Atomic Properties of Selected Biomolecules: Quantum Topological Atom Types of Hydrogen, Oxygen, Nitrogen and Sulfur Occurring in Natural Amino Acids and Their Derivatives

Paul L. A. Popelier* and Fiòna M. Aicken[a]

Abstract: Molecular electron densities are generated at B3LYP/6-311+G(2d,p)//HF/6-31G(d) level for 57 molecules, including one conformation of each naturally occurring amino acid and smaller derived molecules. The electron densities are partitioned into atomic fragments according to the approach of quantum chemical topology (QCT). A set of 547 unique topological atoms is obtained, containing 421 hydrogens, 63 oxygens, 57 nitrogens and 6 sulfurs. Each atom is described by seven properties: volume, kinetic energy, monopole, dipole, quadrupole, octupole and hexadecapole moment. Cluster analysis groups atoms into atom types based on their similarity expressed in the discrete 7D space of atomic properties. Using a separation criterion we distinguish seven hydrogen, six oxygen, two nitrogen and six sulfur atom types.

Keywords: amino acids · electron density · density functional calculations · quantum chemical topology

Introduction

Quantum chemical topology (QCT)[1–3] is a modern approach that seeks to recover chemical insight from ab initio wave functions. Based on quantum mechanics[4–6] this approach provides a parsimonious procedure to generate a wealth of shapes and properties of atoms as they appear *inside* molecules. QCT uses the molecular electron density to define an atom in a molecule as a sharply bounded three-dimensional subspace. These (quantum) topological atoms do not overlap and collectively exhaust full space. This approach proposes an attractive route for one of the challenges of modern theoretical chemistry: to identify an atom in its chemical environment. The fact that atoms preserve their characteristics under similar chemical surroundings enables chemistry to be a science of rational classification rather than a compilation of disparate facts.

This paper and its companion paper[7] concentrate on the atom types occurring in amino acids. We partition one conformation of each naturally occurring amino acid into unique atoms and then apply cluster analysis to accumulate similar atoms into atom types. This paper focuses on hydrogen, nitrogen, oxygen and sulfur, while the companion paper focuses on carbon only, the richest element in terms of atom-type variety. Cluster analysis detects the similarity between atoms in a discrete space of atomic properties that have been painstakingly obtained by volume integration over topological basins.

The issue of atomic similarity is closely related to that of atomic transferability, although the latter is not explicitly investigated in this paper. Transferability arises when a unique atom is replaced by a similar one or by an average atom representing a subset of atoms, usually referred to as an atom type. In this paper we will rigorously establish atom types purely by means of intrinsic similarity. However, the assessment of an atom type's transferability is application-dependent and would require a separate study.

Several groups have employed QCT to study amino acids[8, 9] or examined the transferability of alkyl chains in aldehydes and ketones,[10] of methyl and methylene fragments in alkyl monoethers[11] and investigated approximate transferability to alkanols[12] and alkanenitriles.[13] The concept of compensatory transferability was recently introduced[14] 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 related to that of the group of Breneman who proposed[15] the so-called "transferable atom equivalent" (TAE) method some time ago, but focuses only on the definition of atom types without modification of atomic properties by adjustment of atomic surfaces.

Again in the context of transferability lines of attack other than QCT were taken, for example in the context of the electrostatic potential of polypeptides,[16] point charge models for amino acid side chains[17] and electrostatic interactions of peptides and amides[18] or in connection with a molecular electron density "lego" approach to molecule building.[19]

[a] Dr. P. L. A. Popelier, F. M. Aicken
Department of Chemistry, UMIST
Manchester, M60 1QD (Great Britain)
Fax: (+44) 161-200-4559
E-mail: pla@umist.ac.uk

  0947-6539/03/0905-1207 $ 20.00+.50/0

Thanks to a thorough understanding of an in-house topological integration procedure[20, 21] used to obtain atomic properties[22] we are able to report a detailed cluster analysis on 421(H)+63(O)+57(N)+6(S) = 547 atoms, drawn from a set of twenty amino acids, supplemented by smaller derived molecules. The companion paper reports on the set of 213 carbon atoms, completing our study of 760 atoms in total.

Theoretical Background

Quantum chemical topology: Quantum chemical topology (QCT) is an approach to extract chemical insight from modern ab initio wave functions. The core of this methodology, called the theory of "atoms in molecules",[2, 3] was pioneered by the Bader group. A brief historical survey of its development is given in a recent literature survey,[23] while a more recent survey[24] witnessed its increasing action radius and popularity. At the heart of QCT is the notion of topological basins, which is reviewed below. This notion also constitutes the hub of the topological study[25] of the electron localisation function (ELF),[26] another strand of QCT.

In view of comprehensive introductions[3, 27] we only review salient points of QCT. The extraction of chemical knowledge in QCT occurs from three-dimensional property densities, such as the electron density ρ, its Laplacian, ELF or kinetic energy densities. Herein we only focus on ρ. Instead of invoking a reference electron density we use the molecular electron density as its own reference. This is accomplished by the gradient of ρ, which is in fact an internal difference. A gradient path is a sequence of infinitesimally short gradient vectors traced in real space, each re-evaluated at the endpoint of the previous one. A gradient path moves in the direction of steepest ascent of the property density (ρ for our purpose) until it reaches an attractor, which typically coincides with a nucleus. The infinite number of gradient paths attracted to a nucleus constitutes a topological basin. According to QCT this basin is identified with a (topological) atom inside a molecule.

An atomic property is obtained as an integral of a property density over the volume of a topological atom. For example, atomic population is defined as the integral of ρ over the atomic volume. Atomic multipole moments are defined within the compact spherical tensor formalism,[28] which yields only three, five, seven and nine components of the dipole, quadrupole, octupole and hexadecupole moment, respectively. These moments[29, 30] 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^{-1}. A topological intermolecular potential, based on these multipole moments, predicts the geometries of van der Waals complexes,[31] a multitude of natural DNA base pairs[32] as well as water clusters and the hydration of amino acids.[33] We proceed with orientationally invariant magnitudes of the multipole moments because a comparison between their components would require keeping track of their orientation and a convention for maximum alignment. In summary each atom is represented by seven (scalar) atomic properties: volume, population, dipole, quadrupole, octupole and hexadecupole moment, and kinetic energy. The latter was obtained by integration of a kinetic energy density, while the atomic volume was obtained by capping the atoms by the $\rho = 0.001$ au contour.[34]

Cluster analysis: An appropriate technique to classify the large number of atoms into atom types is cluster analysis.[35–37] This method visualises associations between variables in a tree structure or *dendrogram*. Figure 1 shows the dendrogram of all hydrogen atoms occurring in our data set (details in "generation Dataset"). At the very bottom of the dendrogram individual atoms appear. As one moves up the diagram more and more atoms become linked: they fuse into larger and larger clusters as their similarity (expressed by the distance measure described in next section) decreases.

