Electronic Energy and Multipolar Moments Characterize Amino Acid Side Chains into Chemically Related Groups

Hugo J. Bohórquez,* Mateo Obregón, Constanza Cárdenas, Eugenio Llanos, Carlos Suárez, José Luis Villaveces, and Manuel Elkin Patarroyo

Fundación Instituto de Inmunología de Colombia-FIDIC, Carrera 50 No. 26-00, Bogotá, Colombia

Received: March 12, 2003; In Final Form: August 4, 2003

We explored amino acid side chains from a quantum mechanical perspective in order to identify molecular similarities and differences, for the purpose of exploring the de novo design of peptidic sequences with desired biochemical reactivities. Charge densities for the 20 genetically encoded amino acids in both α and β conformations were partitioned into molecular fragments, and their electronic properties (charge, energy, and dipole and quadrupole moments) were calculated using atoms in molecules theory. Transferability, as required by this theory, was confirmed for the side chains. Two methods were used to identify similarity: The first mapped each side chain property vector onto frequency differentiated Andrews plots, while the second used a composite measure of vectorial distance and angle as the dependent variable for hierarchical cluster analyses. We found that both methods clustered the side chains into chemically related groups only on the basis of theoretically derived variables. Both fine grained and coarse grained levels of analysis highlighted important emergent properties such as hydropathy, polarity, size, aromaticity, and the presence of carbon chains (aliphatics) and hydroxyl groups (alcohol). These results verify the hypothesis that symmetries of charge densities (as measured by the variables used here) can account for the observed chemical reactivities of amino acids.

1. Introduction

The theoretical study of chemical similarity between amino acids is of major interest due to its importance in protein chemistry, giving rise to replacement rules that are becoming more widely used in drug research.[1−5] These rules have been developed on the assumption that physical and chemical variables group the 20 genetically encoded amino acids into related functional classes. The variables used most commonly are volume, hydrophilicity, surface area, polarity, and charge, among the many that can be defined or measured. Most classifications unsuccessfully group residues according to side chain reactivity. In such works, molecular similarity is determined by different mathematical methods, giving rise to many alternatives for determining classification. Furthermore, these works omit properties which quantum mechanics holds as being fundamental for characterizing molecular interactions. We propose a different approach to amino acid classification by using a new similarity criterion based entirely on electronic density analysis. We hope to overcome some of the above-mentioned problems in determining amino acid similarity, as well as identifying a generalizable method for comparing molecular fragments.

Systematic comparison from a theoretical point of view of different molecules for assessing their specific biological activities is a major task in molecular design. This problem is known as the molecular similarity question. The quantum molecular similarity issue was first formally addressed by Carbó et al.,[6] who suggested overlapping the charge densities of two molecules as a measurement of similarity. During the past two decades, researchers have paid a lot of attention to this definition,[7] remarking on its importance for drug design. Other approaches based on similar ideas use electrostatic potential, electron density *shape* comparisons,[8] or atomic multipole-based similarities.[9] A common feature regarding these proposals is the vagueness concerning how they must be used for comparing submolecular regions. A different approach that is based on the theory of atoms in molecules (AIM) was recently proposed, suggesting a similarity vector space composed of a given molecule's *bond critical points* (BCPs) as obtained from charge density analysis.[10,11] Popelier et al. proposed that the measurement of molecular similarity should be the Euclidean distance in a 3-D space defined by the density, the Laplacian of the charge density, and the ellipticity, (ρ, $\nabla^2\rho$, and ϵ, respectively), calculated where the bond critical points occur. Such a representation has the disadvantage of being basis-set dependent, impeding comparison with data derived from other theoretical calculations.

We propose that electronic polarizations (including monopolar, dipolar, and quadrupolar symmetries) and energy be the molecular representation of choice for comparative studies, mainly because these magnitudes are governed by physical laws, that is confer predictability, since these magnitudes are described by physical axioms, facilitating their calculation, accuracy, and interpretation. Such properties have also been shown to be basis-set independent,[12,13] and they can also be obtained by experimental means. In practice, multipole analysis also has the advantage of facilitating a comparison scheme by only necessitating a few values for each item rather than requiring the contrasting of entire charge densities.[14] Another important factor in our proposal is that the computed values for these properties are both cheaper and virtually always obtainable for any arbitrary molecule (Only in few cases do the theoretical calculations fail to converge, but this is merely a circumstantial computational

* To whom correspondence should be addressed. Telephone: +57-1-3158919 ext 122. Fax: +57-1-3158919 ext 108. E-mail: hugo_bohorquez@fidic.org.co.

10.1021/jp034639b CCC: $25.00
Published on Web 11/04/2003

problem and not a theoretical impasse.), as opposed to experimental means, which are error-prone and costly.

Keeping such considerations in mind, we started our theoretical study of molecular similarity by considering the following hypotheses: (a) biochemical groups can be suitably described by a reduced set of variables obtainable from the ab initio determined electron density $\rho(\mathbf{r})$, and (b) differences between these magnitudes are a measurement of both coarse and fine grained molecular dissimilarity. We show results that validate the above premises applied to the 20 genetically encoded amino acids.

2. Theoretical Definition of Atomic and Group Properties

It is known from experiments that atoms and functional groups of atoms can be transferable to a high degree. When atom properties are evaluated, one can determine the atomic or group contribution to the total properties for a system. Therefore, any transferability probe implies the evaluation of submolecular properties. Building quantum chemical representations of molecules by combining molecular fragments has been the subject of many studies,[15−18] but only AIM theory leads to an unambiguous definition of transferable chemical groups.[19] This theory is firmly rooted in quantum mechanics and yields atomic properties in a rigorous way.[20]

AIM theory also reveals a remarkable stability of atomic properties with respect to basis-set dependency,[21,22] making this theory a suitable tool for rational biomolecular design and comparison. Previous studies suggest that amino acid residues can be treated as transferable electrostatic building blocks that match each other for assembling polypeptide chains.[23−26] It has been shown recently that AIM yields sufficiently accurate electrostatic moments which reproduce the ab initio electrostatic potential.[27] Since our study depends on the definition of atomic properties, a concise explanation of the theory of atoms in molecules follows.

