 1980 Carbó et al.[10] addressed for the first time the fundamental question "How similar is a molecule to another?" from a quantum chemical point of view. On the basis of the assumption that similar molecules must have similar electron distributions, they proposed the matching measure between molecules A and B to be simply $\epsilon_{AB} = \int_V |\rho_A - \rho_B|^2 \, dV$. After some rearrangement, the computation of ϵ_{AB} comes down to rotating and translating the compared molecules in order to maximize the value of the integral $\int_V \rho_A \rho_B \, dV$.

During the last two decades the main idea behind this index has proved its merit in view of the considerable attention it has received in the literature (see ref 11 and references therein]. However we should be aware that focusing all attention on the electron distribution is a choice and that a measure incorporating alternative quantities may prove to be closer to the ultimate similarity measure. In fact some methods are entirely based on the electrostatic potential,[12] the momentum density,[13] or just the 3D *shape* of the electron distribution.[14] But then, again, the role of ρ as the ultimate source of all information about a molecule cannot be underestimated. For example, the electrostatic potential is defined as an integral of ρ weighted by $|\mathbf{r} - \mathbf{r}'|^{-1}$ and the local electronic kinetic energy density $G(\mathbf{r})$ has recently been evaluated from the experimental electron density.[15]

Below we mention a few typical problems that arise in the evaluation of Carbó-like indices. First of all, the index is expensive to compute for ab initio wave functions of reasonable quality (i.e., at least HF/6-31G*) especially because it has to be calculated for every pair of compared molecules. To remedy this problem, a method to approximate ρ by fitting spherical Gaussian functions was proposed.[16] Second, two molecules under comparison have to be superimposed so as to maximize the index. This is a very time-consuming procedure further hampered by multiple (undesirable) maxima. Also the index is dependent on the method chosen for molecular matching.[17] Third, it is not clear if the comparisons should be limited to small regions or performed over the whole molecule. To patch up this difficulty, a method with arbitrary and nonunique fragment densities has been presented.[18] Finally, the measure is severely biased by core density contributions, which led to the introduction of valence electron density similarity measures,[19] addition of nuclear charges to screen the core electronic charge,[20] and also to indices based on the electrostatic potential and field.[21]

In an attempt to formulate a fast, reliable, and therefore useful molecular similarity index that is free of the aforementioned problems, we want to take full advantage of the insight that AIM offers into the electronic structure of a molecule. Since this work is *not* intended to be a technical improvement of the Carbó method (aiming at a CPU time gain or computational stability), we have not revisited our Hammett case study by his method. Instead, we look at the quantum similarity issue independently from a philosophically different point of view, encouraged by observations described in the next section.

4. BCP Space

The integrals appearing in typical quantum similarity measures essentially express that the electron density *in every point of space* contributes to the comparison between two molecules. Although one cannot argue against the completeness of this approach, it leads to chemically unimportant regions (such as the nuclear cores), greatly influencing the similarity measure. Is there any way we could focus on a few remarkable points in the molecule, thereby replacing the integral by a sum? We believe that AIM provides such a set of points in a simple and unbiased way, namely the BCPs.

The BCPs are topologically unique points in space that can easily be located if a robust algorithm is used.[22] This algorithm is an eigenvector following method, which is superior to the

Newton−Raphson method in that it is able to locate critical points starting from a poor guess. It has been shown before that several properties evaluated at the BCP summarize the characteristics of the corresponding bond. For example, the electron density at the BCP, denoted by ρ_b, determines a bond order that yields values of 1.0, 1.6, 2.0, and 2.9 for the C−C bonds in ethane, benzene, ethene, and ethyne, respectively.[23] Also strong correlations have been found between bond energy and ρ_b.[24]

As a further example, the Laplacian of the electron density at the BCP, denoted by $\nabla^2\rho_b$, distinguishes two broad classes of bonds: if $\nabla^2\rho_b < 0$, the bond is a so-called *shared interaction*, but if $\nabla^2\rho_b > 0$, the bond is called a *closed-shell interaction*. Covalent bonds belong to the former class, and ionic bonds, hydrogen bonds, and van der Waals bonds belong to the latter. The distinction between these two types of interaction can be rationalized via the equation $\nabla^2\rho_b = \lambda_1 + \lambda_2 + \lambda_3$. If the positive eigenvalue λ_3 dominates, density is accumulated along the bond path toward the nuclei. If the negative eigenvalues dominate, then the electron density accumulation in the plane perpendicular to the bond path is prominent. This reflects the large charge buildup between two bonded nuclei, which is reminiscent of covalent bonding. The quantity $\nabla^2\rho_b$ provides, of course, more subtle information than the crude classification explained above, as demonstrated in early seminal work on bond properties in hydrocarbons.[25]

A third important quantity describing another facet of the electronic structure of a bond is the ellipticity at the BCP, denoted by ϵ_b or simply ϵ. The ellipticity is defined as $\lambda_1/\lambda_2 - 1$ and is always positive because $\lambda_1 < \lambda_2 < 0$. Since $|\lambda_1| > |\lambda_2|$ the latter corresponds to the "soft" curvature. A contour diagram of ρ in the plane of the eigenvectors corresponding to λ_1 and λ_2 shows a set of nested ellipses (or circles if $\lambda_1 = \lambda_2$). Clearly, λ_2 corresponds to the major axis because in this direction fewer contour lines are crossed per unit length as a result of the soft curvature. The ellipticity measures the susceptibility of ring bonds to rupture and provides a quantitative generalization of the $\sigma-\pi$ character of a bond.