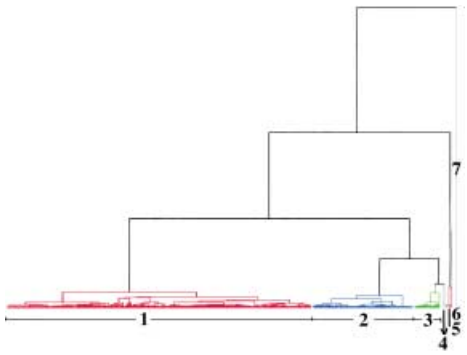

Figure 1. Dendrogram generated by the cluster analysis on hydrogen defining seven atom types.

A given level of similarity is marked by an imaginary horizontal line intersecting the dendrogram. The number of clusters appearing at a given level of similarity is determined by the number of intersections between this horizontal line and the vertical lines in the dendrogram. Cormack's division[35] specifies the type of cluster analysis we applied here as *agglomerative hierarchical* because it assigns a set of entities into a group by a series of successive fusions. First a similarity or distance matrix is constructed, based on the Euclidean distance. Subsequently individuals or groups that are most similar are fused. There are several ways of measuring the Euclidean distance between an individual and a group, or between two groups. This work applies the *average linkage* method, which defines the similarity distance between two groups as the average of the distances between all pairs of individuals, one individual from each cluster.[38] As a result all the objects within a cluster contribute to the inter-cluster similarity. Put differently, each object is, on average, more similar to any other member in the same cluster than to any other member in another cluster. An advantage of this method is that the distribution of individuals within two clusters influences their proximity.

Determination of atom types: Cluster analysis does not supply a criterion deciding the number of clusters the data set should be divided in. Instead it only presents possible ways in which a dataset can be partitioned into clusters by means of a

  0947-6539/03/0905-1208 $ 20.00+.50/0 *Chem. Eur. J.* **2003**, *9*, No. 5

dendrogram. It is achievable, however, to invoke a criterion, external to cluster analysis, that fixes the number of clusters and thereby provides a *representation* of the data set in terms of atom types. This criterion is purely statistical and ensures that each atom type is sufficiently separated from another.

It is convenient to explain the issue of cluster separability and hence atom type separability with an example. Table 1a shows the range of values for each atomic property for the three clusters appearing at the three-cluster level in the hydrogen dendrogram (Figure 1). Each range can be characterised by its mean and a standard deviation, which is justified since large populations of continuous data are generally distributed according to the normal distribution curve,[37] which is represented by a normalised Gaussian function centred 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 outlier criterion to ensure that two clusters are well separated? For each atomic property the mean and standard deviation of all atoms in a given cluster is calculated. 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 ($\Sigma\sigma_{AB} = \sigma_A + \sigma_B$). Table 1b illustrates these values for all three possible cluster pairs. Next the inter-cluster ratio ($\Delta\mu_{AB}/\Sigma\sigma_{AB}$) is calculated, again for each atomic property and between all possible cluster pairs, as shown in Table 1c. If this ratio is larger than three 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. Another interpretation of this separability of clusters (in this case atom types) A and B is that an atom belonging to A can never *also* belong to B because statistically A and B are so remote that this atom is an outlier to B. Loosely speaking one can say that clusters A and B do not overlap to a degree of three standard deviations. If the criterion is relaxed to the inter-cluster ratio $\Delta\mu/\Sigma\sigma > 2$ (95.5% misclassification chance) the clusters are allowed to overlap to a larger extent.

From Table 1c it is clear that all three cluster-pairs are well separated at $\Delta\mu/\Sigma\sigma > 3$ level since there is always at least one atomic property for which $\Delta\mu/\Sigma\sigma > 3$. Cluster 1 and 2 are only separable because of their widely differing dipole moments. On the other hand clusters 2 and 3 are well separated by all atomic properties except the population. All cluster pairs are also separable at the $\Delta\mu/\Sigma\sigma > 2$ level.

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. Moving down a dendrogram (for example Figure 1) increases the number of clusters and the information they contain becomes more specific and detailed. The disadvantage is that the clusters start to overlap more, that is 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 inter-cluster ratio. In other words, if at least one pair of clusters is not separable at a given inter-cluster ratio then the representation is not valid. This procedure is used throughout this paper. For example, the three-cluster representation for hydrogen is valid because all cluster-pairs are well-separated as explained above.

Given a particular dendrogram we are driven towards discovering as many atom types as possible in order to preserve as much chemical information as possible. However, we have to be careful since, when taken to the extreme, this drive to many atom types leads to overlapping and hence nonsensical (ill-defined) atom types. Of course there is also

Table 1a. The means and standard deviations of atomic properties (in au) for all clusters appearing at the three-cluster level in the hydrogen dendrogram (Figure 1).

	Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
cluster	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	43.1	9.6	0.5717	0.0769	0.8620	0.1911	0.1659	0.0126	0.1536	0.0697	0.2627	0.0637	0.249	0.092
2	51.5	2.2	0.6179	0.0075	1.0436	0.0164	0.0670	0.0053	0.2582	0.0033	0.3016	0.0234	0.180	0.111
3	60.5	0.3	0.5752	0.0001	1.0000	0.0002	0.1140	0.0000	0.3577	0.0002	0.0949	0.0003	1.000	0.002

Table 1b. The inter-cluster values $\Delta\mu$ and $\Sigma\sigma$, and their ratio for each atomic property (in a. u.) between each pair of clusters appearing at the three-cluster level in the hydrogen dendrogram (Figure 1).

	Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
clusters	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$	$\Delta\mu$	$\Sigma\sigma$
1,2	8.4	11.8	0.0462	0.0844	0.1817	0.2075	0.0989	0.0180	0.1046	0.0729	0.0389	0.0870	0.069	0.204
1,3	17.4	9.9	0.0035	0.0770	0.1380	0.1913	0.0519	0.0126	0.2041	0.0699	0.1678	0.0640	0.751	0.094
2,3	9.0	2.5	0.0427	0.0076	0.0436	0.0166	0.0471	0.0053	0.0995	0.0035	0.2067	0.0237	0.820	0.113

Table 1c. The inter-cluster ratio $\Delta\mu/\Sigma\sigma$ for each atomic property (in au) between each pair of clusters appearing at the three-cluster level in the hydrogen dendrogram (Figure 1).

$\Delta\mu/\Sigma\sigma$	Volume	Kinetic energy	Population	Dipole	Quadrupole	Octupole	Hexadecupole
1,2	0.7	0.5	0.9	5.5	1.4	0.4	0.3
1,3	1.8	0.0	0.7	4.1	2.9	2.6	7.9
2,3	3.5	5.6	2.6	8.8	28.7	8.7	7.3

  0947-6539/03/0905-1209 $ 20.00+.50/0

the absurd and useless limit in which each individual atom would constitute its own atom type. The danger of overlap and concomitant breakdown of a representation is prevented by the stricter separation criterion, which demands that $\Delta\mu/\Sigma\sigma > 3$. The more relaxed criterion, $\Delta\mu/\Sigma\sigma > 2$, is used for qualitative interpretation.