In the theory of atoms in molecules, the partitioning of 3-D space into atomic regions is based on differential geometry analysis of electron density $\rho(\mathbf{r})$. The 3-D space is divided into nonoverlapping volumes Ω_i, where i identifies each atom in the molecule. In general, these atomic regions are bounded by infinity and by the interatomic surfaces which obey the zero-flux condition,

$$\nabla\rho(\mathbf{r})\cdot\mathbf{n}(\mathbf{r}) = 0 \tag{1}$$

for all points r, where the vector $\mathbf{n}(\mathbf{r})$ is the unit vector normal to the surface at $\mathbf{r}$. Thus, each atom has an exclusive portion of space where its physical properties can be evaluated, which is called the *atomic basin*.

2.1. Atomic Properties and Group Additivity. An atomic property corresponding to the ith atom is defined as being the volume integral of the property density, $G(\mathbf{r})$, over the atomic region Ω_i:

$$G(\Omega_i) = \int_{\Omega_i} G(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{2}$$

One of the most important realizations of AIM, when compared to other partitioning schemes, is that this theory was constructed for preserving the additive characteristics of physical properties.[28] That is, the molecular expected value of the property G is

$$\langle G\rangle = \sum_{i}^{N} G(\Omega_i) \tag{3}$$

where N is the total number of participating atoms in the 3-D volume Ω for property $G(\mathbf{r})$.

By means of this rule for addition, the values of different atom sets (or fragments) that build up a molecule can be calculated. Thus, it is feasible to evaluate physical magnitudes corresponding to a functional group.

In the present work we focus on the evaluation of the charge density moments and the electronic energy for each side chain. The atomic population, $N(\Omega_i)$, can be considered as the zeroth charge density moment,

$$N(\Omega_i) = -e\int_{\Omega_i}\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{4}$$

and the atom's net charge is

$$q(\Omega_i) = Z_i e + N(\Omega_i) \tag{5}$$

where Z_i is the atomic number of the ith atom.

The energy of an atom in a molecule, $E_e(\Omega)$, is purely electronic in origin and is defined as

$$E_e(\Omega_i) = -T(\Omega_i) \tag{6}$$

where the electronic kinetic energy $T(\Omega)$ is defined in the atomic statement of the virial theorem.[19]

The first moment of the charge density, $M(\Omega_i)$, provides a measure of the extent and direction of the atom's charge density dipolar polarization by determining the displacement of the atom's center of negative charge from the position of its nucleus:

$$\mathbf{M}(\Omega_i) = -e\int_{\Omega_i}\mathbf{r}_{\Omega_i}\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{7}$$

The second moment, $Q(\Omega_i)$, gives information on planar distributions of charge, which is particularly prevalent in aromatic groups. This is the quadrupolar polarization of an atomic density. When it is measured with respect to the z-axis, its expression is given by

$$Q_{zz}(\Omega_i) = -e\int_{\Omega_i}(3z_{\Omega_i}^2 - r_{\Omega_i}^2)\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{8}$$

3. Definition of Molecular Similarity

Our hypothesis claims that any two molecular fragments can be compared with physical variables determined from the electron density $\rho(\mathbf{r})$. If the compared molecules exhibit similar values in every transferable property, they can be considered as interchangeable. The limit of perfect transferability for atoms or for groups of atoms can never be attained.[29] This implies that any similarity calculation must be defined as a comparative scale for the molecular set under study. Clearly, closer values for a given property will be interpreted as a greater level of transferability for each fragment *in terms of the specific property*. This means that, while fragment A can be nearly identical in property G to fragment B, there is the possibility that they may be dissimilar in another property G'. Therefore, fragments A and B must be similar in all properties of interest for them to be considered overall similar. Note that properties G and G' must share an underlying nature and yet be mostly nonoverlapping in the property space for them to be considered simultaneously in any similarity analysis. Accordingly, in the current study, all our discriminant variables are directly derived from the electronic properties of the molecules in question.

Since changes in the properties of an atom are a direct response to changes in the atom's charge distribution, we can concentrate on electronic variables for our evaluation of

similarity. Polarizability results from the multipole operators evaluated over the charge distribution; they quantify the implicit symmetry of the charge distribution, and every multipolar moment measures a different symmetry. Thus, we can use polarizability as a valid scenario for molecular transferability probes, since it summarizes the molecule's symmetry (*shape*). Multipole moments are directional variables; that is, they depend on the chosen coordinate system. Consequently, the following approach is valid only if the molecular systems under study are aligned. In what follows, we affirmed that all the molecular systems were aligned in such way that the common atoms bonded to the fragments are similarly located in a common coordinate system. This means that the origin is located at the C_α, with the carboxilic atom and the first atom of the side chain of each residue held in the same respective directions for all amino acids.

With the dipolar polarization vector $\mathbf{M} = (M_x, M_y, M_z)$ and the quadrupole tensor

$$\mathbf{Q} = \begin{bmatrix} Q_{xx} & Q_{xy} & Q_{xz} \\ Q_{xy} & Q_{yy} & Q_{yz} \\ Q_{xz} & Q_{yz} & Q_{zz} \end{bmatrix}$$

we define the vector representing each amino acid as a point in the (energy, charge, dipolar moment, quadrupolar moment) property space:

$$\mathbf{V} = (E, q, M_x, M_y, M_z, Q_{xx}, Q_{xy}, Q_{xz}, Q_{yy}, Q_{yz}, Q_{zz}) \quad (9)$$

Since these variables have different magnitudes and units, it is necessary to choose a normalization method. We performed the following transformation to each variable for normalizing the set of vectors $\{\mathbf{V}\}$:

$$\hat{E} = \frac{|E| - \min\{|E|\}}{\Delta E}$$

with $\Delta E = \max\{|E|\} - \min\{|E|\}$.

$$\hat{q} = \frac{q - \min\{q\}}{\Delta q}$$

with $\Delta q = \max\{q\} - \min\{q\}$.

