J|A|C|S
ARTICLES
Published on Web 01/09/2003

Atomic Properties of Selected Biomolecules: Quantum Topological Atom Types of Carbon Occurring in Natural Amino Acids and Derived Molecules

P. L. A. Popelier* and F. M. Aicken

Contribution from the Department of Chemistry, UMIST, Manchester M60 1QD, England

Received September 5, 2002; E-mail: pla@umist.ac.uk.

Abstract: We seek to recover rigorous atom types from amino acid wave functions. The atom types emerge from a cluster analysis operating on a set of seven atomic properties, including kinetic energy, volume, population, and dipole, quadrupole, octupole, and hexadecapole moments. These properties are acquired by partitioning the molecular electron density into quantum topological atoms. Wave functions are generated at the B3LYP/6-311+G(2d,p)//HF/6-31G(d) level for a sensible conformation of each of the 20 naturally occurring amino acids and smaller derived molecules, which together constitute a data set of 57 molecules. From this set 213 unique quantum topological carbons are obtained, which are linked according to the similarity of their properties. After introducing a statistical separation criterion, our cluster analysis proposes two representations: a cruder one with 5 atom types and a finer one with 21 atom types. The immediate coordination of the central carbon plays a major role in labeling the atom types.

Introduction

The viability of discerning atom types and functional groups is one of the cornerstones of chemistry. That atoms preserve their characteristics under similar chemical surroundings enables chemistry to be a science of rational classification rather than a collection of disparate facts. A clear definition of a property of an atom inside a molecule is a prerequisite for the success of this enterprise. Hence the recovery of atom types from contemporary wave functions is a vital challenge to theoretical chemistry. The definition of atom types enables the parametrization of force fields and endows them with a predictive power that prevents their becoming mere databases. However, in force field design atom types are typically *imposed*, on the basis of basic chemical intuition, rather than directly *derived* from the wave function.

In this paper we focus on amino acids and smaller fragment molecules to be specified later, which we systematically partition into atoms. To this end, we employ quantum chemical topology (QCT) as embodied in the theory of atoms in molecules (AIM).[1] This approach prescribes a partitioning in real space based on the gradient of the electron density. We obtain a large set of unique quantum topological atoms (each described by atomic properties), in which we discover atom types by means of cluster analysis. As such we compute atom types rather than determine them a priori with potential prejudice.

The computational procedure described below also enables in principle the study of transferability, but in this work *we confine ourselves to a detailed and careful description of atomic type clusters.* Several other workers have focused on transferability, for example in the context of the electrostatic potential of polypeptides,[2] point charge models for amino acid side chains,[3] and electrostatic interactions of peptides and amides[4] or in connection with a molecular electron density "Lego" approach to molecule building.[5] Yet others shared our interest in quantum chemical topology to investigate amino acids[6,7] or studied the transferability of alkyl chains in aldehydes and ketones[8] and of methyl and methylene fragments in alkyl monoethers[9] and examined approximate transferability in alkanols[10] and alkanenitriles.[11] The concept of compensatory transferability was recently introduced[12] and illustrated for the linear homologous series of hydrocarbons and polysilanes and for the formation of pyridine from fragments of benzene and pyrazine. The work presented in this paper is closest to that of the group of Breneman who proposed the so-called transferable atom equivalent (TAE) method.[13]

However in this paper we will not focus on the transferability that our atom types offer. Instead the work presented here serves

* To whom correspondence should be addressed. Telephone: +44-161-2004511. Fax: +44-161-2004559.

(1) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon, Oxford, Great Britain, 1990.
(2) Price, S. L.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755−1763.
(3) Chipot, C.; Angyan, J. G.; Maigret, B.; Scheraga, H. A. *J. Phys. Chem.* **1993**, *97*, 9788−9796.
(4) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 4915−4926.
(5) Walker, P. D.; Mezey, P. G. *J. Am. Chem. Soc.* **1994**, *116*, 12022−12032.
(6) Matta, C. F.; Bader, R. F. W. *Proteins: Struct., Funct. Genet.* **2000**, *40*, 310−329.
(7) Matta, C. F.; Bader, R. F. W. *Proteins: Struct., Funct. Genet.* **2002**, *48*, 519−538.
(8) Grana, A. M.; Mosquera, R. A. *J. Chem. Phys.* **2000**, *113*, 1492−1500.
(9) Vila, A.; Mosquera, R. A. *J. Chem. Phys.* **2001**, *115*, 1264−1273.
(10) Mandado, M.; Grana, A. M.; Mosquera, R. A. *J. Mol. Struct. (THEOCHEM)* **2002**, *584*, 221−234.
(11) Lopez, J. L.; Mandado, M.; Grana, A. M.; Mosquera, R. A. *Int. J. Quantum Chem.* **2002**, *86*, 190−198
(12) Bader, R. F. W.; Bayles, D. *J. Phys. Chem. A* **2000**, *104*, 5579−5589.
(13) Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. *Comput. Chem.* **1995**, *19*, 161−179.

10.1021/ja0284198 CCC: $25.00

as an unbiased guide for future force field design or improvement. The ultimate power of atom types should be measured via interaction energies and geometry prediction, which is the subject of future work that needs to be very systematic and incremental in order to be truly successful in the long term. The value of the current study is that it is unbiased by chemical intuition and a rigorous attempt to recover atom types from modern wave functions.