It has been shown before[25] that the descriptors ρ_b, $\nabla^2\rho_b$, and ϵ are very successful at translating the predicted electronic effects of orbitals theories into observable consequences in ρ. The large body of data presented in that work for hydrocarbons (Table 4 in ref 25) is astonishingly consistent and reveals many subtleties despite the elementary basis set. In particular, BCP properties detect conjugation, subtle delocalization effects and hyperconjugation. They distinguish aromatic and anti-aromatic character and parallel bond order and prove that three-membered saturated hydrocarbon rings act like double bonds. The confidence in the main idea behind our molecular similarity program is largely based on the early observations of Bader and co-workers[25] and further observations made by the present author in connection with the well-known drug haloperidol[2] discussed below.

The electron distribution, its Laplacian, and the ellipticity are in fact three components of a so-called chemical descriptor vector.[26] Each vector describes a bond in a three-dimensional BCP space. Of course the dimensionality of the BCP hyperspace can be increased by adding more components such as the kinetic energy density K_b. Thus each molecule is represented by just a handful of numbers, being the components of the vectors describing the molecule's bonds. The basic working hypothesis is that—disregarding several technical issues—the molecule is completely and accurately described in a compact and abstract space called BCP space. As a result, similarity measures are reduced to *discrete* distance-like measures in BCP space without loosing their quantum mechanical basis.

Some advantages to this way of tackling similarity become clear immediately. First there is no need to superimpose molecules in real space, which is a cumbersome procedure. In fact the molecule's absolute orientation is lost in BCP space but in our approach this information is not needed anyway. The only caveat is that two molecules with different handedness are mapped into the same coordinates in BCP space, but this lost information can easily be added to the BCP properties as an extra discrete dimension. Second, comparisons over restricted regions of the molecule are straightforward because of the discrete nature of the representation. Moreover, the possibility of including only a part of the molecule in the similarity calculation is the key to find the active center of a molecule in a chemical environment, as will become clear from our Hammett QSAR example. Finally, our method is in no way biased by the core density but includes it in a more balanced way. Indeed, it has been shown that the core density is an important contributor to BCP properties. If it is omitted such as in semiempirical wave functions[27] (or wave functions obtained with the effective core potential (ECP) or valence density plane wave approximation), then severely distorted or corrupted topologies appear. In other words, the absence of core densities in these models often causes the absence of the BCP, which proves that they influence not just the position but even the presence of the BCP and therefore the BCP representation of the molecule.

We believe that since the core does not dominate BCP space, our similarity measure in BCP space is more discriminative and predictive. This opinion is mainly based on the Hammett QSAR example presented below, but a glance at the representation of the drug haloperidol in BCP space corroborates our view. For details the reader is referred to[2] but here we just reiterate the main point, namely the astounding fine-tuning the BCP space reveals in the classification and characterization of bonds. Figure 2a shows a stick diagram of haloperidol ($C_{21}O_2NH_{23}FCl$), which consists of four fragments: fluorobenzene, chlorobenzene, 4-hydroxypiperidine, and butyraldehyde.

Haloperidol has 51 bonds, each of which is represented as one BCP in a 3D BCP space, spanned by the properties ρ_b, $\nabla^2\rho_b$, and ϵ. The complete representation[2] shows that the BCPs cluster up in 10 well-resolved clusters. Figure 2b shows a representative BCP for each cluster. It has been observed[2] that the $C_{arom}-C_{arom}$ cluster is in fact split in two: the smaller subcluster represents the two pairs of benzene carbon−carbon bonds adjacent to the C−F or C−Cl bond. These four bonds show a somewhat higher ellipticity than the other members of their cluster because halogens are π-donors. Moreover, this fine-tuning is even correct in predicting that fluorine is a stronger π-donor than chlorine, since fluorine causes the largest increase in ϵ. Furthermore, hyperconjugation can be spotted in the structure of the $C_{aliph}-C_{aliph}$ cluster. These and further observations increase the level of confidence in the power of the BCP space to describe the electronic structure of a molecule compactly and reliably.