Each component of the higher multipoles (**M** and **Q**) cannot be treated as an independent variable as in the case of the energy E or the charge q, demanding a more careful transformation:

$$\hat{\mathbf{M}} = (\mathbf{M} - \mathbf{M}_0)\frac{1}{\Delta M}$$

with $\mathbf{M}_0 = (\min\{M_x\}, \min\{M_y\}, \min\{M_z\})$ and $\Delta M = \max\{M_x, M_y, M_z\} - \min\{M_x, M_y, M_z\}$.

$$\hat{\mathbf{Q}} = (\mathbf{Q} - \mathbf{Q}_0)\frac{1}{\Delta Q}$$

where $\mathbf{Q}_0 = (\min\{Q_{xx}\}, \min\{Q_{xy}\}, \min\{Q_{xz}\}, \min\{Q_{yy}\}, \min\{Q_{yz}\}, \min\{Q_{zz}\})$ and $\Delta Q = \max\{Q_{xx}, Q_{xy}, Q_{xz}, Q_{yy}, Q_{yz}, Q_{zz}\} - \min\{Q_{xx}, Q_{xy}, Q_{xz}, Q_{yy}, Q_{yz}, Q_{zz}\}$.

$$\hat{\mathbf{V}} = (\hat{q}, \hat{E}, \hat{M}_x, \hat{M}_y, \hat{M}_z, \hat{Q}_{xx}, \hat{Q}_{xy}, \hat{Q}_{xz}, \hat{Q}_{yy}, \hat{Q}_{yz}, \hat{Q}_{zz}) \quad (10)$$

In this way, each component is defined in [0, 1]. Notice that this procedure ensures that no single component of the higher multipoles (**M** or **Q**) can dominate the direction of $\hat{\mathbf{V}}$, because just one of the components has the maximum value of 1. The largest value of the Euclidean norm $|\hat{\mathbf{V}}|_{\max}$ is less than $\sqrt{3}$ for our set of parameters.

TABLE 1: Side Chain Conformation

	χ_1 (deg)	residue
gauche+	−66.7	Arg Asn Asp Gln His Ile Leu Lys Met Phe Trp Tyr
gauche−	64.1	Thr Ser Cys
trans	183.6	Val

4. Amino Acid Model

The amino acid group can be represented by $|NHC_\alpha H(R)C(=O)|$, where R indicates the side chain. To mimic the protein environment, the $-NH|$ term was blocked with a formyl group and the $-(C=O)|$ term was blocked with an amino group. Thus, the amino acid model studied here is $HC(=O)|NHC_\alpha H(R)C(=O)|NH_2$.

Certainly, to imagine a fixed geometry for a molecule is a coarse approximation at best. We chose nuclear coordinates found in crystallographic observations of protein structure so as to reflect chemical structures found in nature.[30−32]

In our comparative study, we have taken advantage of the fact that the main chains of structured polypeptides exhibit statistical preferences for two conformations, namely, β strand and α helix. As in our previous work,[33] we chose amino acid geometries from those most frequently found in reported databases.[32] Thus, our average geometries are constructed with mean values for bond distances and angles extracted from these datasets. We used these average values for the geometries from standard databases, without performing geometry optimizations upon them, since our aim is to compare aligned side chains as they are found in proteins, not as they could be found in the free state.

The values of the main chain torsion angles ψ and ϕ were taken as $-39°$ and $-65°$, respectively, for α conformations, and as $120°$ and $-130°$, respectively, for β strands. The heterocyclic pyrrolidine ring in proline restricts its backbone conformations, and thus the torsion angles ψ and ϕ for proline were taken as $-39°$ and $-70°$, respectively, for α and as $120°$ and $-70°$, respectively, for β strands. This gives rise to a total of 40 molecules for our study.

The side chain conformation of each amino acid was selected according to the highest frequency reported by X-ray crystallography per conformer,[30,31] as shown in Table 1.

5. Results

The integrals for atomic properties were calculated over each atomic basin as described in section 2. The corresponding calculations were performed by using the AIMPAC suite of routines with the PROMEGA integration method.[34] This method is suitable for complicated systems whose wave functions represent a challenge due to their intricate expression, as amino acids indeed have. Calculating the AIM properties for the full set of atoms in our amino acid model was done in around 1000 h of CPU time on a SGI Power Challenge four-processor machine at our institute (approximately 1 h per atom). The computation error, $L(\Omega)$, was reported as being less than 10^{-4}, which is acceptable for these types of calculations. The molecular wave functions were evaluated using GAUSSIAN94[35] at the HF 6-31G level with polarization functions on those heavy atoms capable of forming hydrogen bonds. Wave functions for all the molecules discussed in this article at the RHF/6-31 G level in Gaussian94[35] format, as well as other material pertinent to this research, can be found at our Web site: http://www.fidic.org.co/biomathematics/.

TABLE 2: Side Chain Properties[a]