After a careful study of the integration procedure used to obtain atomic properties,[14] we are now in a position to report on a detailed cluster analysis of a set of 760 atoms drawn from one conformation from each of the 20 most common naturally occurring amino acids, as well as smaller derived molecules. In this paper we only focus on the 213 carbons, the richest group of atom types, while results on hydrogen, nitrogen, oxygen, and sulfur are reported elsewhere.[15]

Quantum Chemical Topology

We review only one aspect of this increasingly popular approach[16,17] since instructive accounts can be found in textbooks.[18,19] The gradient of the electron density, ρ, is the key to understanding the way a molecule is partitioned. A continuous sequence of infinitesimally short segments of the gradient vector, each time reevaluated at its endpoint, forms a gradient path. A gradient path moves in the direction of steepest ascent in ρ until it reaches an attractor, which typically coincides with a nucleus. The infinite number of gradient paths attracted to one nucleus constitutes an atom. Such an atom is a portion of 3D space that takes a unique shape dictated by its environment. Atomic properties are then obtained as an integral over the atomic volume of a property density. The atomic population is defined as the integral of ρ over the atomic volume, while the atomic kinetic energy is the integral over the kinetic energy density. Note that we are *not* using the total atomic energy, often obtained by multiplying the atomic kinetic energy by a correction factor $(1 + \gamma)$, where γ is the virial ratio $-V/T$, which should be equal to 2 for an equilibrium geometry. For split-level finite basis calculations (leading to small nonvanishing forces on the nuclei), it is always correct to refer to the atomic kinetic energy. The atomic multipole moments are defined within the compact spherical tensor formalism.[20] Hence there are only three, five, seven, and nine components of the dipole, quadrupole, octupole, and hexadecupole moment, respectively. It has been proven[21,22] that these moments suffice to reproduce the atomic electrostatic potential at the "water-accessible surface" with a root-mean-square accuracy of less than 0.1 kJ/mol. They also form the basis of a topological intermolecular potential that predicts the geometries of van der Waals complexes[23] and DNA base pairs.[24] A comparison between the components of the multipole moments would require keeping track of their orientation and a convention for maximum alignment. To avoid these complications, we proceeded with orientationally invariant magnitudes. In summary each atom is represented by seven (scalar) atomic properties. The atomic volume was obtained by capping the atoms by the $\rho = 0.001$ au contour.

There is no deep reason we constrain our work to these seven properties, other than they are popular in the QCT community. The electrostatic potential can be used as another or extra measure to assess similarity, although this involves arbitrary grids. Work in progress explores this avenue for lysine and retinal and for derived molecules.

Dataset Generation

A set of 57 molecules was generated including the 20 most common naturally occurring free amino acids and smaller derived molecules. The latter were generated by cleaving the $C_\alpha-C_\beta$ bond of each amino acid and capping the side chain fragment with a hydrogen atom. Subsequently the $C_\beta-C_\gamma$ bond was cleaved, creating two fragments of which the larger one was again capped with a hydrogen atom. This process of cleaving and capping was applied for consecutive *single* bonds within each side chain until the smallest possible molecule was arrived at. For example, aspartic acid, $H_2N-HC_\alpha(C_\beta H_2C_\gamma(=O)-OH)-COOH$, gives rise to acetic acid, $H-C_\beta H_2C_\gamma(=O)-OH$; formic acid, $HC_\gamma(=O)-OH$; and H−OH or water. We call such a set of molecules derived from a given amino acid a *family* and employ the standard amino acid letter code to label the molecules of the same family. For example, aspartic acid is denoted by *D4*, acetic acid by *D3*, formic acid by *D2*, and water by *D1*. Note that molecular hydrogen could have been a member of the *D* family but appears as a member of the *G* family (Glycine is *G2*) and is hence designated by *G1*. Double bonds and ring structures were left intact, and duplicated molecules were discarded. The bonds occurring in the set of 57 molecules have been characterized before[25] via their so-called bond critical point properties in the context of molecular similarity.

Determination of Atom Types

Via the topological analysis we obtain a multitude of unique atoms, each expressed in a seven-dimensional space of atomic properties. The uniqueness of the atoms illustrates that perfect transferability is an unattainable limit.[26] It is only when we start grouping atoms into clusters that they obtain a degree of transferability. In other words, the information that the atom groups contain is averaged out or "blurred" to such an extent that this information may be safely carried over to a different molecular environment. We accomplish this by a mathematical technique called cluster analysis, which is briefly reviewed in Appendix 1.

By lumping atoms together cluster analysis yields *atom types*, each of which corresponds to a cluster. However cluster analysis itself does not provide a criterion for determining the number

(14) Aicken, F. M.; Popelier, P. L. A. *Can. J. Chem.* **2000**, *78*, 415−426.
(15) Popelier, P. L. A.; Aicken, F. M. Submitted for publication.
(16) Popelier, P. L. A.; Aicken, F. M.; O'Brien, S. E. In *Chemical Modelling: Applications and Theory*, Vol. 1; Royal Society of Chemistry Specialist, Periodical Report; Hinchliffe, A., Ed.; Royal Society of Chemistry: Letchworth, U.K., 2000; Chapter 3, pp 143−198.
(17) Popelier, P. L. A.; Smith, P. J. In *Chemical Modelling: Applications and Theory*, Vol. 2; Royal Society of Chemistry Specialist Periodical Report; Hinchliffe, A., Ed.; Royal Society of Chemistry: Letchworth, U.K., 2002; Chapter 8, pp 391−448.
(18) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education, London, 2000.
(19) Gillespie, R. J.; Popelier, P. L. A. *Chemical Bonding and Molecular Geometry from Lewis to Electron Densities*; Oxford University Press: New York, 2001.
(20) Zare, R. N. *Angular Momentum*; Wiley-Interscience: New York, 1988.
(21) Kosov, D. S.; Popelier, P. L. A. *J. Chem. Phys.* **2000**, *113*, 3969−3974.
(22) Kosov, D. S.; Popelier, P. L. A. *J. Phys. Chem. A* **2000**, *104*, 7339−7345.
(23) Popelier, P. L. A.; Joubert, L.; Kosov, D. S. *J. Phys. Chem. A* **2001**, *105*, 8254−8261.
(24) Joubert, L.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4353−4359.
(25) O'Brien, S. E.; Popelier, P. L. A. *Can. J. Chem.* **1999**, *77*, 28−36.
(26) Bader, R. F. W.; Becker, P. *Chem. Phys. Lett.* **1988**, *148*, 452−458.

