Ab Initio Quality Electron Densities for Proteins: A MEDLA Approach

P. Duane Walker and Paul G. Mezey*,†

Contribution from the Mathematical Chemistry Research Unit, Department of Chemistry, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0

Received March 15, 1994[⊗]

Abstract: A computational technique has been developed for the construction of *ab initio* quality electron density distributions for large peptides and proteins. A database of a specialized set of molecular density fragments is constructed for use with the Molecular Electron Density Lego Assembler (MEDLA) method for building molecular electron densities of biopolymers composed from amino acids. The MEDLA program uses *ab initio* electron densities from the molecular fragment database and a set of atomic coordinates available, *e.g.*, from X-ray diffraction experiments or from a molecular modeling program such as BIOGRAF, to construct the electron density for any specified conformation of the molecule. The current database can be used to compute the electron density distributions for any peptides and proteins that are made up of the 20 most common amino acids. The MEDLA program generates *ab initio* quality, three-dimensional electron densities for much larger molecules than those which could be computed at present using conventional *ab initio* methods. Even for molecules of >1000 atoms, the MEDLA method requires minimal computational time. The method generates the electronic density for the entire molecule or if desired for any specific molecular fragment such as the backbone of a protein. The entire range of the electron density distribution is computed, from which molecular isodensity contour (MIDCO) surfaces for any density threshold value can be constructed using AVS or other visualization packages. The MIDCO surfaces provide a far more realistic description of molecular shape than the commonly used fused sphere Van der Waals surfaces or solvent accessible surfaces based on spherical atom models. In this work, *ab initio* quality electron densities are calculated for simple model peptides, for some important bioactive peptides in low-energy conformations, for the globular protein crambin comprised of 642 atoms in 46 amino acid residues, and for the gene 5 protein, made up of 87 residues with a total of 1384 atoms.

1. Introduction

Bioactive peptides and proteins are of enormous interest to biochemists, pharmacologists, and organic chemists. The secondary structure of large peptides and proteins has usually been depicted by ball and stick or wire frame models for the molecular skeleton, and by fused sphere Van der Waals surfaces for their 3D space filling characteristics.[1] Such models are useful but represent only a drastic oversimplification of the real, fuzzy molecular bodies and their actual, detailed shape features.

In reality, the electron density cloud represents the actual molecular body, hence electron densities offer a better choice for analyzing the shapes of molecules. Both high- and low-density features are important: high-density threshold MIDCOs describe the skeletal, "bonding" features, whereas the low-density threshold MIDCOs describe the space filling aspects and space requirements of molecular bodies.[2-4] Clearly, all shape, size, bonding, and conformational features of molecules are revealed by electron density; however, the computation of the electron densities using conventional *ab initio* techniques is limited to peptides containing <100 atoms, with the exception of periodic systems.

There have been several attempts to circumvent the formidable computational problems. Some techniques use a block diagonalized density matrix which eliminates all diatom interactions.[5] While this approach may be acceptable for the core regions of the density, it is not expected to describe bonding and molecular shape and size features adequately. Even accepting these limitations, the calculation of the electron density for even a small protein using a minimal basis set would still be computationally very expensive.

Bader *et al.* have used the atoms in molecules method[6,7] to generate molecular fragments based on the gradient of the electron density. Contour lines of density cross sections obtained from this approach compare well with *ab initio* results for the small peptides considered in ref 8, except at low density where the contour lines become disjointed. This separation of the contour lines is an artifact of the atoms in molecules method due to the presence of *boundaries* for the constituent fragments in the molecular system. When these fragments of fixed boundaries are placed into a different molecular environment, small gaps (of zero density) and/or local overlaps (of double density) occur in the chemically important bonding region between the fragments.

Bénard *et al.* construct electron density cross sections for large molecules by considering a fragment of the molecule whose constituent atoms satisfy a distance criterion from the plane of the cross section.[9] The aim of the technique is not the modeling

† Also at Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0.

⊗ Abstract published in *Advance ACS Abstracts*, December 1, 1994.

(1) Voet, D.; Voet, J. G. *Biochemistry* Wiley: New York, 1990.

(2) Mezey, P. G. *J. Comput. Chem.* **1987**, *8*, 462.

(3) Mezey, P. G. Three-Dimensional Topological Aspects of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.

(4) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1991**, *12*, 220.

(5) *HYPERCHEM, Release 2*; Autodesk, Inc.: 2320 Marinship Way, Sausalito, CA 94965, 1992.

(6) Bader, R. W. F.; Nguyen-Dang, T. T. *Adv. Quantum Chem.* **1981**, *14*, 63.

(7) Bader, R. W. F. *Acc. Chem. Res.* **1985**, *9*, 18.

(8) Chang, C.; Bader, R. W. F. *J. Phys. Chem.* **1992**, *96*, 1654.

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

0002-7863/94/1516-12022$04.50/0 © 1994 American Chemical Society

of the total density but the generation of difference distributions with respect to a superposition of neutral, noninteracting atoms. The terminal ends of the fragment are tied off with hydrogen atoms and an SCF calculation is done for the fragment. The resulting difference distributions are useful for an assessment of bonding characteristics, for example, in the case of the leuenkephalin peptide, they provide new evidence for the interactions between NH and CO groups.[10] However, this technique requires multiple SCF calculations if one needs more than one cross section, and it is not suitable for our purposes: for computing the entire 3D electron density distribution which is needed for the generation of MIDCOs.

Simplified models for electronic charge distributions, for example, formal charges assigned to atomic centers, have been used for both small and large molecules. For the generation of molecular electrostatic potentials for large systems, techniques based on atomic charges have been proposed.[11]

Recently, the Molecular Electron Density Lego Assembler (MEDLA) technique was developed to construct *ab initio* quality electron density distributions for large molecules from density distributions of small molecular fragments.[12] First, a database of *ab initio* electron densities of various molecular fragments is generated where the fragments are obtained from smaller parent molecules for which direct *ab initio* computation is feasible. Note that these fragment densities, just as the densities of entire molecules, are fuzzy electron distributions which have no boundaries. The large molecule whose density distribution is being constructed is partitioned into fragments which appear in the database. The density distributions for these fragments are sequentially rotated and translated to account for the actual arrangement of the fragment in the target molecule and then added together (superimposed) to model the electron density distribution. When superimposing fragment densities, a mutual interpenetration of the charge clouds occurs, and there is no density gap, no density doubling, and no other accumulation of error at any location of the merged fragments. This method is somewhat reminiscent to building structures using Lego blocks, where the mutual interpenetration of fuzzy fragments takes the role of snapping the blocks together. The electron density distribution calculated using the method was *quantitatively* shown to be very similar to that calculated for entire molecules using conventional *ab initio* packages at the 6-31G** level of basis, in fact, more accurate than direct *ab initio* results at the 3-21G level, while requiring only a small fraction of the computational time.[12]

The accuracy of the applications of the MEDLA method is currently limited to the 6-31G** level of basis, since the density fragment database has been constructed at this *ab initio* level. If, however, a database of more accurate density fragments is constructed, using, for example, a superior basis set, or correlated wave functions, then we expect a similar increase in the accuracy of the MEDLA technique.