	q	M_x	M_y	M_z	Q_{xx}	Q_{xy}	Q_{xz}	Q_{yy}	Q_{yz}	Q_{zz}	E	$\|\mathbf{M}\|$	$\|\mathbf{Q}\|$
Alaα	0.050	0.126	0.213	0.085	−0.009	−0.109	−0.006	−0.189	−0.046	0.198	−39.688	0.262	0.262
Alaβ	0.019	0.118	0.196	0.065	0.167	−0.132	−0.118	−0.244	−0.085	0.077	−39.685	0.238	0.337
Argα	0.058	−0.060	0.048	−0.097	−6.450	5.271	1.996	0.137	−2.280	6.313	−320.808	0.124	10.179
Argβ	0.030	−0.062	0.021	−0.115	−6.340	5.323	1.924	0.175	−2.287	6.165	−320.672	0.132	10.092
Asnα	0.045	−0.337	−0.369	−1.199	−2.881	1.719	−0.170	1.921	0.310	0.961	−207.413	1.299	3.566
Asnβ	0.019	−0.209	−0.328	−1.202	−2.593	2.284	−0.061	2.053	0.313	0.540	−207.337	1.263	3.818
Aspα	0.019	−0.134	0.069	−1.021	−1.519	2.000	0.914	0.986	−0.161	0.533	−227.219	1.032	2.976
Aspβ	−0.007	−0.081	0.132	−1.009	−1.406	2.195	0.921	1.162	0.008	0.244	−227.120	1.021	3.133
Cysα	−0.022	−0.678	0.350	−0.431	−1.859	1.542	3.672	5.547	−0.172	−3.688	−437.530	0.876	7.286
Cysβ	−0.059	−0.689	0.320	−0.492	−1.647	1.444	3.436	5.549	−0.489	−3.902	−437.459	0.906	7.165
Glnα	0.057	−0.498	−0.215	−1.269	−1.685	1.918	2.039	0.020	−1.187	1.665	−246.447	1.380	4.009
Glnβ	0.030	−0.498	−0.250	−1.282	−1.585	1.949	1.974	0.071	−1.188	1.514	−246.334	1.398	3.918
Gluα	0.056	−0.590	−0.695	−0.305	2.381	−0.820	0.229	0.272	1.117	−2.652	−266.277	0.962	3.339
Gluβ	0.024	−0.610	−0.720	−0.300	2.389	−0.767	0.195	0.394	1.059	−2.783	−266.211	0.990	3.377
Glyα	0.009	0.057	0.112	0.047	−0.054	0.235	0.072	0.186	0.161	−0.132	−0.625	0.134	0.389
Glyβ	−0.013	0.071	0.119	0.043	−0.059	0.227	0.055	0.184	0.172	−0.125	−0.639	0.145	0.384
Hisα	0.060	−0.202	−0.049	−0.322	−8.537	7.961	3.335	1.897	−0.924	6.641	−263.406	0.383	13.449
Hisβ	0.027	−0.223	−0.134	−0.279	−8.389	7.876	3.312	1.780	−0.788	6.609	−263.285	0.381	13.278
Ileα	0.096	0.200	0.256	0.153	−0.392	−0.370	0.269	−0.085	−0.016	0.477	−157.092	0.359	0.734
Ileβ	0.068	0.150	0.226	0.168	−0.425	−0.348	0.173	−0.078	−0.128	0.503	−157.015	0.319	0.719
Leuα	0.089	0.161	0.199	0.180	0.345	−0.108	0.371	−0.549	−0.155	0.204	−157.074	0.313	0.734
Leuβ	0.061	0.187	0.134	0.140	0.690	−0.109	0.156	−0.704	−0.284	0.014	−157.003	0.269	0.897
Lysα	0.084	−0.044	−0.089	−0.006	−0.973	−0.711	1.935	1.428	0.440	−0.455	−212.018	0.100	2.837
Lysβ	0.056	−0.046	−0.114	−0.026	−0.875	−0.667	1.844	1.466	0.413	−0.591	−211.917	0.125	2.745
Metα	0.060	0.444	1.073	−0.518	−3.837	3.140	2.413	0.404	0.854	3.433	−515.708	1.271	6.297
Metβ	0.034	0.445	1.053	−0.542	−3.738	3.188	2.353	0.437	0.842	3.301	−515.530	1.265	6.211
Pheα	0.054	0.142	0.175	0.092	3.945	−13.983	2.317	−14.195	3.501	10.249	−269.379	0.243	22.337
Pheβ	0.025	0.116	0.227	0.066	4.051	−13.910	2.199	−14.163	3.500	10.111	−269.256	0.263	22.217
Proα	0.507	−0.052	0.720	0.468	−0.910	−0.423	0.577	0.234	−0.279	0.677	−79.114	0.860	1.297
Proβ	0.538	−0.037	0.751	0.488	−0.944	−0.462	0.452	0.149	−0.279	0.794	−79.019	0.896	1.300
Serα	−0.002	0.533	−0.236	0.476	−0.523	−0.544	0.838	1.281	−0.678	−0.758	−114.457	0.753	1.898
Serβ	−0.027	0.547	−0.261	0.439	−0.352	−0.587	0.705	1.196	−0.778	−0.844	−114.392	0.748	1.854
Thrα	0.016	0.582	−0.245	0.543	−0.142	−0.377	0.785	1.574	−0.485	−1.432	−153.578	0.833	2.087
Thrβ	−0.009	0.602	−0.261	0.522	−0.051	−0.325	0.705	1.630	−0.580	−1.579	−153.494	0.838	2.165
Trpα	0.061	0.382	0.108	−0.931	4.482	−20.815	−0.463	−20.818	−0.632	16.336	−400.232	1.012	32.538
Trpβ	0.057	0.366	0.193	−0.916	4.630	−20.623	−0.507	−20.746	−0.621	16.116	−400.035	1.005	32.285
Tyrα	0.064	0.185	0.133	−1.071	3.158	−14.477	1.498	−14.786	0.282	11.627	−344.144	1.095	22.915
Tyrβ	0.033	0.164	0.195	−1.087	3.246	−14.358	1.374	−14.688	0.280	11.442	−344.000	1.117	22.707
Valα	0.092	0.180	0.234	0.146	−0.147	0.223	0.372	0.348	−0.092	−0.201	−117.971	0.329	0.620
Valβ	0.062	0.130	0.206	0.164	−0.218	0.245	0.319	0.362	−0.180	−0.144	−117.916	0.294	0.626

[a] All values are in atomic units.

5.1. Side Chain Properties. The expected values of the side chain electronic properties were evaluated following eq 3. Table 2 shows the obtained values for net charge q, dipolar moment **M**, quadrupole moment **Q**, and energy E. Every independent component of the vectorial (**M**) and tensorial variables (**Q**) is included.

As seen in Table 2, all side chains but proline's ($q \simeq 0.5e$) are neutral, with their $\bar{q} = 0.036e$ (0.035). Proline is an imino acid with a pyrrolidine group cycled over the backbone at the amino N atom, which explains its nonzero net charge. Because the remaining side chains are essentially neutral, we will omit charge in the ensuing analysis.

This side chain neutrality implies a zero electric field flux across the side chain basin boundaries. Bader et al. have claimed that this feature is a necessary transferability probe for functional groups.[24] Additionally, this charge neutrality implies that, according to our model, high charge density moments (**M**, **Q**) play central roles in electrostatic interactions.

5.1.1. Transferability of Side Chain Properties. The backbone influence on the side chain is evidenced by the change in directionality for multipolar moments between α and β conformations (i.e., change in Cartesian components). Nevertheless, their corresponding magnitudes ($|\mathbf{M}|$ and $|\mathbf{Q}|$ in Table 2) have small changes between the two backbone conformations, indicating that the main influence of conformation is merely a reorientation of the charge density moments.

Minimal variations in atomic properties are necessary for atom transferability. A similar criterion can be argued for molecular fragment transferability, which is evidenced in the three final columns in Table 2. In fact, we found that the average difference between α and β is around 0.15% for energy $\Delta E_{\alpha\beta}$, $\cong$5.7% for the magnitude of the dipole moment $\Delta|\mathbf{M}|_{\alpha\beta}$, and $\cong$4.4% for the quadrupolar moment $\Delta|\mathbf{Q}|_{\alpha\beta}$. These small changes in the side chain property average values reflect the extent of their transferability.

Previous studies with amino acid models reveal that small variations in their nuclei geometry have little effect on the expected values for their molecular properties.[33] On the other hand, when considerable changes in backbone geometry take place, their electrical property values change significantly. In short oligopeptides, changes in the entire molecule properties are accounted for almost entirely by contributions coming from backbone atoms. What we found here is that side chain properties are hardly changed when only the backbone geometries are altered, as can be seen in Table 2. This implies that the backbone and the side chain contribute independently to amino acid property values. Thus, we argue that the side chains studied here can be considered as transferable groups from the AIM theory perspective.