5. Similarity in BCP Space

Once a meaningful description of a molecule in some space has been obtained, there are many ways to measure the similarity between molecules. The chemical descriptor vector described above can be operated upon via a host of mathematical tools such as equivalence, matching, partial ordering, proximity, graph theory, and even group theory.[26] Here we will restrict ourselves to a simple Euclidean distance measure in BCP space. The

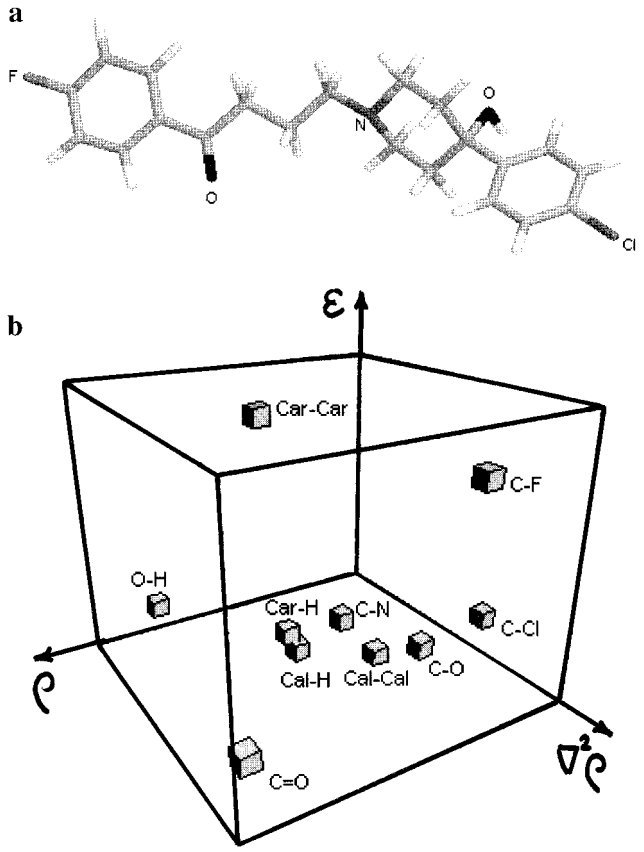

Figure 2. (a) Stick diagram of a conformation of the drug haloperidol. Only atoms that are not carbons (gray) or hydrogens (white) are labeled. (b) Set of 10 representative BCPs of haloperidol represented in a 3D BCP space spanned by the properties ρ_b,$\nabla^2\rho_b$, and ϵ. Each BCP marks one of the following types of bonds: C_{arom}—C_{arom}, C_{arom}—H, C—Cl, C—F, C_{aliph}—C_{aliph}, C—N, C—O, C═O, O—H, and C_{aliph}—H.

distance d_{ij} between two BCPs i and j in our 3D BCP space is defined as follows:

$$d_{ij} = [(\rho_{b,i} - \rho_{b,j})^2 + (\nabla^2\rho_{b,i} - \nabla^2\rho_{b,j})^2 + (\epsilon_{b,i} - \epsilon_{b,j})^2]^{1/2} \quad (1)$$

The distance d(A,B) between two molecules A and B is then defined as a sum of these BCP distances via eq 2. The lower the value d(A,B), the more similar the two molecules are.

$$d(\mathrm{A,B}) = \sum_{i\in A}\sum_{j\in B} d_{ij} \quad (2)$$

Equation 2 raises an important question: which BCPs of A should be compared to which BCPs of B? One answer is to compare every BCP in A with every BCP in B. Although such a total distance provides a valid measure between two *entirely different* molecules, it would not be an effective distance to gauge the similarity between a set of congeneric molecules that typically appears in QSAR studies. As explained in the next paragraph it is straightforward to include only the distances between two *corresponding* BCPs in molecule A and molecule B. It is significant to realize that this is an a priori matching procedure but that it is a perfectly natural and unbiased mode of operation in most QSAR molecular sequences that one could study.

The "raw" distance defined in eq 2 should actually be modified because the three components constituting this distance have different dimensions. In most accounts on clustering, for example, standardization of the variables to zero mean and unit variance is recommended, using the standard deviation from the complete set of entities.[28] This means that a variable x is replaced by $(x - \mu)/\sigma$, where x is ρ_b, $\nabla^2\rho_b$, and ϵ. This can, however, have the serious effect of diluting differences between groups on the variables that are the best discriminators. Preliminary work on the example below, however, shows that our overall conclusion is only moderately dependent on whether the variables are standardized or not.

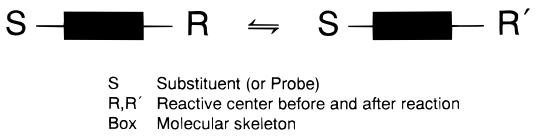

Figure 3. General schematic representation of the essential constituents of a set of molecules studied in connection with a linear free energy relationship.