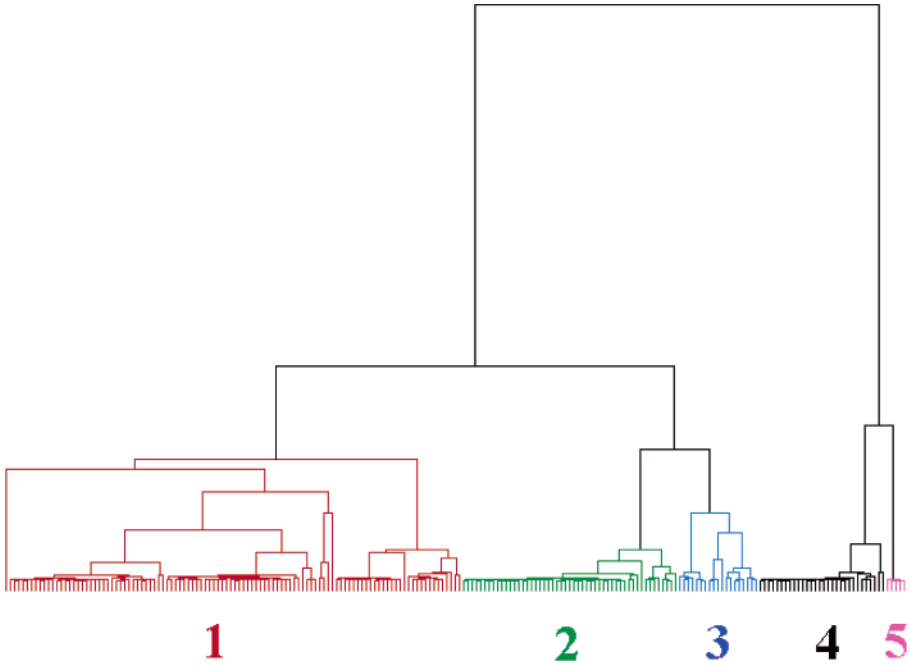

Figure 1. Dendrogram generated by the cluster analysis of carbon defining five atom types.

of clusters one should divide the data set into. Instead it only lays out possible ways in which a dataset can be partitioned into clusters. Nevertheless it is possible to invoke a criterion, external to the cluster analysis, which fixes the number of clusters and hence provides a *representation* of atom types. As explained below via an example, this criterion is purely statistical and ensures that each atom type is sufficiently separated from another. The question of separability of two clusters or (atom types) is best explained via an example, discussed in Appendix 2.

Programs and Computational Methods

All wave functions were generated using the GAUSSIAN94[27] program with **Z**-matrix inputs generated by MOLDEN.[28] The geometries were optimized at the HF/6-31G(d)[29] level with the single point calculation at the B3LYP/6-311+G(2d,p) level.[30,31] After consultation of ref 32, this choice proved to be a good compromise between accuracy and computational cost.[25,33] Matta and Bader found[7] that the HF/6-31+G(d) level recovers the experimental values of the geometric parameters for the side chains of the 20 amino acids with an acceptable degree of accuracy. To maximize conformational similarity among molecules of the same family, each molecule was optimized toward a geometry close to the optimized geometry of another member of the family. All atomic integrations[34,35] were carried out using the program MORPHY98.[36] Some integrations were repeated in order to improve their accuracy, and tables of atomic properties for all atoms are given in Appendix 2 of ref 33.

The hierarchical agglomerative cluster analysis was performed by a program called ClustanGraphics,[37] and initially by SPSS.[38] We use Euclidean distance, which is the default in both computer programs. For the purpose of clustering large datasets, containing more than about 200 cases, it is preferable to use Euclidean distance to calculate the similarity matrix.

Results and Discussion: Classification of Carbon Atom Types

There are 213 unique carbon atoms in our data set. Carbon is a very rich atom in that it is found in many different environments in this particular data set, compared to oxygen or nitrogen for example. Hence it is not surprising that an entire branch of chemistry is devoted to its study. Indeed many different atoms were recovered from the current data set on the basis of amino acids. The stricter intercluster criterion, $\Delta\mu/\Sigma\sigma > 3$, failed for the 6-cluster representation, thereby making the 5-cluster representation the optimal limit. Figure 1 shows the dendrogram for the 5-cluster representation. The number of carbon atoms in each cluster, from one to five, is 108, 51, 19, 30, and 5, respectively, adding up to 213.

The relaxed criterion, given by $\Delta\mu/\Sigma\sigma > 2$, resulted in a 21-cluster representation. The membership of the clusters at *both* the 5- and 21-cluster representation levels is explored in Figure 2. In this figure the carbon atoms are numbered according to their membership in the 21-cluster representation, which will be described below. The numerical labels in Figure 2 are also colored in accordance with the 5-cluster representation as illustrated in the dendrogram shown in Figure 1. For example,

(27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. *GAUSSIAN94*; Gaussian, Inc.: Pittsburgh, PA, 1995.
(28) Schaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123−134.
(29) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.
(30) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785−789.
(31) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(32) Foresman, J. B.; Frisch, A. *Exploring Chemistry with Electronic Structure Methods*; 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996.
(33) Aicken, F. M., Ph.D. Thesis, Department of Chemistry, UMIST, Manchester, Great Britain, 2000.
(34) Popelier, P. L. A. *Mol. Phys.* **1996**, *87*, 1169−1187.
(35) Popelier, P. L. A. *Comput. Phys. Commun.* **1998**, *108*, 180−190.
(36) *MORPHY98*, a program written by P. L. A. Popelier with a contribution from R. G. A. Bone, UMIST, Manchester, England, EU 1998 (http://morphy.ch.umist.ac.uk/ **1998)**.
(37) Wishart, D. *ClustanGraphics* (computer program); Edinburgh, Scotland, GB, 1999.
(38) *SPSS Inc.*, version 10.0.7; Chicago, IL, 2000 (http://www.spss.com).