In this work, a database of 21 molecular fragment densities for use with the MEDLA approach was created which allows for the calculation of electron densities of peptides and proteins comprised of the 20 standard amino acid residues. The database can be augmented as needed. Each of the 21 fragments can have several versions present in the database, where the versions have slightly different nuclear arrangements and/or they differ in the molecular surroundings in their parent molecules. By having a large enough variety for each fragment, and by selecting the version that matches best the actual arrangement of the fragment in the target molecule, high accuracy can be achieved. In practice, unless specialized fragments are required to account for highly distorted configurations, the calculation of the electron densities of a very large class of molecules requires only very few molecular fragments from the database. The peptide electron density can be calculated using fragment densities from the database and atomic coordinate information from crystallographic databases, or directly from a BIOGRAF[13] or similar structure file for any desired conformation of the given molecule. To demonstrate the power of the method, several bioactive peptides and two proteins are analyzed.

The paper is organized as follows. In the next section, the MEDLA approach is explained briefly, where the electron density of a dipeptide, glycinal alanine, is used as an example to show how the method works, and additional test calculations are performed on molecular arrangements with hydrogen bonds and other significant interactions. Section 3 considers a peptide with an ideal α-helical secondary structure. Also demonstrated in this section is the ability of the MEDLA program to calculate the electron density for large molecular fragments such as the backbone of a peptide. In section 4, the electron densities are calculated for three bioactive peptides in low energy conformations, as well as for two proteins, one containing 46 amino acid residues and three sulfur bridges, and a larger one containing 87 amino acid residues. The final section includes some closing comments as well as plans for future work.

2. Molecular Electron Density Assembler for Polypeptides

The method is based on a simple electron density fragment additivity principle.[12] If n is the number of atomic orbitals $\varphi_i(\mathbf{r})$ ($i = 1, 2, \ldots, n$) in an LCAO *ab initio* wave function of a molecule, $\mathbf{r}$ is the position vector variable, and $\mathbf{P}$ is the corresponding $n \times n$ density matrix, then the electronic density $\varrho(\mathbf{r})$ of the molecule is given by

$$\varrho(\mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\,\varphi_i(\mathbf{r})\,\varphi_j(\mathbf{r}) \tag{1}$$

An arbitrary collection of nuclei from the molecule can be used to define the kth fragment $\varrho k(\mathbf{r})$ of the electron density $\varrho(\mathbf{r})$, using the following criterion for the $n \times n$ fragment density matrix P^k_{ij}:

$$\begin{aligned} P^k_{ij} &= P_{ij} \text{ if both } \varphi_i(\mathbf{r}) \text{ and } \varphi_j(\mathbf{r}) \text{ are AO's centered on} \\ &\quad \text{nuclei of the fragment} \\ &= 0.5P_{ij} \text{ if precisely one of } \varphi_i(\mathbf{r}) \text{ and } \varphi_i(\mathbf{r}) \text{ is centered on} \\ &\quad \text{a nucleus of the fragment} \\ &= 0 \text{ otherwise} \end{aligned} \tag{2}$$

The electron density of the kth fragment is defined as

$$\varrho^k(\mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{n} P^k_{ij}\,\varphi_i(\mathbf{r})\,\varphi_j(\mathbf{r}) \tag{3}$$

If the nuclei of the molecule are partitioned into m mutually exclusive groups to generate m fragments, then the sum of the fragment density matrices is the density matrix of the molecule, and the sum of the fragment densities is the density of the molecule:

(10) Wiest, R.; Pichon-Pesme, V.; Benard, M.; Lecomte, C. *J. Phys. Chem.* **1994**, *98*, 1351.

(11) (a) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 4915. (b) Price, S. L.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755.

(12) Walker, P. D.; Mezey, P. G. *J. Am. Chem. Soc.* **1993**, *115*, 12423.

(13) *BIOGRAF*; Biodesign, Inc.; 199 S. Los Robles Ave., Pasadena, CA 91101, 1988.

$$P_{ij} = \sum_{k=1}^{m} P^{k}_{ij} \tag{4}$$

and

$$\varrho(\mathbf{r}) = \sum_{k=1}^{m} \varrho^{k}(\mathbf{r}) \tag{5}$$

The simple additivity rules 4 and 5 are *exact* on the given *ab initio* LCAO level.

Such fragment densities can be combined to form approximate electron density for a different molecule by selecting and arranging fragments so that the nuclear positions match those in the target molecule. This procedure, the Molecular Electron Density Lego Assembler (MEDLA) approach, has been shown[12] to produce approximate electron densities that are *quantitatively* very similar to densities obtained in direct *ab initio* calculations using a 6-31G** level of basis; in fact, the corresponding MIDCOs are found to be visually indistinguishable for a family of small molecules, and more accurate than direct 3-21G *ab initio* results. The technique not only is applicable for building electron densities for large molecules but also serves as a new tool that extends the scope of earlier global and local shape analysis methods of molecular fragments[14] and complete molecules.[15]

One class of molecules ideally suited for this approach is that of the polypeptides made up of the 20 common amino acids. It was found that a database of only 21 types of molecular fragments could yield a peptide containing any of these amino acids in most conformations. The required fragments are shown in Figure 1. The peptide bond is described by either a pair of fragments (**1**,**2**) or a single fragment (**3**). The difference between fragments **15** and **16** is that the sulfur in **15** is bonded to a carbon atom while in **16** the sulfur is part of a sulfur bridge.

Our MEDLA program[16] was adapted to read the atomic coordinate output files from BIOGRAF directly and compute the electron density for the peptide. As an example, the electron density distribution was calculated for a dipeptide, glycinal alanine, in its zwitterion form. In this molecule, seven molecular fragments were required, **5**, **8**, **1**, **2**, **9**, **7**, and **12**. The calculation of the entire density distribution within a cube of edge length of 25 au (atomic unit), with a resolution of 0.3 au, took 31 s on a Kubota 3000 workstation.

Figure 2 shows MIDCOs for four density threshold values for each of two views for the dipeptide. The top view has the plane of the peptide bond system perpendicular to the page, while the bottom view shows the peptide bond in the plane of the page. Note that although only four threshold values are shown in the figure, the entire density distribution, from the very high values around the nuclei to the very low values in the peripheral regions of the molecule, is computed using the MEDLA program, and MIDCOs for any other density thresholds can also be displayed.

The electron density for this molecule could be calculated using conventional *ab initio* techniques albeit requiring much longer computational time. However, as we consider larger molecules, the computational time of the MEDLA approach increases only linearly with the number of molecular fragments,

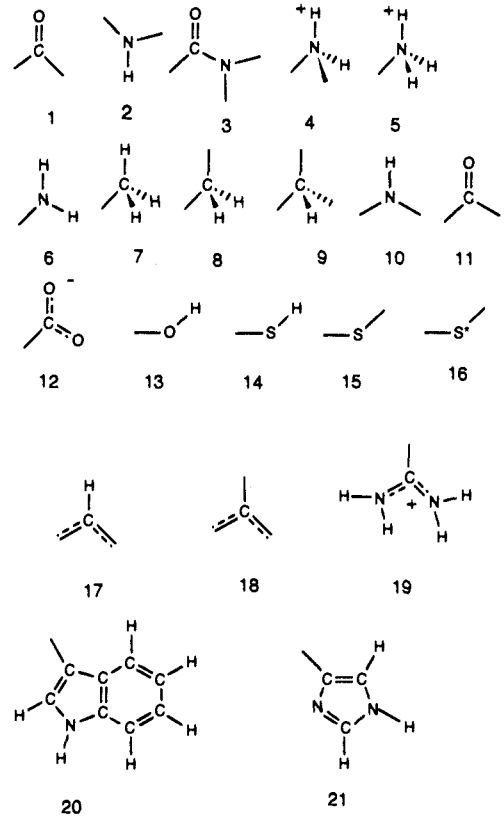


Figure 1. The 21 types of molecular fragments contained in the MEDLA polypeptide electron density fragment database. Only the atoms labeled by letters are contained in the fragment. Note that the entire peptide bond moiety is contained in the molecular fragments (either **1** and **2** or **3**).