6. The Hammett Equation

Here is not the place to extensively review the Hammett equation since it is well-known physical organic chemistry textbook material, but a few key points must be made to clarify the terrain of action of our method. Already in the thirties it was observed that equilibrium constants K of reactions of compounds differing only in a substituent were in fact correlated in a simple way. To express this remarkable correlation, Hammett introduced a substituent constant σ as the logarithm of the ratio of the ionization constant of a substituted benzoic acid (substituent S) to that of benzoic acid itself in water solution at 25 °C as

$$\sigma = \log \frac{K_S}{K_H} = \mathrm{p}K_{a,H} - \mathrm{p}K_{a,S} \quad (3)$$

Obviously, the σ value for benzoic acid itself is zero. Figure 3 summarizes in an abstract way the general situation that a relation such as eq 3 describes. The reaction at hand is S−Ph−COOH + H_2O = S−Ph−COO^- + H_3O^+ where Ph represents the phenyl group. Clearly, the phenyl group is the molecular skeleton marked in Figure 3, COOH is the reactive center R and the substituents are given below, e.g., OCH_3. Hammett's relationship can be generalized to other reactions with different molecular skeletons and sets of substituents, thereby introducing the reaction constant ρ (not to be confused with the electron distribution), which is set to unity for benzoic acid.

If the equilibrium constants for several examples of a particular reaction of aromatic molecules are known, the Hammett equation can be used to estimate the equilibrium constants with *different* ring substituents with known σ constants.[29] The beauty and perhaps mystery of the Hammett equation is that we can predict the acidity of a molecule without a detailed understanding of how the reaction equilibrium is reached. A complete *ab ovo* prediction would require a sophisticated molecular dynamics simulation at a time the liquid structure of even pure water is still not fully understood. Everything seems to be governed by the electronic structure that the substituent transmits to the reactive center via the molecular skeleton. That regularities embodied in the Hammett equation appear at all is even more curious, realizing that the equilibrium constant depends on ΔG, which depends not only on enthalpy ΔH but also on the organization energy $T\Delta S$. One still does not completely understand just why the Hammett equation is so generally successful.[1]

Tables 1 and 2 quote the σ values of a selection of common substituents (or probes) attached to a benzoic acid in *para* and

TABLE 1: Selection of σ Values for a Few Common Substituents Attached to Benzoic Acid in the *Para* Position[a]

substituent	A	B	C	D	E	F
NH_2	−0.57	−0.66	−0.66	−0.66	−0.66	−0.66
OCH_3	−0.28	−0.268	−0.27	−0.268	−0.27	?
CH_3	−0.14	−0.17	−0.14	−0.17	−0.17	−0.17
H	0	0	0	0	0	0
F	0.15	0.062	0.15	0.062	0.06	0.06
Cl	0.24	0.227	0.24	0.227	0.23	0.28
CN	0.70	(1.00)	0.71	0.660	0.63	0.66
NO_2	0.81	0.778	0.78	0.778	0.78	0.78

[a] Sources: A, March, J. *Advanced Organic Chemistry*, 4th ed.; 1992; Table 4, p 244. B, Hammett, L. P. *Physical Organic Chemistry*, 1940; Table I, p 188. C, Isaacs, N. *Physical Organic Chemistry*, 2nd ed; 1995; Table 4.1, p 152. D, Carroll, F. A. *Structure and Mechanism*, 1998; Table 6.8, p 384. E, Traven, V. A. *Frontier Orbitals and Properties of Organic Molecules*, 1992; Table 1.4, p 8. F, Miller, B. *Advanced Organic Chemistry*, 1997; Table 5.1, p 124.

TABLE 2: Selection of σ Values for a Few Common Substituents Attached to Benzoic Acid in the *Meta* Position[a]

substituent	A	B	C[b]	D	E	F
NH_2	−0.09	−0.161	−0.40	−0.16	−0.16	-0.16
CH_3	−0.06	−0.069	−0.07	−0.069	−0.07	-0.07
H	0	0	0	0	0	0
OCH_3	0.10	0.115	0.11	0.115	−0.12	?
OH	0.13	?	0.12	0.121	−0.002	0.12
CN	0.62	0.678	0.59	0.56	0.68	0.56
NO_2	0.71	0.710	0.75	0.710	0.71	0.71

[a] Legend of sources is the same as in Table 1. [b] "Primitive" (unaveraged) values.

meta position, respectively. There are some discrepancies in the listed values depending on the cited source, but all figures agree that the *para* substituents should be ranked as follows: $NH_2 < OCH_3 < CH_3 < H < F < Cl < CN < NO_2$. Discrepancies again occur for *meta* substituents affecting the ranking only once but the following order ensues from the majority of data: $NH_2 < CH_3 < H < OCH_3 < OH < CN < NO_2$. In the next section we show how these rankings can be exactly predicted using the proposed Euclidean distance similarity measure in BCP space. Note that the OCH_3 group appears at two different sides of the H substituent, i.e., in the *para* group compared to the *meta* group. This difference in ranking will be correctly predicted.