A final point concerns the quality or reliability of atomic integration. Working with the same data set previous work[22] estimated an error bar (or "intrinsic error") for each atomic property. If the variance within a cluster is smaller than the intrinsic error estimated for an atom then the clusters are narrower than they can possibly be and misclassification is likely. In cases where the standard deviation of a cluster drops below the value of the intrinsic error (as tabulated in Table 8 of ref. [22]) 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.

Programs and Computational Methods

The program MOLDEN[39] provided the Z matrices for the program GAUSSIAN94,[40] which generated all required wave functions at B3LYP/6-311+G(2d,p) level[41, 42] after optimisation at HF/6-31G(d)[43] level. This choice proved to be a good compromise[44] between accuracy and computational cost.[45, 46] Moreover it was found[9] that at the HF/6-31+G(d) level the experimental values of the geometric parameters for the side-chains of the 20 amino acids are reproduced with an acceptable degree of accuracy. Each molecule of a given family was optimised towards a geometry close to the optimised geometry of another member of the family in order to maximise conformational proximity amongst molecules of the same family. The program MORPHY98[47] carried out all atomic integrations,[21, 48] some of which were repeated with the intention of improving their accuracy. Extensive tables of atomic properties for all atoms are given in Appendix 2 of ref. [46]. The program ClustanGraphics[49] and in an earlier stage the program SPSS[50] performed the hierarchical agglomerative cluster analysis. After separate standardisation for each atomic number the distance between two atoms A and B is Euclidean and is defined as:

$$d_{AB} = \sqrt{\sum_{k=1}^{7} (P_k(A) - P_k(B))^2} \quad (1)$$

where $P_k(A)$ is a property of atom A. It is preferred to use the Euclidean squared distance measure to calculate the similarity matrix for the purpose of clustering large datasets (containing more than about 200 cases).

Results and Discussion

Dataset generation: We produced a set of 57 molecules that includes the twenty most common naturally occurring free amino acids and smaller derived molecules. How the latter set was constructed is best explained by means of an example, as shown in Figure 2. Aspartic acid, H_2N-HC_α(-$C_\beta H_2 C_\gamma$(=O)-OH)-COOH, is cleaved at the C_α–C_β bond and the side chain fragment is capped with a hydrogen atom. Focusing on the side chain, the molecule thus obtained is acetic acid, H-$C_\beta H_2 C_\gamma$(=O)-OH. Subsequently the C_β–C_γ bond is cleaved creating two fragments of which the larger one was again capped with a hydrogen atom. This leads to formic acid, H-C_γ(=O)-OH. The final *single* bond to be cleaved and capped is C_γ-O resulting in water as the last molecule derived from aspartic acid, also designated by the shorthand "D".

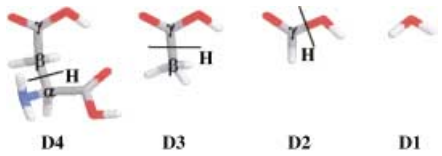

Figure 2. Illustration of the "cleaving and capping" procedure used to generate the data set of amino acids and derived molecules. Aspartic acid (D4) generates the three other members of the D family: acetic acid (D3), formic acid (D2) and water (D1). Explanation in main text.

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. Hence aspartic acid is denoted by D4, acetic acid by D3, formic acid by D2 and water by D1. The molecule H_2 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. Such ambiguities do not influence the outcome of the work described below. The only important matter is that each molecule has a unique name and that we have a sufficiently large set of molecules (with internal similarities) to draw topological atoms from. Note that double bonds and ring structures were left intact, and that duplicated molecules were discarded. Earlier work[45] characterised the bonds occurring in the set of 57 molecules by their so-called bond critical point properties in the context of molecular similarity.

Classification of hydrogen atom types: Hydrogen is the most abundant atom in the data set with 421 individual atoms. The $\Delta\mu/\Sigma\sigma > 3$ separation criterion fails at the 8-cluster representation, leaving the 7-cluster representation as the most detailed but still well-separated and valid representation. The dendrogram for this model is illustrated in Figure 1. The membership of the clusters can be charted as follows:

1 bonded to C
2 bonded to N
3 bonded to O
4 bonded to N and hydrogen-bonded to N or O
5 bonded to S
6 bonded to S and hydrogen-bonded to O
7 bonded to H

Tracing the dendrogram from top to bottom we notice that the first cluster to separate itself from the other clusters is that of hydrogen bonded to H (cluster 7). Next to split off are the hydrogens bonded to S (clusters 5 and 6), followed by the hydrogens bonded to C (cluster 1). Subsequently the hydrogens bonded to N but *not* involved in a hydrogen bond split off (cluster 2). Next the hydrogens bonded to N and involved in a hydrogen bond (cluster 4) split off leaving behind the hydrogens bonded to O (cluster 3). At some point the hydrogens bonded to S (cluster 5) split in two classes: those involved in a hydrogen bond and those that are not. Remarkably the hydrogens bonded to N and involved in a hydrogen bond (cluster 4) split off from the ones hydrogen-bonded to O (cluster 3) rather than to N (cluster 2). This observation seems

0947-6539/03/0905-1210 $ 20.00+.50/0 *Chem. Eur. J.* **2003**, *9*, No. 5

to suggest that the hydrogen bond distorts hydrogens bonded to N so much that they start to resemble hydrogens bonded to an oxygen. One of the previously proposed topological hydrogen-bond criteria[51] supports this statement. Indeed hydrogen bonds are known to reduce the volume of the hydrogen-bonded hydrogen and the oxygen-bonded hydrogens have on average the smallest volume of all hydrogens.

At the 8-cluster level cluster 1 (hydrogens bonded to C) splits in two subclusters that cannot be differentiated by either separation criterion, $\Delta\mu/\Sigma\sigma > 2$ or $\Delta\mu/\Sigma\sigma > 3$. The inter-cluster ratio $\Delta\mu/\Sigma\sigma$ for these two subclusters is far below 2 for all atomic properties, ranging from 0.2 (for both the volume and the population) to merely 1.3 for the hexadecupole moment. Therefore the 7-cluster representation is the only one we discuss. Note that for carbon, whose cluster structure was reported elsewhere,[7] there are two valid representations, one with five atom types (for $\Delta\mu/\Sigma\sigma > 3$) and one with 21 atom types (for $\Delta\mu/\Sigma\sigma > 2$).

Clusters 1 (bonded to C) and 5 (bonded to S) are the least distinguishable types of hydrogen. Only the dipole moment allows a distinction to be drawn between the two clusters ($\Delta\mu/\Sigma\sigma = 8$). This implies that hydrogen atoms perceive these two environments, carbon and sulfur, in a very similar way. It is tempting to relate this observation to the recognised[52, 53] bioisosterism of sulfur and the methylene group.