whereas in a conventional *ab initio* calculation the time required increases at a higher power of the number of atomic basis functions. A comparison of memory requirements is even more favorable for the MEDLA approach. Using conventional *ab initio* approaches, the electron density distributions for the rest of the molecules considered in this study could not be calculated on the Kubota 3000 workstation used in our laboratory; furthermore, we estimate that a similar quality direct *ab initio* electron distribution of the gene 5 protein could not be calculated even on a CRAY supercomputer in less than a century of CPU time. By contrast, the longest CPU time required for a MEDLA calculation in this work was 21 min on our workstation, for the gene 5 protein.

Various tests have been carried out to assess the quality of the MEDLA densities. The purpose of the comparisons of the 6-31G** MEDLA results with direct 3-21G and direct 6-31G** *ab initio* results is to demonstrate that the MEDLA method generates *ab initio* quality densities. Detailed comparisons show that the MEDLA densities are of *ab initio* quality, at a level *better* than direct 3-21G *ab initio* calculations. The MEDLA results are nearly indistinguishable from results at the direct 6-31G** *ab initio* level, the level used for the construction of the MEDLA fragments. In our earlier MEDLA study[12] we have tested the technique only for small molecules with no significant interactions between non-connected fragments. In this work, large systems are considered in conformations which allow considerable interactions between fragments. To determine the accuracy of the method under such conditions, three additional tests were performed.

In the first test, test A, the density distribution for the dipeptide shown earlier was calculated using the direct *ab initio* method with two basis sets (3-21G and 6-31G**), in order to

(14) Mezey, P. G. *Int. J. Quantum Chem. Quant. Biol. Symp.* **1987**, *14*, 127.

(15) Mezey, P. G. *Shape in Chemistry: An Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1993.

(16) Walker, P. D.; Mezey, P. G. *MEDLA 93*; Mathematical Chemistry Research Unit: University of Saskatchewan, Saskatoon, Canada S7N 0W0, 1993.

(17) Good, A.; Richards, W. G. *J. Chem. Inf. Sci.* **1992**, *33*, 112.

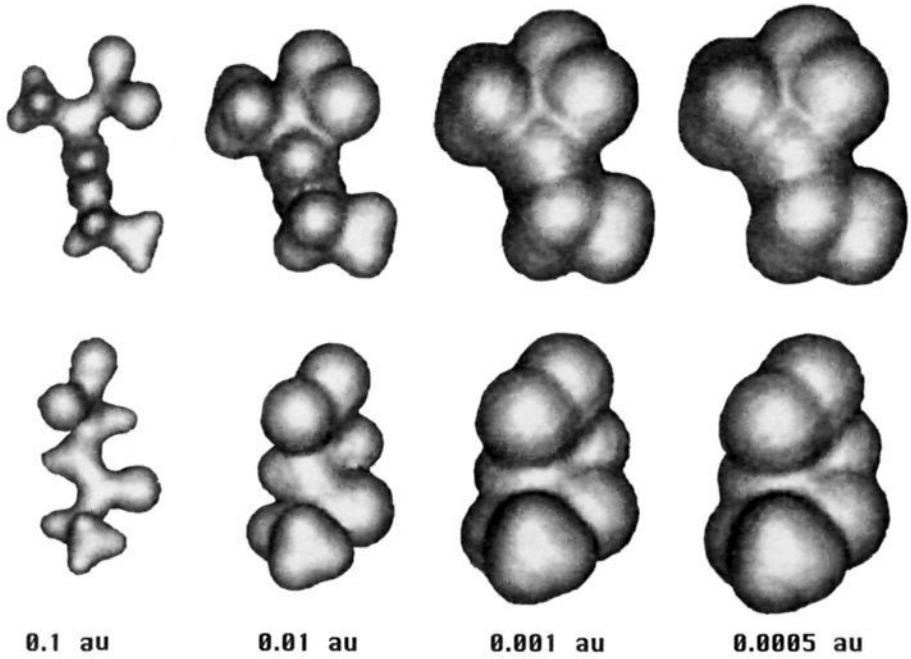


Figure 2. MIDCOs for a dipeptide, glycinal alanine. Two views are shown for each of four threshold values of density. The top view has the plane of the peptide bond perpendicular to the page, while in the bottom view the peptide bond lies in the plane of the page.

assess the reliability of the MEDLA peptide bond system. Figure 3 shows one view of MIDCOs of four different threshold values for the direct *ab initio* 3-21G, 6-31G**, and MEDLA results. The figure clearly shows that the MEDLA MIDCOs approximate the 6-31G** MIDCOs better than the 3-21G MIDCOs do. This conclusion is the most striking for the threshold value of 0.3 au.

Test B assessed the MEDLA representation of an important "nonbonding" interaction: hydrogen bond. The molecular system is a fragment of a helical peptide with a hydrogen bond between the first and fourth amino acid residues. The two ends of the fragment are "tied off" with methyl groups. Figure 4 shows the calculated MIDCO surfaces for four threshold values, for the direct *ab initio* 3-21G and 6-31G** basis results, and for the MEDLA density distributions. According to these results, as most clearly shown by the MIDCOs for the threshold value of 0.007 au, the MEDLA method provides a better approximation of the 6-31G** hydrogen bond than the direct *ab initio* results using the 3-21G basis set. This demonstrates that at the 6-31G** basis set level the *shapes* of the hydrogen bonds are mainly determined by the mutual interpenetration of the densities of the two local moieties making up the hydrogen bond, and can therefore be adequately represented within the MEDLA approach.

Test C compared the representations of the nonbonding interaction between $-SCH_3$ and $-Ph$ fragments. These two fragments are taken in the same arrangement as they appear in a conformation of the pentapeptide, metenkephalin, considered in a later section of this work. In Figure 5, MIDCOs are shown for four values of density. Again, we see the MEDLA MIDCOs match the direct *ab initio* 6-31G** MIDCOs better than the 3-21G MIDCOs do. This is especially obvious at the density threshold of 0.003 au where we first see the interaction resulting in a merger of local charge clouds. This test also indicates that in the given metenkephalin conformation there appears a significant interaction between the benzene ring and the methylene C−H bond adjacent to the sulfur.