According to AIM theory, there are two transferability conditions: (a) the neutrality of the studied group and (b) the conservation of expected values for variables for small changes

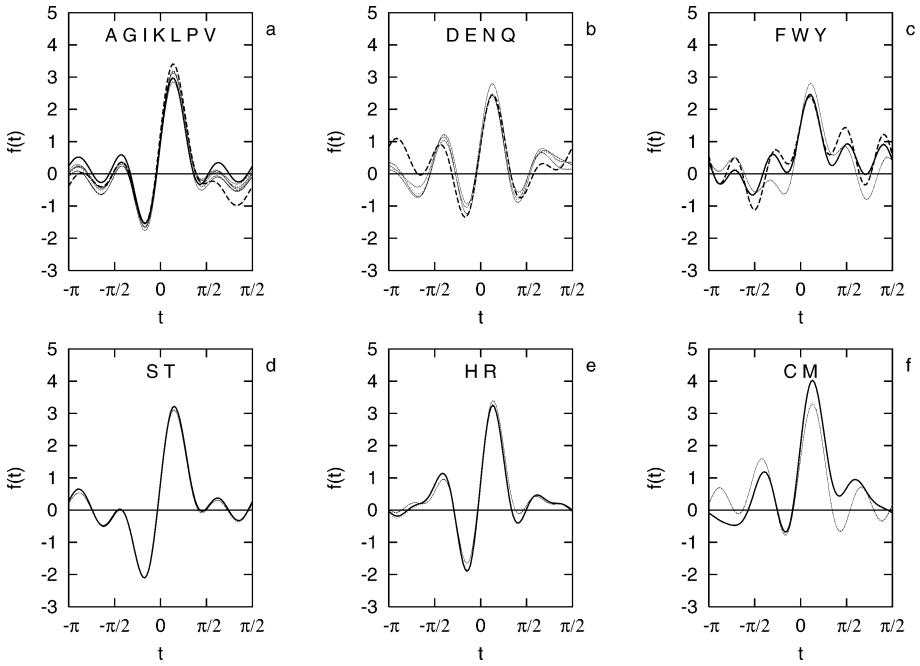

Figure 1. Main groups determined by Andrews plot analysis of the 20 side chain properties (Table 2). Each main group (a−c) was determined using eq 12 as those containing similar curves. Chemically related side chains exhibit similar curves: (a) Aliphatic side chains (thin lines) are almost overlapped except for Pro (bold line), and they appear to share several features with Lys (dashed line). (b) Between side chains containing polar groups, Glu (dashed line) is the less alike. (c) Aromatic side chains (Phe, thin line; Trp, bold line; Tyr, dashed line) share a common overall pattern besides the fact that between them there exist several differences. (d) The two side chains with alcohol groups (Ser, thin line; Thr, bold line) are overlapped, and it is evident that they have a pattern similar to that of the aliphatic group. (e) The basic side chains (Arg, thin line; His, bold line) are quite similar and share many of the characteristics portrayed by the polar group. (f) The two sulfur containing side chains (Cys, thin line; Met, bold line) have certain similarities, but while Cys shares features with the polar group, Met does not.

in nuclei geometry. In our research both conditions are satisfied; the amino acid side chains, excepting proline's, are neutral, and the variables change within narrow limits under variations of nuclei geometry.

5.2. Vector Representation. The normalized variables (eq 10) define a set of 40 points in 10 dimensions (charge was not included because, except for proline, the side chains are principally neutral), demanding a special graphic representation. Andrews plots are a kind of plot that permits representing *n*-dimensional data by a unidimensional function $f(\hat{\mathbf{V}};t)$.[36,37] Basically, different variables are assigned to different aspects of a curve, in this case the amplitudes of different sine and cosine functions.

We used Andrews plots for representing our variables (normalized in such a way that these parameters are unitless) by grouping those with similar curve shapes. This corresponds to mapping the feature vector (in 10-D) onto frequency components. Each depicted function obeys the following equation:

$$f(\hat{\mathbf{V}};t) = \frac{\hat{E}}{\sqrt{2}} + \hat{p}_x \sin(tg) + \hat{p}_y \cos(t) + \hat{p}_z \sin(2t) + \hat{Q}_{xx} \cos(2t) + \hat{Q}_{xy} \sin(3t) + \hat{Q}_{xz} \cos(3t) + \hat{Q}_{yy} \sin(4t) + \hat{Q}_{yz} \cos(4t) + \hat{Q}_{zz} \sin(5t) \quad \text{for} \quad t \in [-\pi, \pi] \quad (11)$$

By inspection of the common valleys, peaks, and *t*-axis intersections, we can unambiguously identify similarities between plots. Hence, we evaluated the integral of the absolute value of the difference function between all pairs of Andrews plots so as to quantitatively estimate their shape similarities:

$$s_{ij} = \int_{-\pi}^{\pi} |f(\hat{\mathbf{V}}_i;t) - f(\hat{\mathbf{V}}_j;t)| \, dt \quad \text{for} \quad i, j \in \{1, 2, ..., 40; i \neq j\} \quad (12)$$

So as to determine group membership, we scanned all side chains *i* and selected those other side chains *j* whose s_{ij} values defined by eq 12 fell into the group $\{s_{ij} | s_{ij} < (\bar{s}_{ij} - (\sigma_{ij}/2))\}$, where $\bar{s}_{ij}$ and σ_{ij} represent the mean and standard deviation of the compared samples, respectively. By this method, we found a group of closest amino acid side chains for each of the 40 molecular fragments we studied.

This procedure defines several subsets of similar plots which are shown in Figure 1. A remarkable feature is the pairing of α and β elements for all the cases under study. In Figure 1, β conformations appear as dotted plots, barely differentiable from the solid and dashed plots representing the corresponding α conformations.

5.3. Andrews Plot Classification. The sets determined by the criteria expressed in eq 12 were classified into two categories: main groups, characterized as those sharing basically the same group of similar cases, and minor groups, which appear to be marginally related to the main groups by several of their elements. Three main groups were found, which clearly are chemically related: aliphatic side chains {Ala, Gly, Ile, Leu, Pro, Val}, polar side chains {Asn, Asp, Gln, Glu}, and aromatic side chains {Phe, Trp, Tyr}. Related to the former group is the subset of aliphatic−alcohol side chains {Ser, Thr}. Related to the second group is a pair of basic side chains {Arg, His}. And finally, besides the sulfur-containing side chains being related each other, the Cys function is more like those pertaining to the polar group than Met. The function corresponding to Lys, which is plotted as a dashed curve in Figure 1a, reveals that this side chain shares several features with the aliphatic side chains, but it is also related to the polar groups, being a kind of intermediate case between these two main groups.