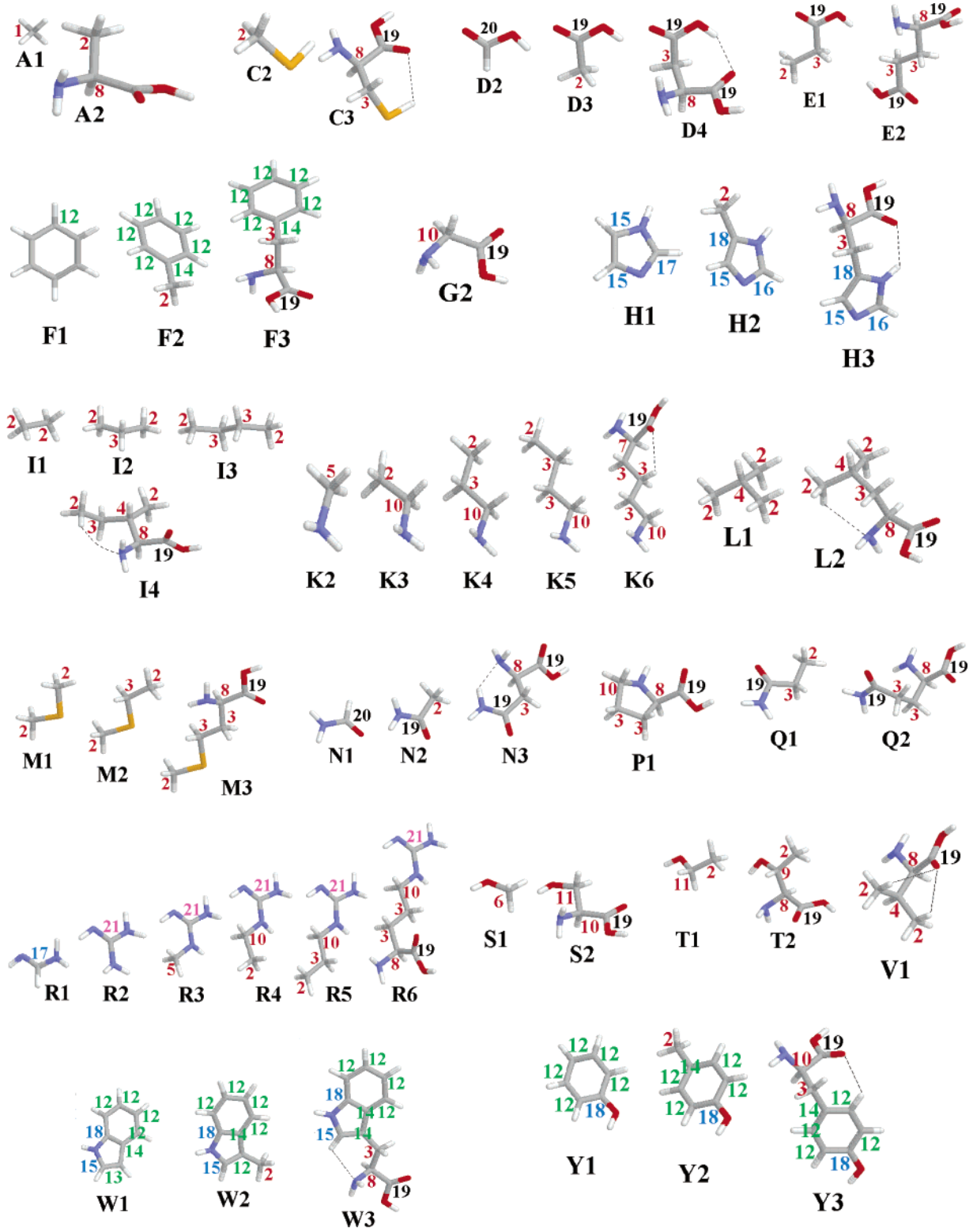

Figure 2. Assignment of carbon's membership to the 5-cluster representation (marked by the colors of Figure 1) and the 21-cluster representation (marked by numerical labels; see main text). Intramolecular hydrogen bonds are marked by a dashed line.

atom types 1 to 11 (according to the 21-cluster representation) are marked in red, which is the color of the first cluster (or atom type) according to the 5-cluster representation.

How can we characterize the clusters in the 5-cluster representation in a direct and unbiased way? This question is equivalent to asking what the atoms belonging to each cluster have in common. Rather than invoking hybridization assignation we use the topology itself to determine the bonded environment of a given atom. According to AIM an atom is bonded to another if it is connected to it via a bond path.[39] Hence we can determine the coordination of an atom by counting the number of bond paths connected to it. The atomic number of the atoms bonded

(39) Bader, R. F. W. *J. Phys. Chem. A* **1998**, *102*, 7314−7323.

Table 1. Mean and Standard Deviations for the Carbon Atom Types of the 5-Cluster Representation

	vol		kinetic energy		population		dipole		quadrupole		octopole		hexadecapole	
cluster	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	59.1	9.8	37.821	0.081	5.897	0.169	0.176	0.159	0.23	0.14	2.76	0.43	3.7	0.8
2	81.3	5.4	37.932	0.014	6.028	0.011	0.131	0.033	1.67	0.09	2.18	0.17	4.9	0.3
3	67.8	8.7	37.672	0.147	5.515	0.259	0.618	0.064	1.61	0.14	2.06	0.31	4.5	0.5
4	36.1	2.8	36.994	0.044	4.457	0.060	0.796	0.038	0.87	0.04	0.98	0.15	2.2	0.3
5	39.3	1.0	37.218	0.010	4.660	0.007	0.088	0.006	1.16	0.02	1.15	0.07	2.8	0.1

Table 2. Correlation Matrix for the Carbon Atom Types of the 5-Cluster Representation

	vol	kinetic energy	population	dipole	quadrupole	octopole	hexadecapole
vol	1						
kinetic energy	0.93	1					
population	0.92	0.99	1				
dipole	−0.30	−0.49	−0.45	1			
quadrupole	0.47	0.16	0.09	0.06	1		
octopole	0.75	0.92	0.94	−0.39	−0.19	1	
hexadecapole	0.98	0.92	0.89	−0.33	0.52	0.74	1

to a given carbon further characterizes the cluster. The following descriptions typify the carbon clusters, where the atoms separated by vertical bars are bonded to the central carbon:

(1) tetracoordinated carbons bonded to any other atom (C or H or N or O or S) (red);

(2) tricoordinated carbons in a ring, bonded to (C or H|C|C) (green);

(3) tricoordinated carbons bonded to (C or H|C|N), (H|N|N), or (C|C|O) (blue);

(4) tricoordinated carbons bonded to (C or H| N or O |O) (black);

(5) tricoordinated carbons bonded to (N|N|N) (purple).