In all three tests, the MEDLA results are virtually indistinguishable from the direct *ab initio* 6-31G** results, whereas

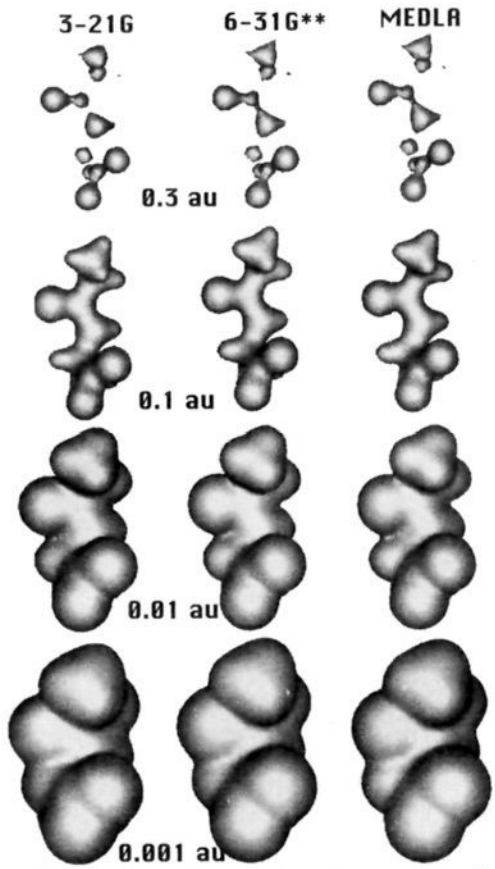


Figure 3. MIDCOs for a dipeptide, glycinal alanine, from direct *ab initio* 3-21G, 6-31G**, and MEDLA density distributions. MIDCOs for four threshold values of density are shown. The direct *ab initio* 6-31G** and MEDLA densities are visually indistinguishable, and both are of better quality than the direct *ab initio* 3-21G electron density. The topological distinctness of the 3-21G electron density is evident at the density threshold of 0.3 au.

the direct *ab initio* 3-21G results show more deviations. According to these findings and similar visual comparisons in

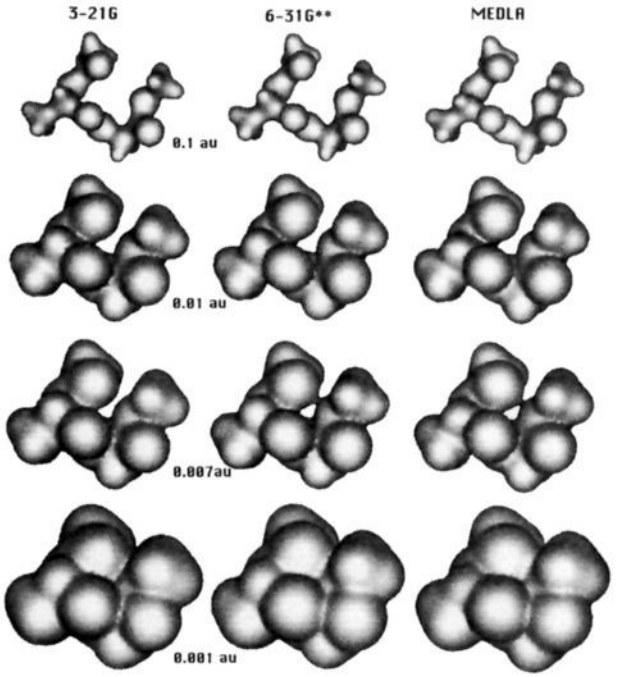


Figure 4. MIDCOs for a fragment of helical peptide obtained from 3-21G, 6-31G**, and MEDLA density distributions. MIDCOs for four threshold values of density are shown. The hydrogen bond completing a "doughnut" is well manifested at the 0.007 au density threshold for the virtually indistinguishable direct *ab initio* 6-31G** and MEDLA densities, which differ markedly from the less accurate 3-21G result, where the "doughnut" is incomplete.

ref 12, the MEDLA electron densities are of better quality than direct 3-21G *ab initio* electron densities.

Whereas these visual comparisons are convincing, the results of the above three tests were also studied quantitatively using two similarity measures. The first is a numerical similarity measure developed by Richards *et al.*[17] for use with two surfaces X and Y,

$$S_{XY} = \frac{B_{XY}}{(T_X T_Y)^{1/2}} \tag{6}$$

where B_{XY} is the number of points in the grid falling inside both surfaces, and T_X and T_Y are the number of points falling within surfaces X and Y, respectively. The MIDCOs considered had threshold values of 0.001 au. Note that by examining T_X and T_Y separately, we can also study the relative volumes of the two surfaces.

A *point-by-point* comparison of densities is the most thorough test of any numerical density representation. Such a test is provided by the direct similarity measure denoted by $L(a,a',X,Y)$. This measure compares two distributions within ranges of electron densities between some thresholds a and a', represented by density "shells" $S(a,a',X)$, where,

$$S(a,a',X) = \{\mathbf{r}:\ a \le \varrho_X(\mathbf{r}) \le a'\} \tag{7}$$

$$L^*(a,a',X,Y) = 1 - \sum_{\mathbf{r}\in S(a,a',X)} [(\varrho_X(\mathbf{r}) - \varrho_Y(\mathbf{r}))/\max(\varrho_X(\mathbf{r}),\varrho_Y(\mathbf{r}))]/n(S(a,a',X)) \tag{8}$$

and

$$L(a,a',X,Y) = [L^*(a,a',Y,X) + L^*(a,a',Y,X)]/2 \tag{9}$$

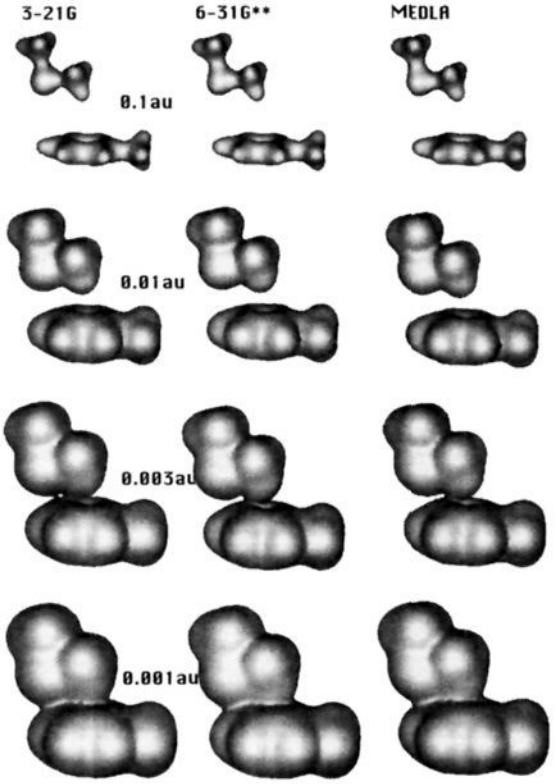


Figure 5. MIDCOs of interacting CH_3-S-CH_3 and $Ph-CH_3$ molecules from direct *ab initio* 3-21G, 6-31G**, and MEDLA density distributions. MIDCOs for four threshold values of density are shown, testing the relative quality of representations of the nonbonding interaction between $-SCH_3$ and $-Ph$ fragments. These two fragments are taken in the same arrangement as they appear in a conformation of metenkephalin. The agreement between the direct *ab initio* 6-31G** and MEDLA density distributions is excellent, whereas the direct *ab initio* 3-21G result deviates slightly from the other two, as shown at the density threshold of 0.003 au.