Andrews plots reveal the similarity existent between the molecular systems by comparing their property vectors **V**. This method is straightforward for establishing such a comparison,

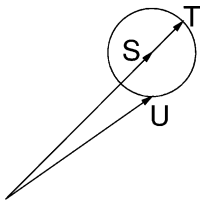

Figure 2. For vectors **S**, **T**, and **U**: $d_{ST} = d_{SU}$, $\theta_{ST} = 0$, and $\theta_{SU} > 0$. Therefore, while **S** is equally distanced to **T** and **U**, **S** shares the same angle as **T**. Hence, using both distance and angle measurements between vectors, we can achieve a more discerning comparison of multidimensional vectors, as in this case, where distance alone could not identify the differences between **T** and **U** from **S**.

giving account for the actual values of variables and requiring minimal mathematical effort. In the following subsection, the analysis of the same database is done by a standard clustering method, to explore by a different approach the classification of the side chains by the electronic multipoles.

5.4. Hierarchical Clustering Analysis. To accomplish a more detailed study of the structure of the groups that can be found with the property vectors **V**, we used a hierarchical clustering method. With this method it is possible to quantitatively compare a set of multidimensional vectors for determining the groups therein formed and their relative connections, which is commonly referred as the *topology* of the set. Two main issues determine topology: the metric and the clustering algorithm. The usual metric is the Euclidean norm, but there exist many other possibilities. The clustering algorithm itself is another key point which also determines the resulting topology. Algorithms such as UPGMA (unweighted pair group method with arithmetic mean) or neighbor joining (NJ) differ in the way they build clusters. We chose NJ as the method for our clustering study because this procedure is adequate for the vector comparison we pursued.[38]

5.4.1. Metric. We propose that the calculation of similarity between systems A and B, characterized by vectors $\mathbf{V}_A$ and $\mathbf{V}_B$, respectively, is given by two measurements, namely, the Euclidean distance, d_{AB}, and the angle between these, θ_{AB}:

$$d_{AB} = |\mathbf{V}_A - \mathbf{V}_B| \tag{13}$$

$$\theta_{AB} = \cos^{-1}\left(\frac{\mathbf{V}_A \cdot \mathbf{V}_B}{|\mathbf{V}_A| \cdot |\mathbf{V}_B|}\right) \tag{14}$$

The inclusion of the angle θ_{AB} between vectors $\mathbf{V}_A$ and $\mathbf{V}_B$ gives a more restrictive similarity definition, as can be seen in Figure 2, justified by the nature of selected variables, where orientation plays an important role. The angle θ_{AB} depends on the choice of the coordinate system, while the distance d_{AB} does not. We have already mentioned that the comparison proposed is referenced to a particular coordinate system (with the origin located at the C_α in the present case), with respect to which all variables are evaluated. To compare d_{AB} and θ_{AB} in the same plane, we chose the normalized values ($d_{AB}/\sqrt{3}$) and (θ_{AB}/π). Each denominator is the theoretical maximum value of each variable.

Because we chose two parameters to be used for the measurement of molecular similarity, $d_{AB}/\sqrt{3}$ and θ_{AB}/π, the Euclidean distance was taken as the value of proximity between vectors $\mathbf{V}_A$ and $\mathbf{V}_B$ (in normalized units):

$$P_{AB} = \sqrt{d_{AB}^2 + \theta_{AB}^2} \tag{15}$$

Clearly, this proximity measurement is smaller if the compared vectors are not only closer but also well aligned.

5.4.2. Side Chain Dendrogram. The similarity indexes (d_{ij}, θ_{ij}), as described in eqs 13 and 14, were computed for each side chain. As an example, the (d_{ij}, θ_{ij}) values for Ala_α appear in Figure 3. In this plot, every other side chain can be located according to its (d_{ij}, θ_{ij}) values from the target side chain (Ala_α in this case). As can be seen, the Euclidean distance alone does not provide enough information for deciding similarity with several other side chains. For example, Pro and Ser exhibit identical d values, making this parameter insufficient to determine the similarity between these two molecules. On the other hand, the angle between Ala and Pro is smaller than that between Ala and Ser, indicating that the Pro side chain is more like Ala than Ser. The proximity measurement P (eq 15) is shorter between Ala and Pro, highlighting the previous assertion.

We used the NJ algorithm for the hierarchical clustering, employing the aforementioned normalized vectors and the proximity measurement P as the distance criteria. These calculations were performed using the MEGA 2.1 software.[39]

The resulting dendrogram appears in Figure 4. This plot consists of many U-shaped chords connecting the vectors in a hierarchical tree. The length of each chord represents the proximity P (eq 15) between the two vectors being connected. In this plot, every major branch, that is, those grouping several vectors from the main branch located at the center, is thicker than the secondary branches (for a brief description of cluster analysis applied to atomic properties, see ref 40). As can be seen, the aforementioned features already identified by Andrews plot analyses appear, but with an additional level of detail. This dendrogram accounts for the average occurrence of neighboring side chains as described by the proximity measurement. Thus, the branched structure obtained contains the information from the 40 proximity plots constructed in a similar manner as that shown in Figure 3.

Side chain conformers in α and β were the closest neighbors in all cases. The side chains are clustered into three major branches (marked with a black dot in Figure 4), with several groups therein. Side chains located inside the main branches (highlighted by different gray tones in Figure 4) share biochemical function, several of which are indicated in Figure 4. These three main branches are exactly the same groups found by Andrews plots. Nevertheless, Lys and Met appear to be isolated points in the classification. This is observable from Figure 1, in which clearly those two side chains share several features with other side chains, but they are apart enough in order to not be included in the resulting main groups.

The larger main group is composed of the aliphatic side chains {{{Ala, Gly}, Pro}, {{Leu, Ile}, Val}, {Ser, Thr}}. The second group is composed of polar side chains {{Arg, His}, {Cys, {Glu, {Asp, {Asn, Gln}}}}}. The other cluster is composed of aromatic side chains {Phe, {Trp, Tyr}}. Inside each major branch, chemically related side chains are paired. For example, Arg and His are the only basic side chains with π electrons on them, and Ser and Thr are those residues with alcohol groups in their side chain. Ala, Pro, and Gly are nested into the same branch, being that they have the shortest side chains. Ile, Leu, and Val have the most hydrophobic side chains. On the other hand, Tyr and Trp are paired, sharing the fact that their side chains include polar groups, thus being less hydrophobic than Phe.