7. Example of the Application of BCP Space

We have looked at a set of *para* and a set of *meta* substituted benzoic acids, which we will discuss in turn. All ab initio wave functions were obtained at the B3LYP/6-311+G**//B3LYP-6-311+G** [30,31] level using GAUSSIAN94.[32] The topological analysis of the electron distribution was performed using MORPHY98.[33] Once all the BCPs are acquired, it is straightforward to find a one-to-one correspondence between the BCPs of each molecule with respect to the BCPs of another molecule. The maximal common subset of BCPs for which this correspondence can be established is the union of the phenyl ring and the carboxyl group because the substituents differ widely in atom type. We measured the distance between all the molecules using eq 2, yielding a matrix (for an example, see Table 3).

How do we interpret a matrix of distances between molecules in terms of a one-dimensional ranking? The full extent of this problem cannot be tackled in this work, but it suffices here to focus on one molecule (i.e., one column in the distance matrix) and use the distances with respect to that molecule to rank all molecules. A natural choice for the reference molecule is the first or last member of a sequence, such as the NH_2 or the NO_2 substituent. Even if this substituent is not known a priori to be the bound of a sequence, it will emerge automatically from inspecting the distance matrix. Table 4 represents the substituent sequences with respect to NH_2.

TABLE 3: Matrix Containing the Distances in BCP Space between the *Para* Substituted Benzoic Acids[a]

	NH_2	OCH_3	CH_3	H	F	Cl	CN	NO_2
NH_2	0							
OCH_3	0.039	0						
CH_3	0.066	0.031	0					
H	0.090	0.052	0.027	0				
F	0.135	0.103	0.098	0.080	0			
Cl	0.158	0.128	0.123	0.104	0.025	0		
CN	0.261	0.232	0.228	0.207	0.130	0.105	0	
NO_2	0.305	0.276	0.272	0.251	0.175	0.149	0.044	0

[a] Only the 3 BCPs of the COOH group contribute to the distances.

The main point proven in Table 4 is that the experimental substituent sequence is *only* reproduced if we restrict our similarity measure (i.e., distance in BCP space) to BCP contributions (see eq 2) from the COOH group. Any inclusion of BCPs from the phenyl group will seriously disrupt the sequence. Consequently, our method points out where the reactive center is for a given QSAR. Table 3 shows the distance matrix for all eight molecules based on contributions from the carboxyl group alone. It is clear from this table that the experimental sequence is perfectly reproduced with respect to any substituent. In other words each column confirms the ranking obtained with respect to NH_2. The same is true had we restricted ourselves to C═O only. However, if we include only the O—H BCP then the substituents F and Cl are swapped if we take the molecules with the following substituents as a reference: F, Cl, H, and CH_3. Further observations confirm that it is better to take the molecules at either side of the activity scale (weakest and strongest) as references to be able to rank the substituents. In other words, minor anomalies in the reproduction of the experimental ranking may occur if the distances are computed with respect to molecules with moderate activity (i.e., ones from the center). The main problem is how to rigorously obtain a one-dimensional ranking from a (two-dimensional) matrix. Perhaps the substituents ought to be represented in a two-dimensional space, as first suggested by Craig, and our distances correlated to the "experimental" distances appearing in the Craig plot.[34] A complete understanding probably requires an analysis using multidimensional scaling (MDS).[35]

To further test the success of our method, we have generated the wave functions of five more *para* substituted benzoic acids: $COCH_3$, CHO, phenyl (Ph), OH, and O^-. The sources cited in Table 5 quote an experimental σ value ranging from 0.44 to 0.52 for $COCH_3$ and [0.42, 0.45] for CHO. Both $COCH_3$ and CHO are invariably bracketed by CN and Cl, which is exactly what our method predicts with respect to any reference molecule. The experimental σ range quoted for phenyl is [−0.01, +0.05], many values being extremely close to zero. Therefore phenyl and hydrogen are hard to distinguish. In the distance matrix for phenyl six columns predict that Ph is bracketed by H and CH_3, and one column predicts that Ph is bracketed by H and F. In other words, Ph is predicted to lie on the wrong side of H with respect to most substituents. If we adhere to the most recent experimental σ value of −0.01 (source D in Table 5) then we predict Ph to be on the right side of H. The difference between Ph and H is so subtle that it is at the limit of what our method can currently offer.

TABLE 4: Ranking of *Para* Substituted Benzoic Acids According to Their σ Values Based on Experiment and the Distance Similarity Measure in BCP Space Restricted to Various Subsets of BCPs[a]

BCP subset	no. of BCPs	substituents[b]							
		NH_2	OCH_3	CH_3	H	F	Cl	CN	NO_2
OH	1	NH_2	OCH_3	CH_3	H	F	Cl	CN	NO_2
C═O	1	NH_2	OCH_3	CH_3	H	F	Cl	CN	NO_2
COOH	3	NH_2	OCH_3	CH_3	H	F	Cl	CN	NO_2
C−COOH	4	NH_2	OCH_3	F	CH_3	Cl	H	CN	NO_2
C_6	6	NH_2	CN	CH_3	OCH_3	H	Cl	NO_2	F
C_6H_4	10	NH_2	CH_3	OCH_3	CN	H	Cl	NO_2	F
C_6−COOH	10	NH_2	OCH_3	CH_3	Cl	CN	H	NO_2	F
C_6H_4−COOH	14	NH_2	OCH_3	CH_3	Cl	H	CN	NO_2	F
S−C_6H_4−COOH	15	NH_2	CN	H	OCH_3	CH_3	NO_2	F	Cl

[a] The ranking is relative to NH_2. The letter S denotes a general substituent attached to the phenyl ring (see Figure 3). [b] Experimental sequence.