where $n(S(a,a',X))$ is the number of grid points falling in density shell $S(a,a',X)$. The quantity $L^*(a,a',X,Y)$ is one minus the average relative difference, if Y is compared to X on the grid, within the range a, a'. This similarity measure is not symmetric: $S(a,a',X)$ can differ from $S(a,a',Y)$ and therefore $L^*(a,a',X,Y)$ is not necessarily equal to $L^*(a,a',Y,X)$. For this reason an average of the two values is used for the similarity measure $L(a,a',X,Y)$. If the threshold values for $S(a,a',X)$ are chosen as nearly identical ($a \sim a'$), then measured $L(a,a',X,Y)$ compares individual MIDCOs. In this work four ranges of a, a' are used: 10−0.001, 10−0.1, 0.1−0.01, and 0.01−0.001 au. Table 1 shows the calculated values of these two similarity measures for the three tests, comparing both the 3-21G and the MEDLA density distributions with the 6-31G** distributions. We see that for the first similarity measure, the MEDLA electron densities have a higher computed similarity with the 6-31G** distributions than do the 3-21G densities. The more sensitive measure $L(a,a',X,Y)$ also shows better agreement between the MEDLA distributions and the 6-31G** distributions than between the 3-21G and 6-31G** results for most ranges of a and a'. Within the density range where the hydrogen bond is most clearly manifested (0.01 to 0.001 au), the hydrogen-bonded system shows $L(a,a',X,Y)$ being significantly higher between the MEDLA and 6-31G** results then between the 3-21G and 6-31G** results. This range of density is where the 3-21G results compare the worst for all three structures. Since this is the density range usually associated with formal "molecular volumes", we expect that the MEDLA method will provide

Table 1. Computed Similarities for MEDLA MIDCOs vs Direct *ab Initio* MIDCOs for the Three Test Systems A (Figure 3), B (Figure 4), and C (Figure 5), Using Two Similarity Measures, S_{XY} and $L(a,a',X,Y)$[a]

3-21G/6-31G**	S_{XY}, %	L(0.001,10,X,Y)	L(0.1,10,X,Y)	L(0.01,0.1,X,Y)	L(0.001,0.01,X,Y)
test A	96	0.88	0.95	0.94	0.84
test B	97	0.91	0.96	0.93	0.87
test C	98	0.94	0.96	0.95	0.92
MEDLA/6-31G**	S_{XY}, %	L(0.001,10,X,Y)	L(0.1,10,X,Y)	L(0.01,0.1,X,Y)	L(0.001,0.01,X,Y)
test A	98	0.93	0.97	0.94	0.92
test B	98	0.93	0.96	0.93	0.91
test C	99	0.97	0.98	0.97	0.96

[a] Four ranges of density thresholds a and a' were used for the second similarity measure.

Table 2. Computed Similarities for MEDLA vs Direct *ab Initio* MIDCOs for Two Water Molecules Experiencing Two Types of Nonbonding Interactions for a Range of Distances

	% of nonbonding interaction	
distance,[a] Å	hydrogen bonds	O−O repulsion
1.5	98	94
2.0	98	98
2.5	98	99
3.0	98	98

[a] The distances listed are between the O atom in one water molecule and a H atom in the other (for the hydrogen bonds) and between the two O atoms (for the repulsive interactions).

reasonable approximations of formal molecular volume measures as well. The highest quantitative agreement between the 3-21G and 6-31G** results occurs within a high density range (10.0−0.1 au) of small overall volume; nevertheless, as we have previously seen in Figure 3, the 3-21G MIDCOs have major topological differences with the 6-31G** MIDCOs in this range, whereas the MEDLA MIDCOs have the same topology as the 6-31G** MIDCOs.

As an additional test of the method, the effect of changing the distance between the fragments involved in nonbonding interactions was studied for two water molecules as they were brought in close proximity to each other. Hydrogen bonds were studied by computing the similarity of the MEDLA *vs* standard *ab initio* density distributions as a function of the distance between the H atom in a water molecule from the O atom in another water molecule, over a range of 1.5−3.0 Å. The effects of repulsive interactions were also tested in the same manner, but this time the distance variation between the two O atoms was considered over the same range. In all cases, the geometries were fully optimized with the exception of the constrained distance. Table 2 shows the computed similarities of the MEDLA and direct *ab initio* distributions for the different distances. Note that all but one of the computed similarities were above 97% when using the similarity measure S_{XY} of Richards. For a severe repulsive interaction (where the O−O distance was 1.5 Å) the computed similarity dropped to 94%. Note, however, that such pathological internuclear distances and such strong repulsive interactions do not occur in any of the conformations of the molecules studied in this work. Based on these tests and the results of ref 12, we are satisfied that the MEDLA approach does indeed produce *ab initio* quality electron density distributions for the molecules considered in this work.

3. Electron Density for an α-Helix

The minimal computational time required for the MEDLA calculations allows for the exploration of the electron density distributions of ideal secondary structures of polypeptides such as the α-helix. Figure 6 shows MIDCOs for a polypeptide with 13 amino acid residues [the sequence of the amino acids is $(Gly)_2$-$(Ser)_{10}$-Gly]. This figure also displays another appealing feature of the MEDLA approach, that of being able to calculate the density for any significant fragment of the molecule. MIDCOs are shown for both the complete molecule and the backbone of the polypeptide. Two views are shown for each isodensity surface. The backbone MIDCOs are shown on the left side of the figure for each threshold value.

The views on the top clearly show the helical nature of the polypeptide. Here the N terminus is at the bottom, while the C terminus is at the top. The bottom views for the higher threshold value MIDCOs show a circular channel going down the center of the helix. Note that in both the complete molecule and the backbone, the holes are equivalent, indicating that the R groups arranged along the outside of the helical backbone have negligible influence on the electron density within the hole. The MIDCOs for the smaller threshold values of density would yield molecular volumes similar to those estimated experimentally.

From the infinitely many threshold values that can be selected for the MIDCOs, only two are chosen in this figure. These MIDCOs show both the skeletal features and the space filling characteristics of the molecule. Conventionally, the skeletal features of these peptides have been described by either a wireframe or ball and stick models, while the space filling characteristics are usually modeled by a system of interpenetrating spheres or formal, solvent accessible surfaces based on them. It is evident from Figure 6 that wireframe models or a set of spheres are unable to accurately model the shape and the actual topologies of the MIDCOs shown. Neither of the conventional models could adequately show the hydrogen bonding required for the stability of the helix. By contrast, hydrogen bonds are clearly manifested in a range of MIDCOs between the two threshold values shown in the figure. MIDCOs for threshold values where hydrogen bonds are clearly recognizable are shown for several of the peptides in this study.

The computational time required for the calculation of the entire electron density distribution for this helical peptide was <5 min on our workstation using the MEDLA approach. This speed is a remarkable feature of the method, considering that the MEDLA approach produces electron density distributions of *ab initio* quality at or near the 6-31G** level of basis for a molecule that has a formula of $C_{36}O_{24}N_{13}H_{63}$ with 137 atoms in total. At present, a direct *ab initio* calculation for this molecule is out of the reach of all but supercomputers and even then the required CPU computational time would be several days.