At this crude level of distinction between atoms the tetracoordinated carbons (cluster 1) all belong to the same cluster, regardless of the atom type they are bonded to. Hence we could simply label this cluster as sp^3 carbons. Similarly the second cluster can be labeled as sp^2 carbons bonded to hydrogen or carbon, and the third, as sp^2 carbons bonded to at least one heteroatom (N or O). All carbons of the third cluster belong to a ring except the carbon in molecule *R1*. The fourth cluster contains carboxylic (83%) and amidic (17%) carbons, while the fifth cluster encompasses guanidinic carbons. We note that C and H are interchangeable in view of the 3-fold occurrence of |C or H|. It is difficult to summarize the clustering at this level any further, but one can state the following. Once coordination has split off the sp^3 cluster (cluster 1), the remaining clusters (2−5) are distinguished by the number (0, 1, 2, or 3) of bonded heteroatoms, where a single oxygen perturbs the carbon as much as two nitrogens.

Table 1 charts the mean and standard deviations of all the properties for each of the five different types of carbon. The largest difference between the tetracoordinated carbons (cluster 1) and all tricoordinated carbons (clusters 2−5) occurs in the value of the quadrupole moment. This observation is compatible with the fact that the quadrupole moment is the charge density analogue of a π population in the orbital model.[1,40]

The population decreases from the maximum value of 6.028 au found for sp^2 carbons bonded to C and H (cluster 2) to the minimum value of 4.457 au found for carboxylic and amidic sp^2 carbons (cluster 4). The proximity of heteroatoms to the sp^2 carbons generally results in a lowered population as expected from electronegativity considerations. As expected, the charge transfer from carbon toward oxygen is larger than that toward nitrogen, which would explain why the minimum carbon population occurs in cluster 4 rather than 5. Indeed the guanidinic carbons do not end up with the lowest population despite being bonded to three nitrogens. The observation that the population of sp^3 carbons is lower than that of sp^2 carbons (not attached to heteroatoms) is consistent with the statement that the electronegativity of C relative to H decreases with the degree of saturation.[1] Carboxylic and amidic carbons (cluster 4) and guanidinic carbons (cluster 5) are indistinguishable by their volumes and octopole and hexadecapole moments. Although, technically, their population can separate them, they differ by only 0.203*e*. It is clear that heteroatoms induce a considerable dipole moment, which is however severely diminished by a nearly symmetric bonding environment, as in the guanidinic carbons ($\mu = 0.088$ au).

Table 2 shows the Pearson correlation coefficient[41] r computed for each pair of atomic properties, each property averaged over all carbons in a given cluster (as given in Table 1). The highest correlation is detected between the kinetic energy and the population ($r^2 = 0.99$). Similar correlations have been pointed out before.[13,42,43] Unfortunately this high correlation is of little practical use since the associated spread is as much as 0.05 au or 130 kJ/mol. Another respectable and perhaps surprising correlation, not observed before, is that between the volume and the hexadecapole moment ($r^2 = 0.98$).

We have been careful in avoiding the tag "*aromatic* carbons" to designate cluster 2. To test whether the label *aromatic* is justified, we obtain the atomic properties of carbons in two

(40) Bader, R. F. W.; Chang, C. *J. Phys. Chem.* **1989**, *93*, 2946−2956.

(41) This coefficient is defined as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2)^{1/2}}$$

where $\bar{x}$ and $\bar{y}$ are atomic properties and $\bar{x}$ and $\bar{y}$ their respective means. The sum runs over the number of clusters (i.e. atom types). The values of r lie within [0, 1] (Spiegel, M. R. *Theory and Problems of Statistics*; McGraw-Hill: New York, 1972).

(42) Thompson, T. R. Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY, 1993; p 904.

(43) Cioslowski, J.; Varnali, T. *Int. J. Quantum Chem.* **1999**, *72*, 331−339.

Table 3. Characterization and Description of the Carbon Atom Types for the 21-Cluster Representation

5[a]	21[a]	coordination[b]	description comments	no.[c]	color
1	1	[H|H|H|H]	methane (i.e. methyl bonded to H)	1	red
1	2	[H|H|H|C,S]	methyl bonded to C or S	37	
1	3	[H|H|C|C,S]	methylene bonded to C or S	32	
1	4	[H|C|C|C]	tertiary carbon	4	
1	5	[H|H|H|N]	methyl bonded to N	2	
1	6	[H|H|H|O]	methyl bonded to O	1	
1	7	[H|C|C|N]	C_α in lysine (K6)	1	
1	8	[H|C|C|N]	C_α (in 16 amino acids)	16	
1	9	[H|C|C|O]	C_β in threonine (T2) (secondary alcohol)	1	
1	10	[H|C,H|C|N]	C_α (in G2, S2, or Y3) and methylene bonded to N	11	
1	11	[H|H|C|O]	methylene bonded to O	2	
2	12	[C|C|C,H]	olefinic, in conjugated ring	42	green
2	13	[C|C|H]	olefinic, in conjugated ring	1	
2	14	[C|C|C]	olefinic, in conjugated ring	8	
3	15	[H|C|N]	enaminic carbon (C_2 in indole) ($C_{4,5}$ in imidazole)	7	blue
3	16	[H|N|N]	C_2 in substituted imidazole and C in R1	3	
3	17	[H|N|N]	C_2 in imidazole	1	
3	18	[C|C|N,O]	bridge C in indole; α-C in phenol; C_5 in substituted imidazole	8	
4	19	[C|N,O|O]	amidic/carboxylic bonded to C	28	black
4	20	[H|N,O|O]	amidic/carboxylic bonded to H	2	
5	21	[N|N|N]	guanidinic	5	purple