Table 2 lists the average values and standard deviations for all atomic properties of each cluster. The largest hydrogens are those bonded to C (except for the one bonded to H) with a volume of 49.1 au, while the smallest hydrogens are bonded to O, with a volume of 21.7 au, less than half the maximum value. We observe the lowest hydrogen population when bonded to O, the lowest but one when bonded to N, then C, followed by H and S. This order is a mirror image of the ranking of Pauling's electronegativity values. When inverted the populations obey the order **S** < H < C < N < O, while the electronegativity scale yields H < C < **S** < N < O. The only mismatch is the position of S. Constructing an alternative electronegativity scale based on QCT populations is enticing, especially since charge transfer is a phenomenon more directly related to the textbook's way of interpreting Pauling's definition as "the ability of an atom in a molecule to attract shared electrons to itself".[54] His and other scales are based on energy differences, and are hence a less direct measure for charge transfer than population patterns. Moreover electronegativity scales are mostly used to predict charge transfer; they may as well be based on it.

The properties of hydrogens involved in hydrogen bonds, such as in clusters 4 and 6, can be compared with the properties of their non-H-bonded analogues in clusters 2 and 5, respectively. This comparison reveals a decrease in the volume, population and the dipole moment of the hydrogen-type upon formation of a hydrogen bond (for example, going from cluster 2 to 4, and from 5 to 6). The effect on the properties observed is in keeping with the previously proposed hydrogen bond criteria.[51]

The correlation between hydrogen's atomic properties are given in Table 3, which lists Pearson correlation coefficients.[55] Again the highest correlation is between the kinetic energy and the population ($r = 0.98$). After removing cluster 7 (bonded to H) this correlation increases to $r = 0.99$. High correlations are also observed for the volume and the quadrupole moment ($r = 0.96$) and the volume and the population ($r = 0.95$). The ordering exhibited by the population, is reflected in the magnitudes of the quadrupole moments in line with a correlation coefficient of $r = 0.91$.

Table 3. Correlation matrix for the hydrogen atom types defined at the 7-cluster representation.

	Volume	Kinetic Energy	Population	Dipole	Quadrupole	Octupole	Hexadecupole
volume	1						
kinetic energy	0.90	1					
population	0.95	0.98	1				
dipole	−0.61	−0.62	−0.70	1			
quadrupole	0.96	0.82	0.91	−0.72	1		
octupole	0.20	0.59	0.46	−0.29	0.08	1	
hexadecupole	0.50	0.14	0.26	−0.08	0.58	−0.66	1

Classification of oxygen atom types: According to both the strict and relaxed criteria ($\Delta\mu/\Sigma\sigma > 3$ and $\Delta\mu/\Sigma\sigma > 2$) the 7-cluster representation is not valid. Hence we describe the 63 oxygen atoms of the data set in terms of a 6-cluster representation. Figure 3 shows the dendrogram of the cluster analysis on oxygen. The clusters can be characterised as follows (the oxygen in question marked in bold):

1) hydroxyl oxygen in the carboxyl group of amino acids: R-$C_{\beta \text{ or } \gamma}$ (=O)**O**H.
2) phenol oxygen (Ar-**O**H) or hydroxyl oxygen in R-C(=O)**O**H (where R is specified below).
3) alcohol oxygen bonded to alkyl group: R-**O**H.

Table 2. Mean and standard deviations of the atomic properties of hydrogen in the 7-cluster representation.

		Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
Cluster		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	H-C	49.1	1.7	0.6182	0.0089	0.9816	0.0207	0.160	0.006	0.198	0.019	0.290	0.041	0.22	0.09
2	H-N	32.4	1.4	0.4972	0.0128	0.6508	0.0254	0.186	0.008	0.061	0.010	0.235	0.025	0.34	0.06
3	H-O	21.7	1.3	0.3798	0.0112	0.4212	0.0157	0.159	0.006	0.026	0.005	0.106	0.018	0.24	0.03
4	N/O ··· H-N	24.3	0.5	0.4548	0.0010	0.5664	0.0054	0.153	0.004	0.086	0.011	0.225	0.027	0.22	0.04
5	H-S	52.7	0.4	0.6185	0.0090	1.0495	0.0140	0.070	0.001	0.258	0.004	0.312	0.013	0.12	0.01
6	O ··· H-S	48.2	0.3	0.6159	0.0001	1.0260	0.0002	0.059	0.000	0.259	0.000	0.270	0.000	0.35	0.00
7	H-H	60.5	0.3	0.5752	0.0001	1.0000	0.0002	0.114	0.000	0.358	0.000	0.095	0.000	1.00	0.00

  0947-6539/03/0905-1211 $ 20.00+.50/0

4) keto oxygen in the carboxyl group: R-C(=**O**)OH.
5) amide oxygen: R-C(=**O**)NH_2.
6) oxygen in water: H_2**O**.

We imagine a horizontal line intersecting three vertical "tree" lines in the dendrogram of Figure 3. At this level of similarity we perceive three "superclusters", the left one being

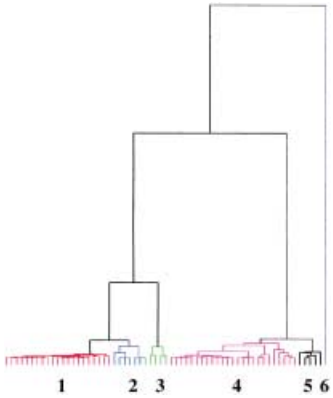

Figure 3. Dendrogram generated by the cluster analysis on oxygen defining six atom types.

formed by a fusion of clusters 1, 2 and 3. This supercluster contains the hydroxyl oxygens, or alternatively, in terms of bonded partners, the oxygens bonded to carbon and hydrogen. The middle supercluster is formed by a coagulation of clusters 4 and 5, and hence encompasses the keto oxygens (doubly bonded to carbon, and appearing in the carboxy and the amide group). An equivalent tag for this supercluster, again in terms of immediate bonding partners, is the set of all oxygens bonded to one carbon. The supercluster on the right is highly dissimilar to the other two ("independent early on") and contains just the oxygen in water, which is obviously bonded to two hydrogens.

Figure 4 shows a chart of only 49 molecules rather than the full set of 57. This is because we focus on O, N and S and the remaining $57 - 49 = 8$ molecules consist only of H and C. Figure 4 assigns the membership of O, N and S to their respective clusters, as shown in the corresponding dendrograms (Figures 3, 5 and 6).