4. Electron Densities for Bioactive Peptides

In this section, the electron densities of several bioactive peptides and two proteins will be analyzed. The first two molecules considered are peptides that are found in the brain, two enkephalins, our bodies natural analgesics. These two peptides each contain five amino acid residues, the first four of which are identical in the two peptides, with only the amino acid at the C terminus being different. These peptides are

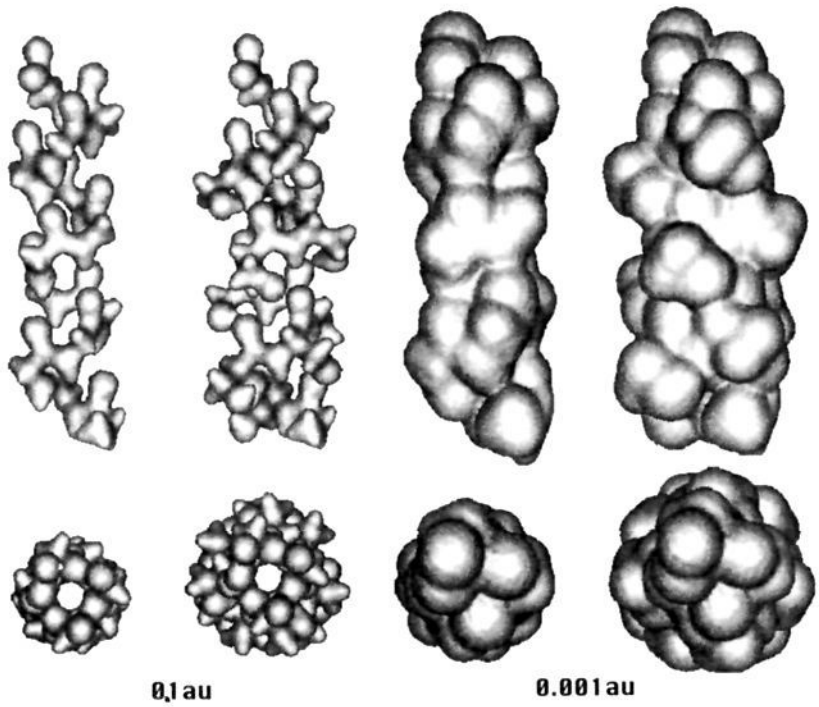
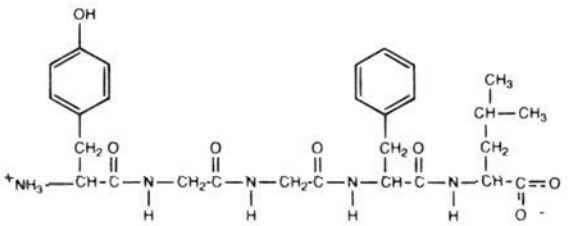


Figure 6. MIDCOs for both the backbone and the complete molecule of a helical peptide comprised of 13 amino acid residues, $(Gly)_2$-$(Ser)_{10}$-Gly. Two views of each MIDCO are shown. The bottom view is perpendicular to the top view.



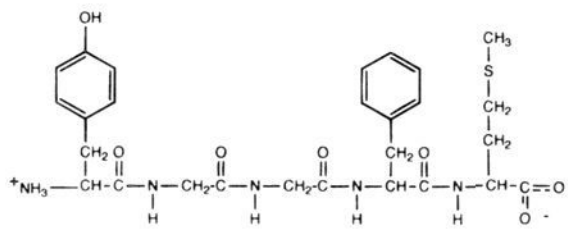


Figure 7. The structural formuli for the two enkephalin peptides. The two peptides are identical with the exception of the C-terminal amino acid. The computed MEDLA electron densities of these two pentapeptides, leuenkephalin (Tyr-Gly-Gly-Phe-Leu) and metenkephalin (Tyr-Gly-Gly-Phe-Met), contain 34 and 35 MEDLA fragments, respectively.

referred to as leuenkephalin (Tyr-Gly-Gly-Phe-Leu) and metenkephalin (Tyr-Gly-Gly-Phe-Met) where the first three letters in the name from the symbol for the unique *differing* amino acid in the two molecules. Figure 7 shows the structure for the two peptides as well as the number of molecular fragments needed for the MEDLA calculation. To find a low-energy conformer, quenched molecular dynamics were carried out for the two peptides for 10 ps at 600 °C using the BIOGRAF software package. The resulting density distributions for these conformers were then calculated using the MEDLA program. The computational time required for the MEDLA calculations was approximately 2 min for each peptide.

Figures 8 and 9 show two views of each of three MIDCOs for leuenkephalin and metenkephalin, respectively. The top views allow the C and N termini to be clearly distinguished, while the bottom views are perpendicular to the top views. All the features of these two peptides cannot be seen in these two views; evidently, several views are required for a detailed, visual shape analysis. This shows the disadvantages of trying to analyze molecular shape using only visualization, especially if restricted to a few 2D projections. For detailed, reliable, and reproducible shape analysis, these MIDCOs should be studied using nonsubjective, nonvisual, computer-based techniques, based on the shape group methods[2-4,14,15] or adaptations of alternative techniques.[18-20]

Figures 8 and 9 demonstrate the large effect the C terminal amino acid in the peptides has on the secondary structure. The conformer for leuenkephalin is fairly extended with the C and N termini being a large distance from each other. However, in Figure 9, the two termini of the conformer of metenkephalin are very close to each other, and in fact are interacting with each other. This folding back of the last residue is likely due to the interaction between the phenyl ring from the Phe residue with the sulfur from the Met residue in metenkephalin. This interaction is easily recognized in the 0.001 au MIDCO shown in Figure 9 and was also analyzed in more detail in section 2.

The next peptide studied was bradykinin with nine amino acid residues. The sequence for the amino acids is Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg. A low-energy conformer was found for bradykinin using the same procedure as for the enkephalins. The molecule contains 59 of our MEDLA fragments and the resulting computational time required to calculate the electron density distribution for this conformer was <5 min on our workstation. Three MEDLA MIDCOs of the resulting electron distribution are shown in Figure 10. Again, two views of each MIDCO are displayed.

(18) Carbó, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185.
(19) Hodgkin, E. E.; Richards, W. G. *J. Chem. Soc., Chem. Commun.* **1986**, 1342.
(20) Leicester, S.; Bywater, R.; Finney, J. L. *J. Mol. Graphics* **1988**, *6*, 104.

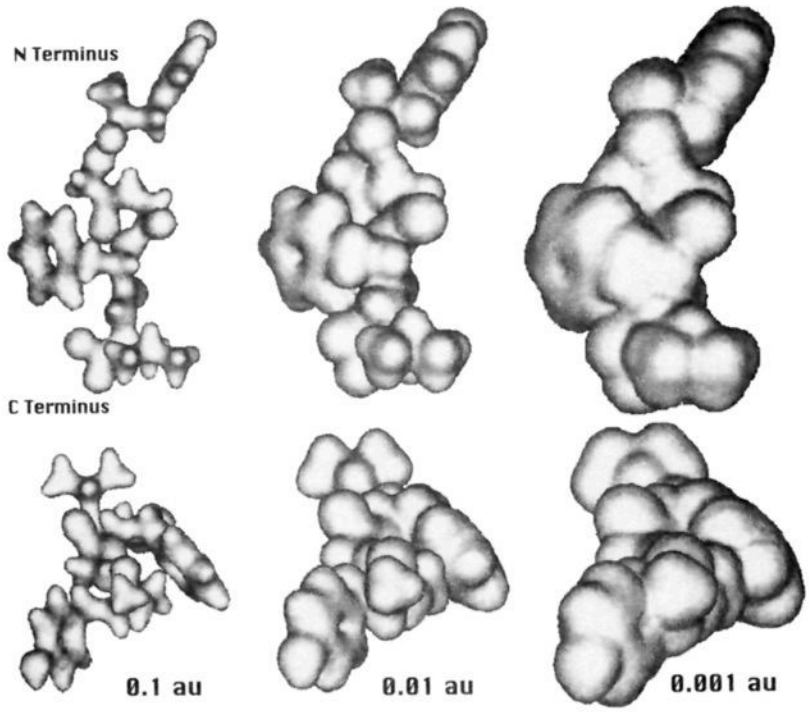


Figure 8. MIDCOs for a low-energy conformer of leuenkephalin, Tyr-Gly-Gly-Phe-Leu. The terminal ends of the peptide are indicated in the figure. Two perpendicular views are shown for each MIDCO.