The resulting dendrogram shows quantitative individual similarities as well as group dissimilarities that are chemically relevant. While every side chain belongs to a given functional

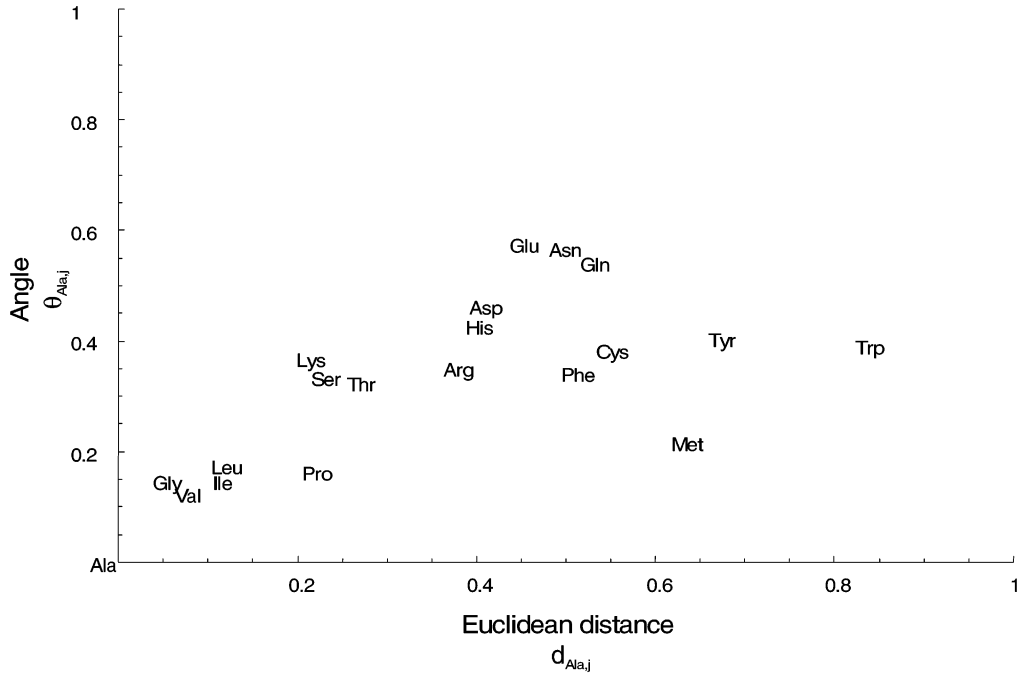

Figure 3. Proximity plot in normalized units for Ala_α. While several side chains are located at the same distance *d* from Ala, they differ in the value of the angle θ, as in the case of Lys and Pro. Consistently, the proximity measurement *P* (eq 15) between Ala and Pro is shorter than that to Lys.

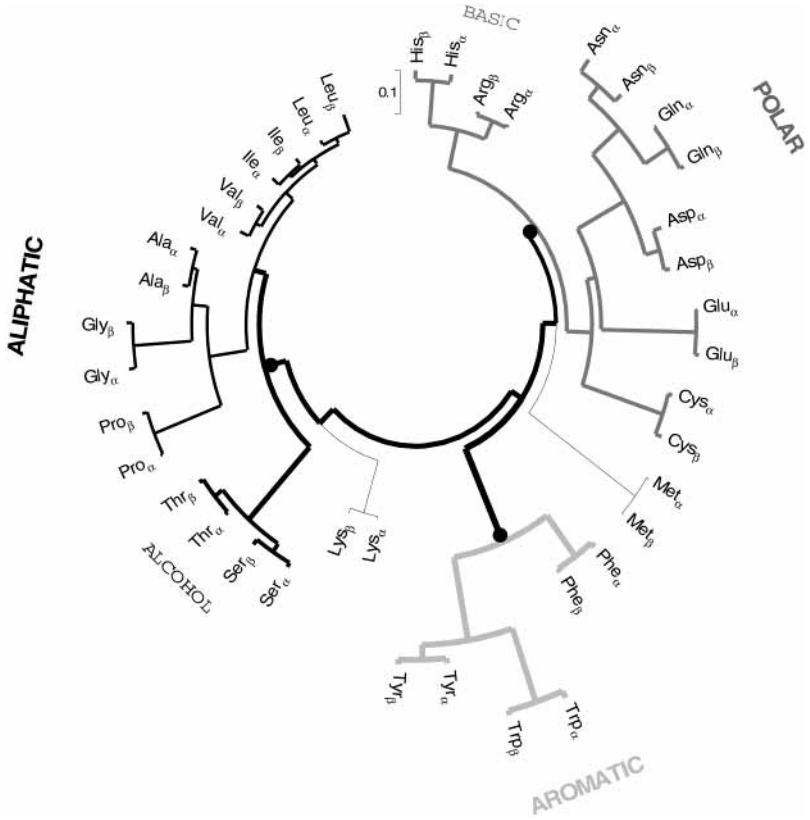

Figure 4. Dendrogram generated by the neighbor joining cluster analysis of side chain types.

class, they exhibit idiosyncratic behavior; that is, there are no synonymous replaceable side chains.

Another interesting finding is that our methods are able to simultaneously classify both small and large molecular fragments, as evidenced by the inclusion of Gly, a one-atom fragment, with the aliphatics, and the inclusion of Trp, with 18 times more nuclei than Gly, with the aromatics. This is in contrast with other methods that have to leave Gly aside as an outlier.

The consistency between physical variables and biochemical data verified by two independent methods (Andrews plot and hierarchical clustering classification) opens the possibility for

using the same strategy in the de novo design of functional groups, using our approach and database as a starting point.

6. Conclusions

This research has investigated theoretical similarities and transferability between side chains of genetically encoded amino acids in two backbone conformations. Each side chain was studied by the expected values of electronic energy and dipole and quadrupole moments. Two main differences with previous amino acid classifications have been addressed in the present work. First, we used a small set of theoretical variables which are all electronic in nature and consider only side chain contributions; and second, we used two independent methods for similarity determination between side chains obtaining the same relevant groups.

We found that each of these physical observables is related to primary physicochemical features. For example, molecular size was accounted for by the expected value of energy; the presence of polarizable groups is highlighted by their dipolar moments; and, the presence of π electrons is directly related to higher quadrupole moments. However, no single variable accounts for the overall side chain classification.