[a] Numerical label of cluster according to the 5-cluster (Figure 1) or 21-cluster representation. [b] Vertical bars separate the atoms bonded to carbon. Alternatives are separated by a comma. [c] Numbers of atoms in each cluster of the 21-cluster representation.

antiaromatic molecules: cyclobutadiene and cyclooctatetraene.[44] Via its "classification function" the program ClustanGraphics is able to establish which cluster is closest to a new entity. Interestingly we discover that the carbon atoms from the antiaromatic systems are assigned to cluster 2. This conclusion is supported by the fact that all the properties of the two antiaromatic entities lie within 3σ of the mean of cluster 2. Hence the integrated properties do not discern between aromaticity and antiaromaticity. Aromaticity is a notoriously difficult concept to extract from wave functions, and indeed many approaches have been proposed.[45] Although a structural and a magnetic aromaticity index has been successfully related to topological properties,[46] aromaticity is not reflected in the present cluster analysis of carbon's atomic properties. The main message from this excursion is that one should avoid assigning false chemical labels to clusters. In other words, since all carbons of cluster 2 are part of aromatic systems, it is tempting to designate them as such. However a correct cluster label is actually based on the number and type of bonded partners, i.e., the immediate coordinating environment. In a similar vein one may be tempted to designate cluster 3 as one containing all sp^2 carbons in aromatic systems attached to at least one heteroatom. However, this cluster can again not be designated as aromatic because it also contains the carbon in iminoamine (R1), the only atom of this cluster not occurring in a ring.

Intramolecular hydrogen bonding found in molecules C3, D4, K6, N3, V1, and Y3 gives rise to spatially extended topological rings (with six or seven members) that contain sp^2 carbons bonded to a heteroatom. The carbon atoms partaking in such rings are classified in cluster 4 along with the rest of the carbonyl carbons (i.e. carboxylic and amidic), not appearing in a ring. Therefore the current classification is not sensitive enough to detect participation in extended rings. Instead classification is predominantly informing us about the quantity and nature of the neighboring bonded atoms.

We now discuss a more detailed atomic classification based on the $\Delta\mu/\Sigma\sigma > 2$ criterion. The atom types of the 21-cluster representation are described in Table 3. Cluster 1 (sp^3) disintegrates into 11 distinct entities, cluster 2 (sp^2,C/H) into 3 subclusters, cluster 3 (sp^2, N/O) into 4, and cluster 4 (amidic and carboxylic) into 2, and cluster 5 (guanidinic) remains undivided. One should bear in mind that the proximity of the clusters is reflected in their numerical labels. For example, clusters 12, 13, and 14 each contain atom types that are very similar to each other. As such Table 3 displays a gradient of gradually changing carbons from top to bottom. Atoms more electronegative than carbon, such as N and O, become more prominent neighbors as one moves toward the bottom. Guided by the number of atoms in each cluster, given in Table 3, one can recover each cluster in this table from Figure 1. For example, near the bottom of the dendrogram there is a fork that separates cluster 4 into a large group of 28 atoms (left) and a small group of only 2 atoms (right). Detailed inspection of the whole dendrogram enables one to map each cluster of Figure 1 onto Table 3. The first cluster appears at the utmost left and the 21st at the utmost right, and the order (or numerical label) of the clusters in Table 3 corresponds to going from left to right in Figure 1.

Whereas before, at the 5-cluster level, coordination was recovered, we now obtain useful information about the deeper functionality within each of the groups. A first glance at Table 3 reveals that many clusters refer to highly specific carbon environments for which no straightforward functional group names are available. For example, cluster 18 refers to carbons that are part of a conjugated ring system to which a nitrogen or an oxygen is bonded. On the other hand many well-known functional groups such as nitro groups are not recovered because they do not occur in amino acids. Easily identifiable carbons are found in cluster 2, which can be labeled as a methyl group.

(44) Conventional organic chemistry shows that a molecule must also be cyclic and planar to be aromatic. While cyclobutadiene is planar, cyclooctatetraene exists as a tub. A topological analysis of this molecule reveals the presence of two rings, made up of the same constituent atoms, bounding a cage critical point. The existence of a cage critical point bounded by only two rings has been discussed as a mathematical possibility by Bader. However until now the minimum number of rings observed has been three, e.g. bicyclo[1.1.1]pentane.

(45) Garratt, P. J. *Aromaticity*; Wiley: New York, 1986.
(46) Howard, S. T.; Krygowski, T. M. *Can. J. Chem.* **1997**, *75*, 1174−1181.

Methyl carbons bonded to a carbon are not distinct from those bonded to a sulfur. The identical (Pauling) electronegativity of carbon and sulfur most likely contributes to this fact. Methyl carbons bonded to N (cluster 5) however are clearly distinguished from those bonded to O (cluster 6) or H (cluster 1). A similar situation is found with the methylene group, where methylene carbons bonded to C or S (cluster 3) are well-separated from those bonded to N (cluster 10) or O (cluster 11).

The majority of C_α carbons form a cluster of their own (cluster 8 containing 16 members), while C_α of lysine is the only member of cluster 7 and the C_α appearing in glycine, serine, and tyrosine belong to a third cluster (number 10), which also includes all methylene carbons bonded to N. As far as we can see, there is nothing peculiar in the local conformation of these four C_α carbons.