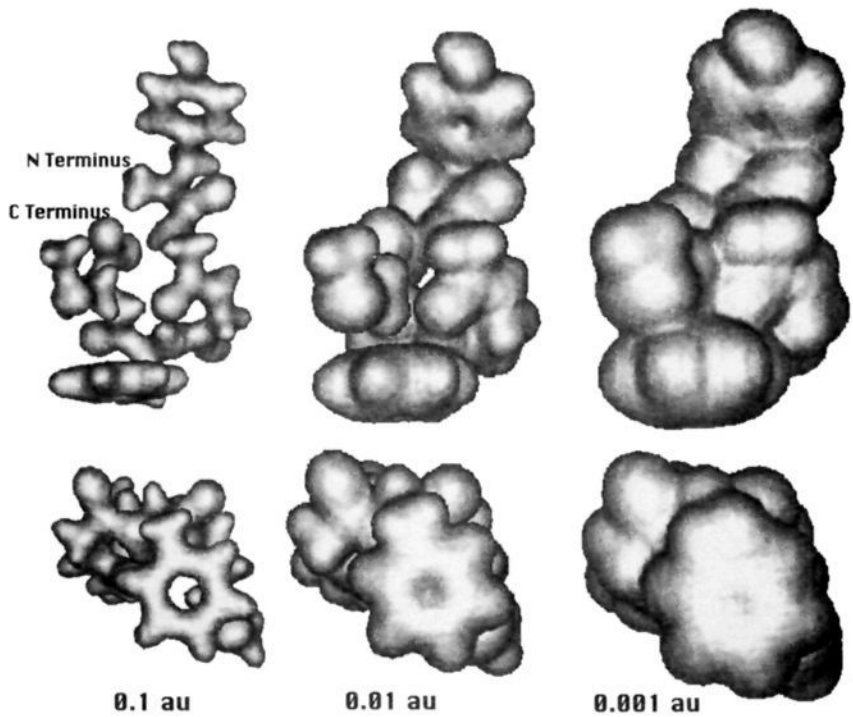


Figure 9. MIDCOs for a low-energy conformer of metenkephalin, Tyr-Gly-Gly-Phe-Met. The terminal ends of the peptide are indicated in the figure. Two perpendicular views are shown for each MIDCO.

The conformation shown for bradykinin is fairly extended. The top views in Figure 10 show the backbone of the peptide with the most clarity. The two proline residues that are beside each other do cause some "kinks" in the backbone of the molecule. The low-density MIDCO shows features which we believe would be poorly represented by fused sphere Van der Waals surfaces or solvent accessible surfaces. In the given conformation of the molecule, there are interactions between the first proline residue and the second phenylalanine residue as well as between the same proline and the serine residue. The deformation in the density caused by these interactions cannot be modeled properly by overlapping spheres centered on the atoms, such as those in fused sphere Van der Waals surfaces, but is clearly shown in Figure 10. Details of all essential topological features are well represented by MEDLA MIDCOs, whereas fused sphere Van der Waals models are inadequate for precise analysis of molecular shape.

The next molecule to be considered is the protein crambin. This protein has 46 amino acid residues and 3 disulfide linkages with 656 atoms in total. The conformer shown is a low-energy

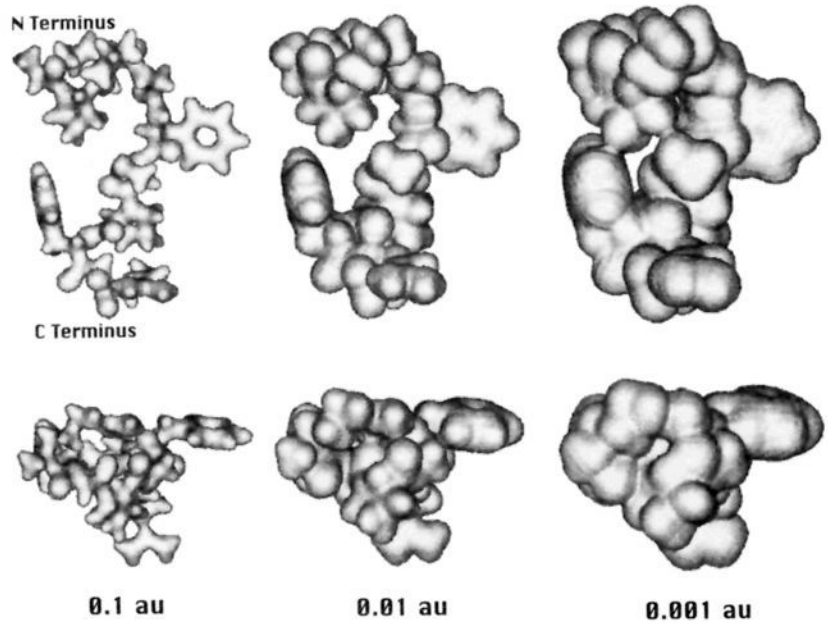


Figure 10. MIDCOs for a low-energy conformer of bradykinin, Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg. The two termini of the peptide are specified in the figure. Two perpendicular views are shown for each MIDCO. There are 59 MEDLA fragments used in the construction of the electron density of this molecule.