Cluster 1 consists of twenty-one atoms: the hydroxyl oxygen of the C_βOOH group in each of the twenty amino acids and the hydroxyl oxygen of C_γOOH in aspartic acid (D4). Cluster 2 contains seven members: three phenol oxygens (in Y1, Y2 and Y3) and four hydroxyl oxygens, three appearing in COOH and bonded to a pure hydrocarbon chain (in D2, D3 and E1), and one appearing in glutamic acid (E2) and bonded to C_γ. The latter occurrence is not surprising because the $NH_2C_\alpha HC_\beta$OOH group is too far away from C_δOOH to influence the oxygen in C_δOOH markedly. This confirms that methylene groups act as a buffer, and hence the hydroxyl oxygen in C_δOOH "perceives" the rest of the amino acid as a pure hydrocarbon chain. Cluster 3 encompasses four hydroxyl oxygens, two in the alcohol group of the serine family (S1 and S2) and two in the alcohol group of the threonine family (T1 and T2). Cluster 4 is the largest class with 25 members, all keto oxygens, one in each amino acid's C_βOOH group, and five extra atoms in COOH group occurring in D2, D3, D4, E1 and E2. This cluster contains *all* keto oxygens in COOH, since there are twenty-five COOH groups in the data set. Consequently the keto oxygen atom type is less sensitive to its environment than the hydroxyl oxygen atom type spread out over clusters 1 and 2, presumably because it is bonded to only one atom instead of two.

Table 4 lists the average values and standard deviations of the atomic properties of each oxygen cluster of the 6-cluster representation. The volume monotonically increases from clusters 1 to 6. As a result keto oxygens (clusters 4 and 5) are larger than hydroxyl oxygens (clusters 1, 2 and 3). The keto oxygens have a slightly larger population than the hydroxyl ones and have the highest kinetic energy. The dipole moments of the keto oxygens are roughly twice as high as of any other oxygen type. The dipole moment of the keto oxygen and of the carbon to which it is bonded both oppose the dipole moment that arises from the charge transfer term.[56]

Table 5 provides the correlation coefficients between the different clusters defined within the 6-cluster representation. The highest positive correlation is observed between the dipole moment and the population ($r = 0.97$), and the highest negative correlation between dipole and quadrupole moments ($r = -0.96$). This matrix is completely different to the one calculated for hydrogen (Table 3).

Classification of nitrogen atom types: There are 57 unique nitrogen atoms in the data set. The three-cluster representation is not valid by both separation criteria, $\Delta\mu/\Sigma\sigma > 2$ and $\Delta\mu/\Sigma\sigma > 3$. Hence we end up with a two-cluster representation, the dendrogram of which is shown in Figure 5. This atom type representation is only valid by the weaker separation criterion ($\Delta\mu/\Sigma\sigma > 2$). Cluster 2 cannot be split any further because the newly formed cluster (cluster 3, utmost right) overlaps (by both separation criteria) in every atomic property with cluster 1, which renders the 3-cluster representation invalid. The fact that the cluster 3 overlaps with cluster 1 rather than cluster 2 is unexpected. Normally a representation is declared invalid when two sub-clusters generated from the *same* cluster cannot be distinguished, in this case clusters 2 and 3. In this sense nitrogen behaves in a unique way compared to the other atoms.

Figure 4 shows the occurrence of the atom types of cluster 1 and 2 in the data set of molecules. Cluster 1 consists of thirty-five nitrogens in total. They are all tri-coordinated, the bonding partner always being a hydrogen or a carbon. Since there is no occurrence of a nitrogen bonded to three carbons, there are three possible types of tri-coordinated nitrogens left. The nitrogen can be bonded to three hydrogens, which occurs in ammonia (K1). The second possibility is that the nitrogen is bonded to two hydrogens and one carbon (CNH_2). This group, the primary amines, is the largest group containing the "backbone"(i.e., bonded to C_α) amine group of each of the twenty amino acids, five extra amine groups of the lysine family (K2, K3, K4, K5 and K6) and five amine groups in the arginine family (R2, R3, R4, R5 and R6). The third and final possibility is the secondary amines (CC′NH), four of which appear in the arginine family (R3, R4, R5 and R6).