TABLE 5: Experimental σ Values for the Five Substituents the Brackets of Which Were Theoretically Predicted[a]

substituent	A	B	C	D	E	F
O^-	−0.81	?	?	?	−0.52	−0.81
OH	−0.38	?	−0.22	−0.37	−0.36	−0.37
phenyl (Ph)	0.05	0.009	0.05	−0.01	0.01	?
$COCH_3$	0.44	?	0.47	0.502	0.52	?
CHO	?	?	0.45	?	0.45	0.42

[a] Legend of sources is the same as in Table 1.

TABLE 6: Ranking of *Meta* Substituted Benzoic Acids According to Their σ Values Based on Experiment and the Euclidean Distance Similarity Measure in BCP Space Restricted to Various Subsets of BCPs[a]

BCP subset	no. of BCPs	substituents[b]						
		NH_2	CH_3	H	OCH_3	OH	CN	NO_2
OH	1	NH_2	CH_3	OCH_3	H	OH	CN	NO_2
C═O	1	NH_2	CH_3	H	OCH_3	OH	CN	NO_2
C−C(OOH)	1	NH_2	CH_3	H	OCH_3	OH	NO_2	CN
COOH	3	NH_2	CH_3	H	OCH_3	OH	CN	NO_2
C−COOH	4	NH_2	CH_3	H	OCH_3	OH	CN	NO_2
C_6	6	NH_2	CH_3	CN	OCH_3	H	OH	NO_2
C_6H_4	10	NH_2	CH_3	CN	OCH_3	H	NO_2	OH
C_6−COOH	10	NH_2	CH_3	CN	OCH_3	H	OH	NO_2
C_6H_4−COOH	14	NH_2	CH_3	CN	OCH_3	H	NO_2	OH
S−C_6H_4−COOH	15	NH_2	H	CH_3	OCH_3	NO_2	OH	CN

[a] The ranking is relative to NH_2. The letter S denotes a general substituent attached to the phenyl ring. [b] Experimental sequence.

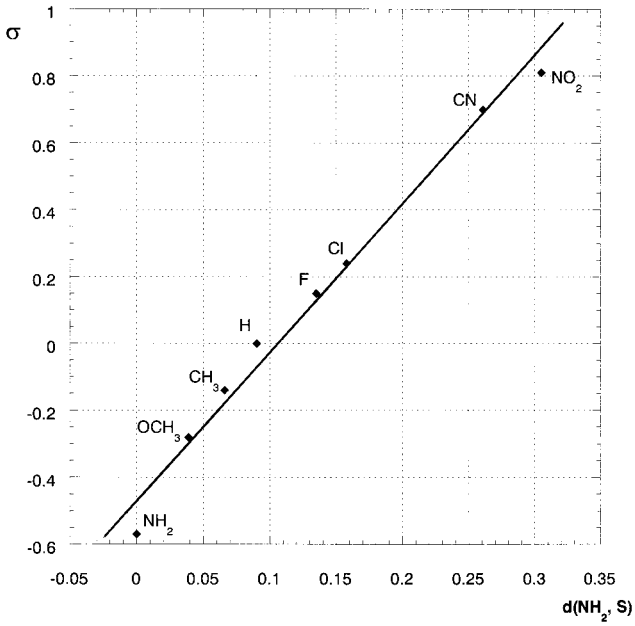

Figure 4. Simple regression analysis for the eight original *para* substituted benzoic acids (column A in Table 1). The experimental σ parameter is plotted against the proposed similarity distance $d(NH_2,S)$, where S is a substituent. The distance is computed via eq 2 and the reference substituent is NH_2, which has the lowest activity. The correlation coefficient is 0.993.

Until 1960 the experimental σ value for OH and O^- was unavailable because of experimental difficulties.[36] Hine was able to present values for these substituents indirectly (using a product rule) and proposed $\sigma(OH) = -0.21$ and $\sigma(O^-) = -0.71$, although other values have also been quoted (Table 5). When we accept Hine's OH value, our method correctly predicts that this substituent is bracketed by OCH_3 and CH_3. The case of O^- is interesting because it offers for the first time the possibility of extrapolation rather than interpolation. Indeed, the experimental σ value puts it outside the original [NH_2, NO_2] bracket, i.e., left of NH_2. We completely recover the correct experimental sequence from the point of view of the O^- substituent because the distance between O^- and NH_2 is the smallest of all and increases monotonically through the sequence $OCH_3 < CH_3 < H < F < Cl < CN < NO_2$.