On the basis of our proposed 2-fold similarity criteria (eq 15) used in the NJ classification, and by the comparison of Andrews plots, the same result appears: chemically related side chains are nested into several groups resembling an amino acid classification by functional traits, with subsets ordered by properties such as hydropathy, polarity, size, aromaticity, and presence of carbon chains (aliphatic) and hydroxyl groups (alcohol). The parallels between our classification and the latter denote the usefulness of our approximation and promises for the application of this methodology to other situations.

We have found that molecular fragment neutrality, a necessary requirement for transferability, is preserved in all side chains but proline, irrespective of backbone conformation. As an additional transferability requirement, the changes in the values of the variables for α and β conformations were small in comparison to the differences between side chains for these variables, highlighting the independence of the molecular fragments from external influences.

Previous studies on amino acids using AIM established the effect of geometry on atomic properties directed toward the theoretical construction of polypeptides.[24−26] The results reported here go further by exploring the consistency between electronic properties and biochemical trends, emphasizing the utility of AIM for comparative purposes in biomolecular design. Because of the generality of the physical variables used, we propose that our results bring another view to these observables and, thus, open a new approach to the determination of biomolecular reactivity for drug design based on electronic comparative analyses.

Acknowledgment. We wish to express our gratitude to the Colombian Ministry of Public Health and Protection for financial support and to Colciencias for their supervision of this research project. We also wish to highlight the thorough and thoughtful feedback of the members of the Grupo de Química Teórica of the Universidad Nacional de Colombia.

References and Notes

(1) Sneath, P. *J. Theor. Biol.* **1966**, *12*, 157−159.

(2) Dayhoff, M.; Schwartz, R.; Orcutt, B. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1978; Vol. 5, Chapter 22, pp 345−352.

(3) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. *J. Protein Chem.* **1985**, *4* (1), 23−55.

(4) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. *J. Protein Chem.* **1985**, *4* (1), 265−297.

(5) Stanfel, L. *J. Theor. Biol.* **1996**, *183*, 195−205.

(6) Carbó, R.; Layda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(7) Cioslowski, J.; Fleischmann, E. *J. Am. Chem. Soc.* **1991**, *113*, 64−67.

(8) Walker, P.; Arteca, G.; Mezey, P. *J. Comput. Chem.* **1991**, *12*, 220−230.

(9) Burgess, E.; Ruell, J.; Zalkow, L.; Haugwitz, R. *J. Med. Chem.* **1995**, *38*, 1635−1640.

(10) Popelier, P. *J. Phys. Chem. A* **1999**, *103* (4), 2883−2890.

(11) O'Brien, S.; Popelier, P. *Can. J. Chem.* **1999**, *77*, 28−36.

(12) Price, S.; Adrews, J.; Murray, C.; Amos, R. *J. Am. Chem. Soc.* **1992**, *114* (76), 8268−8276.

(13) Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; Dover: Mineola, New York, 1996.

(14) Grant, J.; Pickup, B. In *Computer Simulation of Biomolecular Systems*; van Gunsteren, W., Weiner, P., Wilkinson, A., Eds.; Kluwer/Escom: Kluwer Academic Publishers: Dordretch, The Netherlands, 1997; Vol. 3, Chapter 1, pp 150−176.

(15) Walker, P.; Mezey, P. *J. Am. Chem. Soc.* **1994**, *116*, 12022−12032.

(16) Jelsch, C.; Pichon-Pesme, V.; Lecomte, C.; Aubry, A. *Acta Crystallogr.* **1998**, *D54*, 1306−1318.

(17) Pichon-Pesme, V.; Lecomte, C.; Wiest, R.; Benard, M. *J. Am. Chem. Soc.* **1992**, *114*, 2713−2715.

(18) Savin, A.; Nesper, R.; Wengert, S.; Fässler, T. *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 1808−1832.

(19) Bader, R. *Atoms in Molecules: A Quantum Theory*; Vol. 22 of The International series on monographs on chemistry; Oxford University Press: Oxford, 1995.

(20) Bader, R. *Phys. Rev. B* **1994**, *49*, 13348−13356.

(21) Wiberg, K.; Rablen, P. *J. Comput. Chem.* **1993**, *14*, 1504−1518.

(22) Angyan, J.; Jansen, G.; Loos, M.; Hattig, C.; Hess, B. *Chem. Phys. Lett.* **1994**, *219*, 267.

(23) Chang, C.; Bader, R. *J. Phys. Chem.* **1992**, *96* (4), 1654−1662.

(24) Popelier, P.; Bader, R. *J. Phys. Chem.* **1994**, *98*, 4473−4481.

(25) Matta, C. F.; Bader, R. W. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 310−329.

(26) Matta, C. F.; Bader, R. W. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 519−538.

(27) Popelier, P. *Mol. Phys.* **1996**, *87*, 1169−1187.

(28) Bader, R.; Bayles, D. *J. Phys. Chem. A* **2000**, *104* (23), 5579−5589.

(29) Bader, R. *Chem. Phys. Lett.* **1988**, *148*, 452−458.

(30) Lee, K.; Xie, D.; Freire, E.; Amzel, L. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 68−84.

(31) Schrauber, H.; Eisenhaber, F.; Argos, P. *J. Mol. Biol.* **1993**, *230*, 592−612.

(32) Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, E.; Brice, M.; Rogers, J.; Kennard, O.; Shimanouch, T.; Tasumi, M. *J. Mol. Biol* **1977**, *112*, 535−542.

(33) Cárdenas, C.; Obregón, M.; Llanos, E.; Machado, E.; Bohórquez, H.; Villaveces, J.; Patarroyo, M. *J. Comput. Chem.* **2002**, *26* (6), 631−646.

(34) Biegler-König, F. W.; Bader, R.; Tang, T. *J. Comput. Chem.* **1982**, *3*, 317.

(35) 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. *Gaussian '94*, revision D.4; Gaussian Inc.: Pittsburgh PA, 1995.

(36) Andrews, D. *Biometrics* **1972**, *28*, 125−136.

(37) Everitt, B.; Dunn, G. *Applied Multivariate Data Analysis*; Oxford University Press: New York, 1992.

(38) Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: Oxford, New York, 2000; Chapter 6, pp 87−111.

(39) Kumar, S.; Tamura, K.; Jakobsen, I.; Nei, M. *Bioinformatics* **2001**, *17*, 1244−1245.

(40) Popelier, P. L. A.; Aicken, F. M. *J. Am. Chem. Soc.* **2003**, *125*, 1284−1292.