  0947-6539/03/0905-1212 $ 20.00+.50/0 *Chem. Eur. J.* **2003**, *9*, No. 5

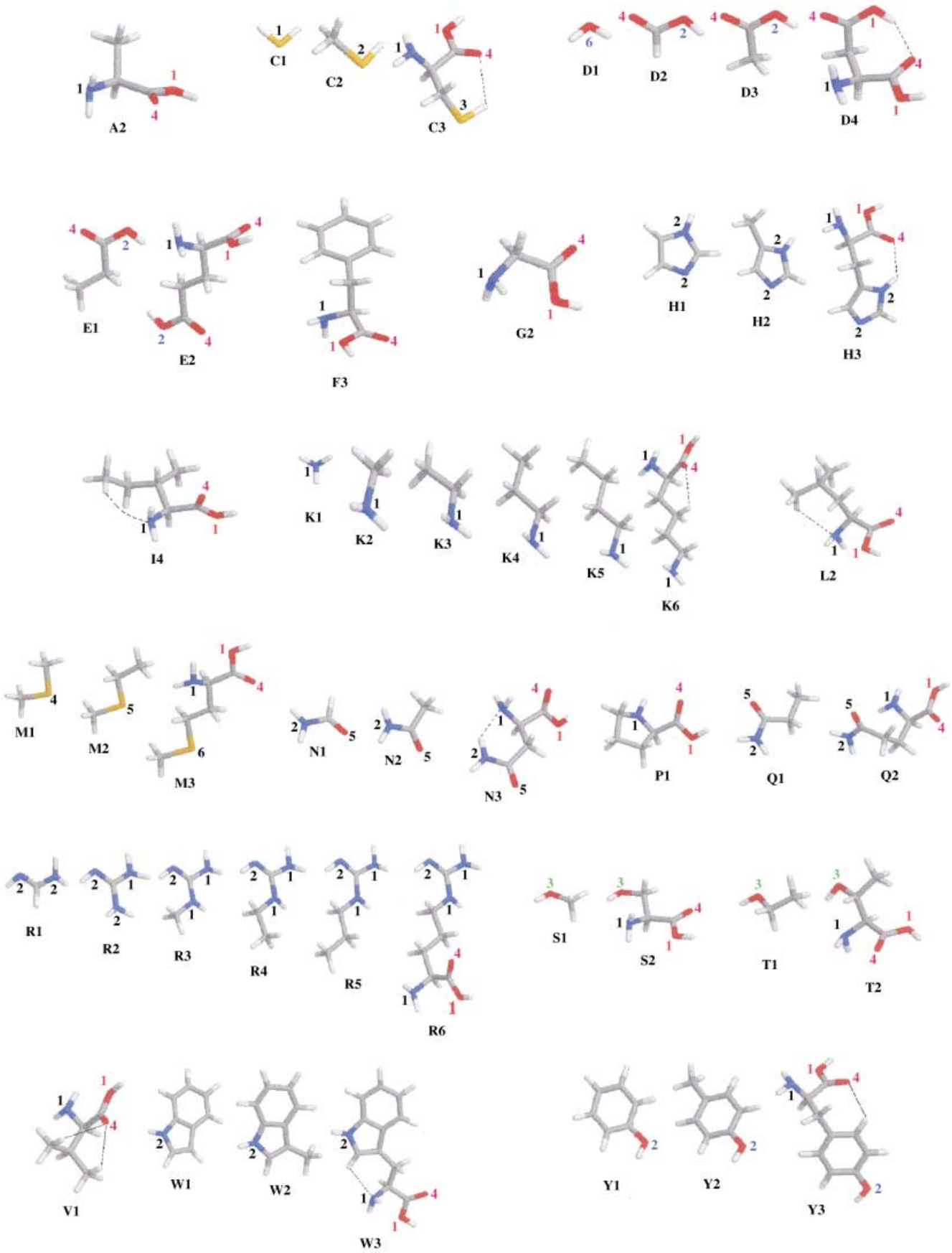

Figure 4. Assignment of the membership of O, N and S to their respective clusters. The labels of the molecules refer to their family designation, which is marked by the standard letter classification of amino acids. The numerical labels of the atoms refer to the respective dendrograms of O (Figure 3), N (Figure 5) and S (Figure 6). Intramolecular hydrogen bonds are marked by a dashed line.

  0947-6539/03/0905-1213 $ 20.00+.50/0

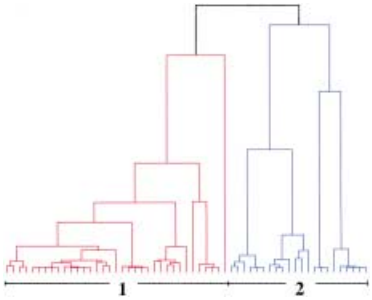

Figure 5. Dendrogram generated by the cluster analysis on nitrogen defining two atom types.

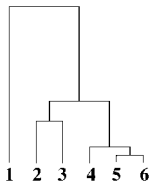

Figure 6. Dendrogram generated by the cluster analysis on sulfur defining six atom types.

Clearly the numbers add up to thirty-five since $1(NH_3)+(20+5+5)(CNH_2)+4(CC'NH)=35$.

Cluster 2 consists of twenty-two nitrogens in total, of which nine are bi-coordinated and thirteen tri-coordinated. There are three possible bonding patterns to C and H for bi-coordinated nitrogens: NH_2, CNH and CNC′. The NH_2 situation does not occur in the dataset. Six nitrogens are of the CNH pattern (in R1, R2, R3, R4, R5 and R6) and three of pattern CNC′ (in H1, H2 and H3). In the tri-coordinated subset we encounter six nitrogens of the CC′NH pattern (in N1, N2, N3, W1, W2 and W3) and seven of the CNH_2 pattern (all five amide nitrogens in N1, N2, N3, Q1, Q2 and two amine groups in R1 and R2). The latter two amino groups could be considered as "pseudo-amidic" if a H-N= group is seen as equivalent to O=. This makes sense because organic chemistry textbooks recognise the imine group as the nitrogen analogue of the carbonyl group. In summary, the total number of twenty-two "cluster 2" nitrogens is recovered since $6(CNH)+3(CNC')+6(CC'NH)+7(CNH_2)=22$.

An important question is whether a simple chemical description or label can characterise each cluster. Hybridisation cannot be introduced as a uniform tag to discriminate the clusters because cluster 2 is a mixture of sp^2 and sp^3 nitrogens. However, all nine aromatic ring nitrogens belong to cluster 2 and cluster 1 consists solely of sp^3 nitrogens. Since QCT is a largely orbital-free approach we explored this question again, now beyond topological coordination (i.e., numbers of topologically bonded atoms). In the spirit of investigating the immediate environment of an atom type we screened the conformation of $48=57-9$ tri-coordinated nitrogens. An appropriate average out-of-plane angle measured nitrogen's environment, a low value indicating near-planarity. This dihedral angle, which involves the central nitrogen and its three bonded neighbours, is 38° for perfect tetrahedrality (e.g. in K1). We find that all nitrogens in cluster 1 yield an out-of-plane angle between 22 and 38°, where the primary amines (CNH_2) tend to have larger angles than the secondary amines (CC′NH). The out-of-plane angles of the thirteen tri-coordinated nitrogens of cluster 2 range from 0 to 24°, and cluster into three groups: seven planar (∼0°), three mildly distorted (∼10°) and three nearly tetrahedral (∼20°). The out-of-plane angle is successful in separating the tri-coordinated nitrogens of cluster 2 [0–24°] and cluster 1 [22–38°], given the poor overlap between the out-of-plane ranges. This finding confirms that the geometry of the immediate chemical environment dominates the properties of a given atom, and hence determines the atom type, based on the wave function. In summary, cluster 1 is the set of nitrogen with a nearly tetrahedral tri-coordinated environment, and cluster 2 is the set of bi-coordinated nitrogens or largely planar tri-coordinated nitrogens. Table 6 furnishes the average atomic properties and the standard deviations for the two clusters. The standard deviations for cluster 2 are consistently higher than those for cluster 1

Table 4. Mean and standard deviations of the atomic properties of oxygen in the 6-cluster representation. The symbol # represents the number of oxygen atoms in each cluster, totalling 63.

		Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
cluster	#	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	21	119.2	0.7	75.6348	0.0049	9.1238	0.0026	0.200	0.011	0.658	0.019	0.57	0.04	2.14	0.12
2	7	121.6	0.9	75.5821	0.0294	9.1017	0.0044	0.196	0.029	0.705	0.012	0.54	0.08	2.41	0.12
3	4	123.4	2.1	75.4888	0.0207	9.0882	0.0121	0.128	0.007	0.783	0.035	2.91	0.12	0.72	0.04
4	25	134.6	2.6	75.6919	0.0034	9.1734	0.0131	0.421	0.011	0.405	0.026	1.24	0.04	0.93	0.30
5	5	139.1	1.5	75.6487	0.0078	9.1734	0.0077	0.366	0.012	0.474	0.030	1.17	0.02	0.67	0.22
6	1	149.5	0.1	75.4082	0.0001	9.1029	0.0001	0.162	0.001	0.862	0.000	1.20	0.00	3.72	0.01

Table 5. Correlation matrix for the oxygen atom types defined at the 6-cluster representation.

	Volume	Kinetic energy	Population	Dipole	Quadrupole	Octupole	Hexadecupole
volume	1						
kinetic energy	−0.34	1					
population	0.30	0.78	1				
dipole	0.28	0.79	0.97	1			
quadrupole	0.00	−0.92	−0.94	−0.96	1		
octupole	0.01	−0.41	−0.29	−0.27	0.24	1	
hexadecupole	0.30	−0.59	−0.50	−0.51	0.66	−0.47	1

0947-6539/03/0905-1214 $ 20.00+.50/0 *Chem. Eur. J.* **2003**, *9*, No. 5

Table 6. Mean and standard deviations of the atomic properties of nitrogen in the 2-cluster representation.

	Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
cluster	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	111.0	8.8	54.9293	0.0544	7.9592	0.0283	0.285	0.056	1.371	0.125	1.97	0.29	4.90	0.61
2	119.7	16.9	55.0896	0.0721	8.0894	0.0354	0.142	0.111	1.335	0.198	1.40	0.39	3.06	0.85

Table 7. Mean and standard deviations of the atomic properties of sulfur in the 6-cluster representation.

	Volume		Kinetic energy		Population		Dipole		Quadrupole		Octupole		Hexadecupole	
cluster	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	223.2	0.2	397.6327	0.0001	15.9171	0.0004	0.718	0.001	3.4283	0.0008	2.788	0.008	5.81	0.04
2	212.1	0.2	397.6509	0.0001	15.9529	0.0004	0.792	0.001	3.3924	0.0008	2.100	0.008	5.38	0.04
3	210.0	0.2	397.6731	0.0001	15.9731	0.0004	0.778	0.001	3.2720	0.0008	2.237	0.008	5.94	0.04
4	200.1	0.2	397.6694	0.0001	15.9741	0.0004	0.855	0.001	3.3151	0.0008	1.796	0.008	3.71	0.04
5	200.1	0.2	397.6889	0.0001	15.9946	0.0004	0.833	0.001	3.3441	0.0008	1.626	0.008	4.60	0.04
6	198.4	0.2	397.6849	0.0001	15.9781	0.0004	0.824	0.001	3.3361	0.0008	1.514	0.008	4.43	0.04

signifying that cluster 1 contains a more homogeneous collection of atoms, as expected from the previous discussion. The dipole moment of nitrogens in cluster 1 is roughly double that of the nitrogens in cluster 2 although the standard deviations are quite high, as expected since cluster 1 consists of tetrahedral tri-coordinated nitrogens and cluster 2 of nearly planar ones. The only property that really discriminates between cluster 1 and 2 is the population since this is the only property for which $\Delta\mu/\Sigma\sigma$ just meets the cut-off value of 2.0.

A final comment about the difficulty in classifying nitrogens is necessary. We acknowledge that the bonding environment of nitrogens in our data set only includes C and H, which have very similar electronegativities. On the other hand carbon, the element with the richest set of atom types comes across C, N, S, O and H as bonding partners in our data set. In order to obtain a richer picture for nitrogen we should include functional groups such azines, oximes, hydrazones, azides, indoles, aziridines, anilines, nitroso compounds and nitro groups. However, since they do not feature in natural peptides they have not been studied here. It would be interesting, however, to extend this type of cluster analysis, perhaps even into the realm of rather esoteric inorganic moieties.

Classification of sulfur atom types: Sulfur occurs only in the cysteine family (C1, C2 and C3) and in the methionine family (M1, M2 and M3), resulting in a total of six unique atoms. Following the usual procedure of monitoring inter-cluster ratios for all possible pairs of clusters reveals that all representations are valid, by both separation criteria $\Delta\mu/\Sigma\sigma > 3$ and $\Delta\mu/\Sigma\sigma > 2$. This means that there are never any overlap problems and that the simple dendrogram shown in Figure 6 can be meaningfully discussed at the "6-cluster" representation. In other words, each atom is its own type. The labels in Figure 6 correspond to the ones in Figure 4, which shows the position of atoms in the molecules of the data set. Comparison of these figures allows us to follow the progression of the cluster analysis. The first atoms to fuse in the dendrogram are 5 and 6, which occur in methylethyl thioether (M2) and methionine (M3), respectively. The sulfur most similar to this pair is not unexpectedly 4, occurring in dimethylthioether (M1). We anticipated the two thiol sulfurs (2 in methanethiol, C2 and 3 in cysteine, C3) to fuse. This cluster then joins with [4,5,6] since the sulfur (1) in H_2S (C1) is obviously most distinct.

Table 7 yields the mean and standard deviations of the atomic properties of sulfur in the 6-cluster representation. Note that the standard deviations coincide with the intrinsic integration error as explained above in the section on the "Determination of atom types", thus preventing them to vanish, which would lead to infinite inter-cluster ratios.

The correlation coefficients are charted in Table 8. The most significant correlation, after that between volume and octupole, is that between the kinetic energy and the population ($r = 0.97$). A least-squares regression between these two properties shows that even this high correlation is too crude to usefully predict the kinetic energy of sulfur from its population, since the prediction error is of the order of 30 kJ mol^{-1}.

Table 8. Correlation matrix for the sulfur atom types defined at the 6-cluster representation.

	Volume	Kinetic energy	Population	Dipole	Quadrupole	Octupole	Hexadecupole
volume	1						
kinetic energy	– 0.91	1					
population	– 0.92	0.97	1				
dipole	– 0.95	0.78	0.85	1			
quadrupole	0.62	– 0.73	– 0.74	– 0.55	1		
octupole	0.98	– 0.89	– 0.89	– 0.92	0.49	1	
hexadecupole	0.82	– 0.54	– 0.57	– 0.88	0.25	0.77	1

Conclusion

The combination of cluster analysis and quantum chemical topology enables the computation of atom types from modern ab initio wave functions. A large number of topological atoms were obtained by partitioning the electron densities of all

  0947-6539/03/0905-1215 $ 20.00+.50/0

natural amino acids and smaller derived molecules. Each atom in the total set of 547 unique atoms (421 hydrogens, 63 oxygens, 57 nitrogens and 6 sulfurs) is described by seven properties: volume, kinetic energy, monopole, dipole, quadrupole, octupole and hexadecapole moment. A statistical separation criterion defines seven hydrogen atom types, six oxygen atom types, two nitrogen atom types and six sulfur atom types. The coordination and immediate environment of the central atom is paramount in the design of atom type labels. Nitrogen is the most difficult to categorise, most likely due to the small variations in its chemical environment in amino acids. An electronegativity scale based on charge transfer is suggested.