We now discuss the results for the *meta* substituted benzoic acids. Note that we have deliberately included OH and OCH_3 because they appear at the NO_2 side of H instead of at the NH_2 side in the case of *para* benzoic acids. This is important to prove that our method is reliable in predicting such differences, as indeed it is. Table 6 shows substituent sequences based on various subsets of contributing BCPs. Again we retrieve the main result that the reactive center COOH perfectly matches the experimental sequence. Just as in the *para* case, the C═O sequence shares this property but the O−H sequence fails to do so. It is curious to see that a BCP subset including the COOH group and the C−C bond attaching COOH to the phenyl ring also matches the experimental sequence unlike in the *para* case. The ranking based on the C−C BCP alone does not reproduce the experimental sequence, however. In summary, that the experimental sequence is only recovered from BCPs belonging to the reactive center holds in both *para* and *meta* cases. However, a fully automated and exhaustive search corroborating this conclusion is warranted.

Finally, we have performed a simple one-dimensional linear regression of the experimental σ values versus the computed similarity measure of eq 2. Figure 4 shows a regression plot of the eight original *para* substituted benzoic acids (column A in Table 1) against the proposed similarity distance $d(NH_2, S)$, where *S* is a substituent. The distance is computed with respect to the NH_2 substituent. The Pearson correlation coefficient for this particular fit is 0.993, but similar values have been obtained from alternative sources of experimental values (see legend of Table 1). The lowest correlation obtained is 0.962 for the experimental sequence marked *C* in Table 1 versus the distances computed with respect to the most active substituent, i.e., $d(NO_2,$

S). In summary, regression analyses yield excellent results, enabling fully quantitative QSAR predictions to be made from BCP space.

8. Arising Issues

Ultimately, we want to use our similarity measure in the context of drug design. For example, Lawrence and co-workers have synthesized a few dozen substituted (*E*)-1-phenylbut-1-en-3-ones and tested their cell growth inhibitory properties in terms of the antitumor activity index IC_{50}.[37] With present day computers it is perfectly feasible to obtain wave functions for all these compounds within a week and to represent them in BCP space. Most encouraging results have already been obtained for the Lawrence QSAR where we were able to confirm a conjecture about the active center of the drug.[38] We expect our similarity index to be successful in other anticancer drugs and to be able to bracket a new substituted drug by substituted drugs already present in current experimental sequences. Of course, regression analyses, which are ubiquitous in QSAR work, will be performed in more detail. That BCPs are reliable to measure inductive and mesomeric effects in aromatic rings caused by various substituents was already known to Bader and Chang[39] in the late eighties based on a careful study of electrophilic aromatic substitution and the Taft resonance parameter σ_R^o. Here we have put forward the main idea behind our approach, which was illustrated via a simple but powerful example.

As our intended research program unfolds, several issues appear that need careful investigation. Some are already under investigation and will be published in due course. The following four questions call for systematic study: What is the actual dimension of BCP space? Or, more precisely, which BCP properties contribute to the best possible reproduction of an experimental sequence? Also we must ensure that the included properties are actually independent. It may turn out that BCP spaces of different dimensionalities are needed for different QSARs. Principal component analysis (PCA)[40] and multidimensional scaling (MDS) are appropriate statistical tools to tackle these concerns. The second question is how reliable BCP space really is on a practical level, in particular with regard to large systems. Basis set variation, transferability studies, and cluster techniques will help to settle this matter. Some issues raised in the first two questions have already been investigated.[41] The third question is whether the Euclidean distance (including the standardization of the variables) is the best similarity measure for our purposes. Indeed, alternative measures[26] operating in BCP space may have a higher discriminative capability. The final question is how conformational changes appear in BCP space. We have deliberately quenched this degree of complexity by looking at fairly rigid systems, but many molecules such as neurotransmitters are known to be conformationally flexible.

9. Conclusion

Molecular similarity measures are important to guide us in the hunt for new medicines and agrochemicals. Quantum similarity measures have been proposed before under the hypothesis that molecular properties can ultimately be reduced to the electron distribution. We believe that they are unnecessarily cumbersome and biased by chemically unimportant regions. Consequently, we propose a novel quantum similarity measure in BCP space. This abstract space is based on the theory of atoms in molecules, which enables one to rigorously extract chemical information from a wave function and represent it compactly and reliably. Using a simple distance in a 3D BCP space (with components ρ_b, $\nabla^2\rho_b$, ϵ) we measure the similarity between molecules. Given a set of substituted congeners, the resulting distance matrix allows us to rank the molecules according to their activity. *The experimental activity sequence will only be reproduced if the distance measure is confined to contributions from the BCPs from the common active center of the molecules.* This approach is general and can be applied whenever an experimental QSAR is available. We have illustrated the success of this approach on the oldest and probably best known QSAR: the Hammett equation for *para* and *meta* substituted benzoic acids. In view of the scope of the present research program, several issues could only be touched upon. However, our approach is so simple and promising that we plan to apply the method to real life medicinal questions.