The carbons in clusters 12, 13, and 14 all appear in a conjugated ring system but can be differentiated as follows. The cluster-12 carbons are all bonded to H, except the C_3 carbon in the methyl-substituted indole W2, which is bonded to the methyl group. There is only one carbon atom in cluster 13, which occurs in the 3-position of indole. Cluster-14 carbons appear at the substitution point in six-membered conjugated rings (possibly containing N or O), including the bond of the C_3 atom in the indoles, where the five-membered ring is viewed as a substitution to the six-membered ring. Moreover the C_3 carbon in W3 is of atom type 14.

An equally subtle differentiation occurs between the clusters 15, 16, and 17, where specific positions in well-known compounds, such as imidazole and indole, are singled out as distinct atom types.

The most unexpected cluster is perhaps number 18, which contains atoms as different as the carbon bridge in indole, the α carbon in phenol, and C_5 in substituted imidazole. This cluster houses carbon atoms bonded to both N and O, a property it shares with clusters 19 and 20. Cluster 19 contains the keto carbon both in CC(=O)NH_2 and C**C**(=O)OH, probably because OH and NH_2 are isoelectronic. In a similar vein cluster 20 contains the keto carbon both in H**C**(=O)NH_2 and H**C**(=O)-OH, showing that the presence of H as opposed to C differentiates the keto carbons rather than the presence of the amino versus the hydroxyl group. It is tempting to associate this observation with the similarity in chemical behavior (e.g. acid-catalyzed nucleophilic addition) between carboxylic and amidic groups, which can be related to the fact that they appear in one cluster.

Table S2 of the Supporting Information reports the mean values and standard deviations for the carbon atom types deduced at the 21-cluster level. Compared to the other elements (H, N, O, S) studied and reported on elsewhere, carbon spans a large range in its properties, emphasizing its flexibility and the richness of its appearance. For example, the volume varies from 89.3 au (C_3 in indole, cluster 13) to 35.4 au (in C-bonded amidic and carboxylic group, cluster 19), which corresponds to a factor of 2.5. The population varies from 6.055 in methane to 4.456 in cluster 19, a difference of almost 1.6*e*. Remarkably the same carbon in CC(=O)NH_2 or C**C**(=O)OH (cluster 19) has the largest dipole moment. Adding to this that a cluster-19 carbon has on average the lowest kinetic energy, it is clear that such a carbon is the most perturbed of all.

Some of the trends and extreme values observed are reinforced by the correlation that exists between atomic properties. Table S3 shows the Pearson correlation coefficient r computed for each pair of atomic properties, each property averaged over all carbons in a given cluster of the 21-cluster representation (given in Table 3). A comparison between Tables 2 and S3 (correlation matrix of the 5-cluster and 21-cluster representation, respectively) shows that the correlations between average atomic properties drawn from the clusters of the 21-representation are generally poorer. Two-thirds of the 21 correlation coefficients decrease going from the 5-cluster to the 21-cluster representation, emphasizing poor local correlation. However the correlation coefficients between the dipole moment and all other properties except for the hexadecapole increase.

Some other correlations are not tabulated but discussed now. For example, within the tetracoordinated cluster (number 1 in the 5-cluster representation) there is a high correlation between the dipole moment and the population ($r = -0.99$), the kinetic energy and the population ($r = 0.98$), the kinetic energy and dipole moment ($r = -0.98$), and the quadrupole and dipole moments ($r = 0.95$). All these properties can be related to the electronegativity of the bonding neighbors, where carbon is more electronegative than hydrogen at the current level of calculation. We generally observe that as the number of electronegative atoms bonded to each atom type increases, the population and kinetic energy tend to decrease, while the dipole and quadrupole moments tend to increase. The only high correlation that survives if the tricoordinated carbons are added is that between the kinetic energy and the population. This correlation is illustrated in Figure S1. As mentioned above in connection with Table 2, here again this correlation has little predictive power because the spread amounts to 105 kJ/mol. Furthermore the high correlation is valid for a large population span, whereas the *local* fine structure around a population of six electrons for example shows a very poor local correlation. Similar observations have been made before in the context of bond critical point properties versus bond length for this data set.[25]

Finally we comment in some detail on the relation of our work to that of Breneman et al.[13,42] Their results are established from 6000 integrated atoms, appearing in an undisclosed set of molecules, calculated at the HF/6-31+G* level. The 110 sample molecules on which their TAE method was tested have very few molecules in common with our data set. Whereas their integration procedure was less accurate, we used the same clustering technique. However Breneman et al. introduced 18 atomic properties, including surface properties, and a different multipole representation. They discovered a total of 36 carbon atom types, starting from an a priori division based on hybridization, a bias that we avoided. It is not straightforward to compare our atom types with theirs, partially because we did not cover groups such as NO_2, Cl, C≡N, CHO, or C≡C. There are identical atom types such as the tertiary carbon (cluster 4), which corresponds to the "S1405" atom type in their notation. Also, quite remarkably, the amidic and carboxylic carbons appear together in their atom type "S1310", reminiscent of our cluster 19. On the other hand we distinguish atom types that they do not distinguish. For example, their methylene ("S1402") is divided over three clusters in our work, i.e., methylene bonded to C or S (cluster 3), or bonded to O (cluster 11) or to N (cluster 10).

Conclusion

Cluster analysis operating on 213 unique quantum topological carbon atoms, each represented by seven properties, exhibits a tree structure determining computable ways to define carbon atom types. Carbon, appearing in 20 amino acids and smaller derived molecules, is a malleable element, molded by a variety of chemical environments into a wealth of atom types. Based on two different cluster-separability criteria, our systematic analysis reveals a coarse atom type representation of 5 clusters, and a finer one containing 21 clusters. The immediate coordination of the central carbon plays a major role in labeling the atom types.

Appendix 1: Cluster Analysis

Cluster analysis[47−49] is widely used in fields such as sociology, zoology, linguistics, anthropology, and others. It is appropriate to classify a large number of objects into classes based on some measure of similarity. In our case the objects are topological atoms, the classes atom types and the similarity a resemblance in atomic properties. Cluster analysis is able to visualize associations between variables in a tree structure or *dendrogram* (e.g. Figure 1). A horizontal line intersecting a dendrogram marks a fixed level of similarity. The number of intersections between this horizontal line and the vertical lines in the dendrogram indicates the number of clusters appearing at a given level of similarity. According to Cormack's division[47] we applied *agglomerative hierarchical* cluster analysis, which assigns a set of entities into a group by a series of successive fusions.