[1] R. F. W. Bader, *Encyclopedia of Computational Chemistry*, *Vol. 1* (Ed.: P. von R. Schleyer), Wiley, Chichester, **1998**, p. 64.
[2] R. F. W. Bader, *Atoms in Molecules. A Quantum Theory*, Oxford University Press, Oxford (UK), **1990**.
[3] P. L. A. Popelier, *Atoms in Molecules. An Introduction*, Pearson Education, London (UK), **2000**.
[4] R. F. W. Bader, T. T. Nguyen-Dang, *Adv. Quantum Chem.* **1981**, *14*, 63 – 124.
[5] R. F. W. Bader, *Pure Appl. Chem.* **1988**, *60*, 145 – 155.
[6] R. F. W. Bader, P. L. A. Popelier, *Int. J. Quantum Chem.* **1993**, *45*, 189 – 207.
[7] P. L. A. Popelier, F. M. Aicken, *J. Am. Chem. Soc.* **2003**, *125*, in press.
[8] C. F. Matta, R. F. W. Bader, *Proteins*: *Struct. Funct. Genet.* **2000**, *40*, 310 – 329.
[9] C. F. Matta, R. F. W. Bader, *Proteins*: *Struct. Funct. Genet.* **2002**, *48*, 519 – 538.
[10] A. M. Graña, R. A. Mosquera, *J. Chem. Phys.* **2000**, *113*, 1492 – 1500.
[11] A. Vila, R. A. Mosquera, *J. Chem. Phys.* **2001**, *115*, 1264 – 1273.
[12] M. Mandado, A. M. Graña, R. A. Mosquera, *J. Mol. Struct. Theochem.* **2002**, *584*, 221 – 234.
[13] J. L. Lopez, M. Mandado, A. M. Graña, R. A. Mosquera, *Int. J. Quantum Chem.* **2002**, *86*, 190 – 198.
[14] R. F. W. Bader, D. Bayles, *J. Phys. Chem. A* **2000**, *104*, 5579 – 5589.
[15] C. M. Breneman, T. R. Thompson, M. Rhem, M. Dung, *Comput. Chem.* **1995**, *19*, 161 – 179.
[16] S. L. Price, A. J. Stone, *J. Chem. Soc. Faraday Trans.* **1992**, *88*, 1755 – 1763.
[17] C. Chipot, J. G. Angyan, B. Maigret, H. A. Scheraga, *J. Phys. Chem.* **1993**, *97*, 9788 – 9796.
[18] C. H. Faerman, S. L. Price, *J. Am. Chem. Soc.* **1990**, *112*, 4915 – 4926.
[19] P. D. Walker, P. G. Mezey, *J. Am. Chem. Soc.* **1994**, *116*, 12022—12032.
[20] P. L. A. Popelier, *Mol. Phys.* **1996**, *87*, 1169 – 1187.
[21] P. L. A. Popelier, *Comput. Phys. Commun.* **1998**, *108*, 180 – 190.
[22] F. M. Aicken, P. L. A. Popelier, *Can. J. Chem.* **2000**, *78*, 415 – 426.
[23] P. L. A. Popelier, F. M. Aicken, S. E. O'Brien, in *"Chemical Modelling*: *Applications and Theory"*, *Vol. 1* (Ed.: A. Hinchliffe), Royal Society of Chemistry Specialist, Periodical Report, Chapter 3, pp. 143 – 198, **2000**.
[24] P. L. A. Popelier, P. J. Smith, in *"Chemical Modelling*: *Applications and Theory"*, *Vol. 2* (Ed.: A. Hinchliffe), Royal Society of Chemistry Specialist Periodical Report, Chapter 8, pp. 391 – 448, **2002**.
[25] B. Silvi, A. Savin, *Nature* **1994**, *371*, 683 – 686.
[26] A. D. Becke, K. E. Edgecombe, *J. Chem. Phys.* **1990**, *92*, 5397 – 5403.
[27] R. J. Gillespie, P. L. A. Popelier, *Chemical Bonding and Molecular Geometry from Lewis to Electron Densities*, Oxford University Press, New York, USA, **2001**.
[28] R. N. Zare, *Angular Momentum*, Wiley Interscience, **1988**.
[29] D. S. Kosov, P. L. A. Popelier, *J. Chem. Phys.* **2000**, *113*, 3969 – 3974.
[30] D. S. Kosov, P. L. A. Popelier, *J. Phys. Chem. A* **2000**, *104*, 7339 – 7345.
[31] P. L. A. Popelier, L. Joubert, D. S. Kosov, *J. Phys. Chem. A* **2001**, *105*, 8254 – 8261.
[32] L. Joubert, P. L. A. Popelier, *Phys. Chem. Chem. Phys.* **2002**, *4*, 4353 – 4359.
[33] P. L. A. Popelier, M. Devereux, unpublished results.
[34] R. F. W. Bader, M. T. Carroll, J. R. Cheeseman, C. Chang, *J. Am. Chem. Soc.* **1987**, *109*, 7968 – 7979.
[35] B. S. Everitt, *Cluster Analysis*, 3rd ed., Edward Arnold, London (GB), **1993**.
[36] M. R. Anderberg, Cluster *Analysis for Applications*, Academic Press, New York (USA), **1973**.
[37] L. Livingstone, *Data Analysis for Chemists*, 1st ed., Oxford University Press, Oxford (UK), **1995**.
[38] D. Wishart, *ClustanGraphics Primer "A Guide to Cluster Analysis"*, Clustan, Edinburgh (UK), **1999**.
[39] G. Schaftenaar, J. H. Noordik, *J.Comput. Aided Mol. Design* **2000**, *14*, 123 – 134.
[40] GAUSSIAN94, 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**.
[41] C. Lee, W. Yang, R. G. Parr, *Phys. Rev.* **1988**, *B37*, 785 – 789.
[42] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648 – 5652.
[43] P. C. Hariharan, J. A. Pople, *Theor. Chim. Acta* **1973**, *28*, 213.
[44] J. B. Foresman, A. Frisch, *Exploring Chemistry with Electronic Structure Methods*, 2nd ed., Gaussian, Pittsburgh, USA, **1996**.
[45] S. E. O'Brien, P. L. A. Popelier, *Can. J. Chem.* **1999**, *77*, 28 – 36.
[46] F. M. Aicken, UMIST (Manchester), **2000**.
[47] MORPHY98, a program written by P. L. A. Popelier with a contribution from R. G. A. Bone, UMIST, Manchester (UK), **1998**, see http://morphy.ch.umist.ac.uk/1998.
[48] P. L. A. Popelier, *Mol. Phys.* **1996**, *87*, 1169 – 1187.
[49] D. Wishart, Edinburgh, Scotland (UK), **1999**.
[50] SPSS Inc., version 10.0.7, Chicago (USA), **2000**, http://www.spss.com.
[51] U. Koch, P. L. A. Popelier, *J. Phys. Chem.* **1995**, *99*, 9747 – 9754.
[52] A. Korolkovas, *Essentials of Molecular Pharmacology*: *Background for Drug Design*, Wiley, New York, **1970**.
[53] A. Burger, *Medicinal Chemistry*, 3rd ed., Wiley, New York, **1970**.
[54] L. R. Murphy, T. L. Meek, A. L. Allred, L. C. Allen, *J. Phys. Chem. A* **2000**, *104*, 5867 – 5871, claim that this is a misreading of Pauling's definition, as if electronegativity is an intrinsic, in situ, molecular property.
[55] This coefficient is defined as

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

where x and 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 value of r lies within [0,1] M. R. Spiegel, *Theory and Problems of Statistics*, McGraw-Hill, New York (USA), **1972**.
[56] T. Slee, A. Larouche, R. F. W. Bader, *J. Phys. Chem.* **1988**, *92*, 6219 – 6227.

Received: August 30, 2002 [F4380]

  0947-6539/03/0905-1216 $ 20.00+.50/0