Acknowledgment. Mr. S. E. O'Brien is thanked for helpful discussions and for preparing Figures 1 and 4. Dr. D. Kosov is thanked for his help with Figure 2b.

References and Notes

(1) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR*; ACS professional reference book; American Chemical Society: Washington, DC, 1995.
(2) Popelier, P. L. A. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman & Hall: London, 1995; p 174.
(3) Moos, W. H.; Green, G. D.; Pavia, M. R. *Annu. Rep. Med. Chem.* **1993**, *28*, 315.
(4) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E., Jr.; Blaney, J. M. In *Reviews in Computational Chemistry*, Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1997; Vol. 10.
(5) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431.
(6) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon: Oxford, U.K., 1990.
(7) Bader, R. F. W.; Popelier, P. L. A.; Keith, T. A. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 620.
(8) Bader, R. F. W.; Beddall, P. M. *J. Chem. Phys.* **1972**, *56*, 3320.
(9) Bader, R. F. W. *Acc. Chem. Res.* **1985**, *18*, 9.
(10) Carbó, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185.
(11) Solà, M.; Mestres, J.; Carbó, R.; Duran, M. *J. Am. Chem. Soc.* **1994**, *116*, 5909.
(12) Richards, W. G.; Hodgkin, E. E. *Chem. Br.* **1998**, *24*, 1141.
(13) Cooper, D. L.; Mort, K. A.; Allan, N. L.; Kinchington, D.; McGuigan, C. *J. Am. Chem. Soc.* **1993**, *115*, 12615.
(14) Duane-Walker, P.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1991**, *12*, 220.
(15) Abramov, Y. A. *Acta Crystallogr.* **1997**, *A53*, 264.
(16) Hodgkin, E. E.; Richards, W. G. *J. Chem. Soc., Chem. Commun.* **1986**, 1342.
(17) Dean, P. M. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds. Wiley: New York, 1990; p 211.
(18) Lee, C.; Smithline, S. *J. Phys. Chem.* **1994**, *98*, 1135.
(19) Bowen-Jenkins, P. E.; Richards, W. G. *J. Chem. Soc., Chem. Comm.* **1986**, 133.
(20) Richard, A. M.; Rabinowitz, J. R. *Int. J. Quantum Chem.* **1987**, *31*, 309.
(21) Hodgkin, E. E.; Richards, W. G. *Int. J. Chem.* **1987**, *14*, 105.
(22) Popelier, P. L. A. *Chem. Phys. Lett.* **1994**, *228*, 160.
(23) Wiberg, K. B.; Bader, R. F. W.; Lau, C. D. *J. Am. Chem. Soc.* **1987**, *109*, 985. Bader, R. F. W.; Slee, T. S.; Cramer, D.; Kraka, E. *J. Am. Chem. Soc.* **1983**, *105*, 5061.
(24) Boyd, R. J.; Choi, S. C. *Chem. Phys. Lett.* **1986**, *129*, 62.
(25) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. *J. Am. Chem. Soc.* **1983**, *15*, 5061.
(26) Johnson, M. A. *J. Math. Chem.* **1989**, *3*, 117.
(27) Hô, M.; Schmider, H.; Edgecombe, K. E.; Smith, V. H., Jr. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1994**, *28*, 215.
(28) Everitt, B. *Cluster Analysis*, 2nd ed.; Wiley: New York, 1980.
(29) Miller, B. *Advanced Organic Chemistry;* Prentice Hall: Englewood Cliffs, NJ, 1997.
(30) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(32) Gaussian 94, Revision B.1; M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanayakkara, M. Challacombe, C. Y. Peng, P. Y. Ayala,

W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople; Gaussian, Inc.: Pittsburgh, PA, 1995.

(33) MORPHY98, a program written by P. L. A. Popelier with a contribution from R. G. A. Bone (UMIST: Manchester, England, EU, 1998); http://www.ch.umist.ac.uk/morphy.

(34) Craig, P. N. *J. Med. Chem.* **1971**, *14*, 680.

(35) Kruskal, J. B.; Wish, M. *Multidimensional Scaling*; Sage: 1978.

(36) Hine, J. *J. Am. Chem. Soc.* **1960**, *82*, 4877.

(37) Ducki, S.; Hadfield, J. A.; Hepworth, L. A.; Lawrence, N. J.; Liu, C−Y.; McGown, A. T. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 3091.

(38) S. O'Brien and P. Popelier, unpublished results.

(39) Bader, R. F. W.; Chang, C. *J. Phys. Chem.* **1989**, *93*, 2946.

(40) Livingstone, D. *Data Analysis for Chemists. Applications to QSAR and Chemical Product Design;* Oxford University Press: Oxford, U.K., 1995.

(41) O'Brien, S.; Popelier, P. *Can. J. Chem.* **1999**, *77*, 28.