First a similarity or distance matrix is constructed on the basis of Euclidean distance. After standardization to a mean of 0 and a standard deviation of 1 for each atomic number, the distance between two atoms A and B is defined as $d_{AB} = [\sum_{k=1}^{7}(P_k(A) - P_k(B))^2]^{1/2}$, where $P_k(A)$ is a property of atom A. Then fusion takes place between individuals or groups that are most similar. This rule does not involve a threshold in distance; it simply means that entities are fused one by one as the similarity (or distance) parametrically increases, until all entities are fused. This fusion happens automatically in the program Clustan-Graphics.

There are several ways of measuring the Euclidean distance between an individual and a group or between two groups. We used the *average linkage method*, which defines the distance or proximity between two groups as the average of the distances between all pairs of individuals, one individual from each cluster.[50] Hence all the objects within a cluster contribute to the intercluster similarity. In other words, each object is, on average, more similar to any other member in the same cluster than to any other member in another cluster. This method has the merit that the distribution of individuals within two clusters influences their proximity.

Appendix 2: Separability of Clusters

Table S1a (Supporting Information) shows the range of values of atomic properties (in atomic units) for two clusters appearing at the three-cluster level in the carbon dendrogram (Figure 1). It is clear that the range of dipole moments of clusters 1 and 3 overlap considerably, whereas the populations are well-separated. Each range can be characterized by its mean and a standard deviation. This is justified because large populations of continuous data are generally considered to be distributed according to the normal distribution curve,[49] represented by a normalized Gaussian function centered at the mean value μ and with a width determined by the standard deviation σ. In a normal distribution any data point found outside the 3σ-interval from the mean is considered to be an outlier. How can we use this cutoff criterion to ensure that two clusters are well-separated? For each atomic property we calculate the mean and standard deviation of all atoms in a given cluster. Given two clusters A and B, we then calculate the difference of the means ($\Delta\mu_{AB} = \mu_A - \mu_B$), the sum of the standard deviations ($\Sigma\sigma_{AB} = \sigma_A + \sigma_B$) and their *intercluster ratio* ($\Delta\mu_{AB}/\Sigma\sigma_{AB}$), again for *each* atom property. If this ratio is larger than 3 *for at least one atomic property*, or $\Delta\mu_{AB}/\Sigma\sigma_{AB} > 3$, we judge the two clusters A and B to be *separable*. This means that according to the $\Delta\mu/\Sigma\sigma > 3$ criterion 99.7% of the populations of both clusters are free from the possibility of being misclassified. If the criterion is relaxed to $\Delta\mu/\Sigma\sigma > 2$ (95.5% misclassification chance), the clusters are allowed to overlap to a larger extent. This procedure is illustrated in Table S1b. As expected, the population has a high intercluster ratio because it was well-separated. On the other hand the volume, dipole, and hexadecupole moment have a low intercluster ratio and hence do not contribute to the separation of the two clusters. In any event, cluster 1 and cluster 3 are separable according to both criteria, $\Delta\mu/\Sigma\sigma > 3$ and $\Delta\mu/\Sigma\sigma > 2$. A more complete analysis involving cluster 2 as well demonstrates that all cluster pairs (1 and 2, 1 and 3, and 2 and 3) are well-separated.

The determination of a single and definite number of clusters or atom types is elusive. However one can propose an "optimal" number of clusters in terms of chemical interpretation. As one moves down a dendrogram (e.g. Figure 1), the number of clusters increases and the information they contain becomes more specific and detailed. The drawback is that the clusters start to overlap more; i.e., they become harder to distinguish as separate entities. Each criterion (i.e. $\Delta\mu/\Sigma\sigma > 3$ or $\Delta\mu/\Sigma\sigma > 2$) gives rise to a *representation*, which contains a number of atom types depending on the dendrogram to which the criterion is applied. *The condition for a representation to be valid is that each possible pair of clusters is separable at a preset value of the intercluster ratio.* In other words, if at least one pair of clusters is not separable at a given intercluster ratio, then the representation is not valid. This procedure is used throughout this paper. On one hand we are driven toward discovering as many atom types as possible in order to preserve as much chemical information as possible. On the other hand, when taken to the extreme, this drive leads to overlapping and hence nonsensical atom types. This breakdown is prevented by the stricter separation criterion (demanding that $\Delta\mu/\Sigma\sigma > 3$) or by the more relaxed criterion (that $\Delta\mu/\Sigma\sigma > 2$).

A final point is related to the quality of atomic integration. Previous work[14] has proposed an error bar (or "intrinsic error") for atomic properties computed for the current set of molecules. When the variance within a cluster is smaller than the intrinsic error estimated for an atom, the clusters are narrower than they can possibly be and misclassification is likely. For instances where the standard deviation of a cluster falls below the value

(47) Everitt, B. S. *Cluster Analysis*, 3rd ed.; Edward Arnold: London, 1993.
(48) Anderberg, M. R. *Cluster Analysis for Applications*; Academic Press: New York, 1973.
(49) Livingstone, L. *Data Analysis for Chemists*, 1st ed.; Oxford University Press: Oxford, Great Britain, 1995.
(50) Wishart, D. *ClustanGraphics Primer*: *A Guide to Cluster Analysis*; Clustan Ltd.: Edinburgh, Great Britain, 1999.

of the intrinsic error (as tabulated in Table 8 of ref 14), the intrinsic error is reported instead. If an atom type contains only one atom, the standard deviation of the cluster does not vanish but equals the intrinsic error of the atom.

Supporting Information Available: Table giving atomic properties, intercluster values, mean and standard deviations of atomic properties, and the correlation matrix for carbon atom types and a figure showing the kinetic energy vs population for the carbon atom types. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0284198