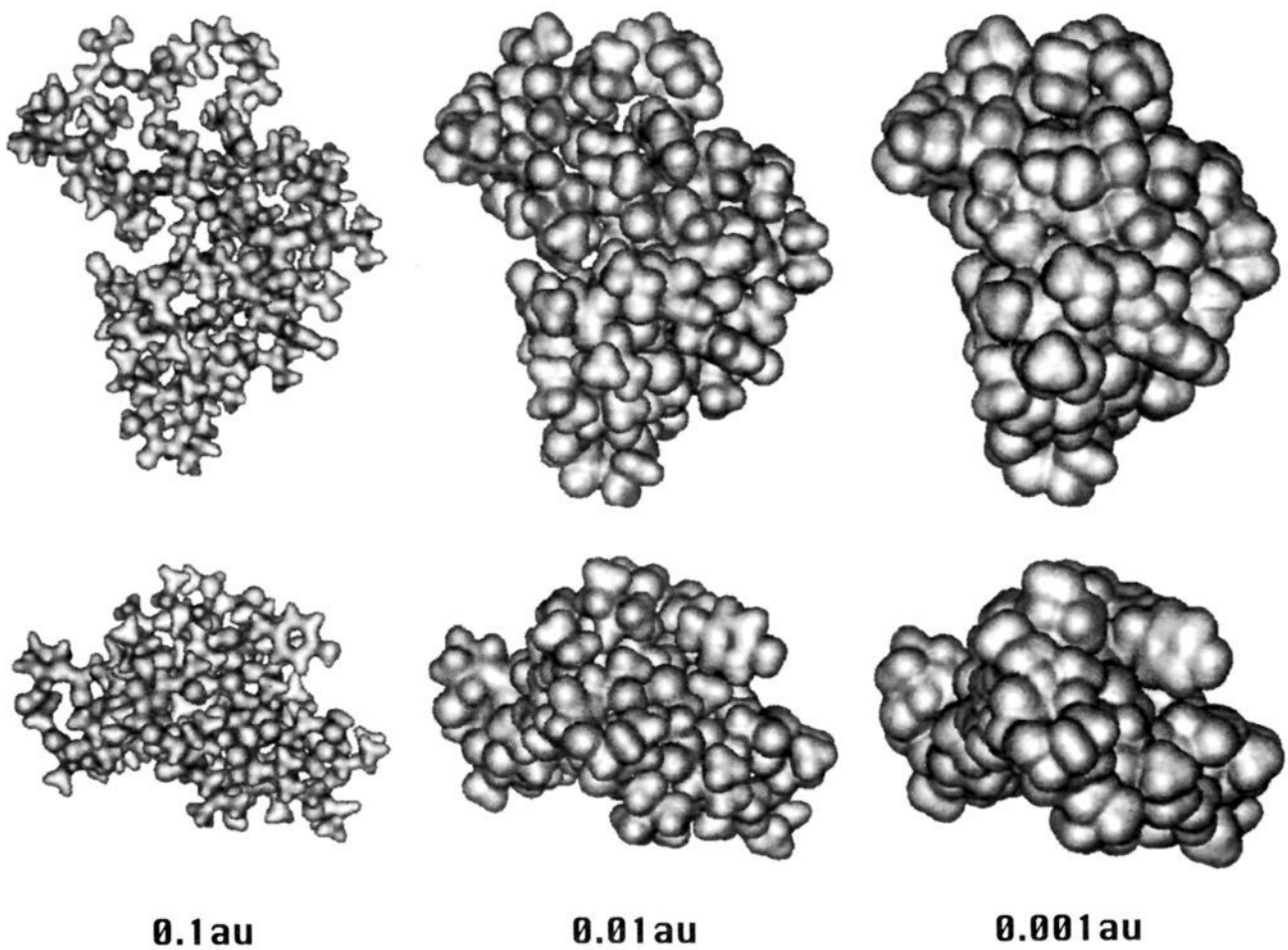


Figure 11. Results of *ab initio* quality electron density calculations for a low-energy conformer of the protein crambin of 656 atoms in 46 amino acid residues. Two perpendicular views are shown for each MIDCO.

conformer; however, no effort was made to find the global minimum. The calculation of the electron density distribution using the MEDLA program required 11 min of CPU time on our workstation. At present, the electron density of this molecule could not be calculated using conventional *ab initio* programs at the 6-31G** level of basis due to memory

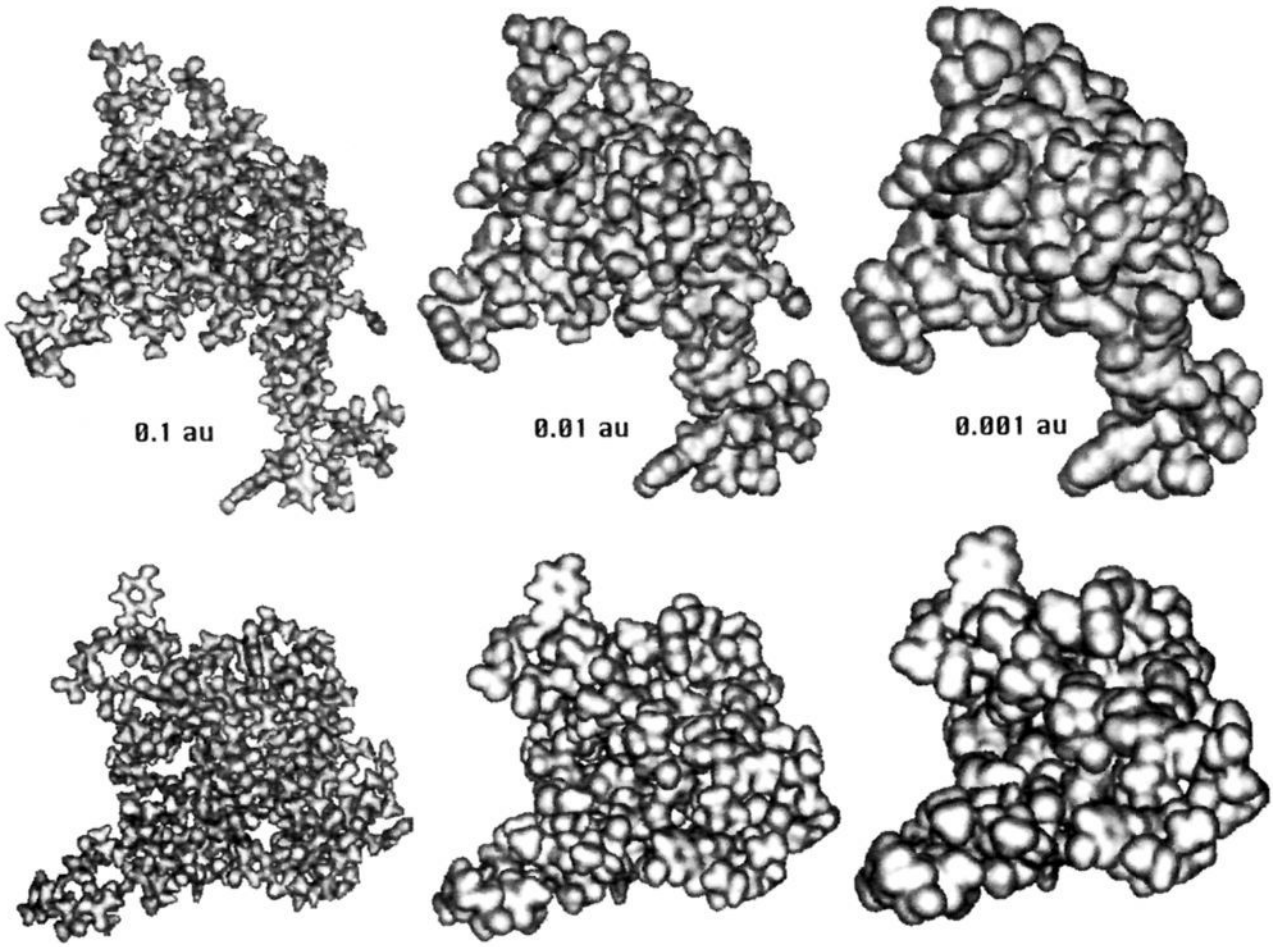


Figure 12. Results of *ab initio* quality electron density calculations for a low-energy conformer of the gene 5 protein from bacteriophage M13. This protein has 87 amino acid residues and a total of 1384 atoms. Two perpendicular views are shown for each MIDCO.

constraints. Even if these constraints were to be circumvented, we estimate that a direct calculation would take more than a decade of CPU time on a CRAY supercomputer. Before the introduction of the MEDLA method,[12,16] the calculation of an *ab initio* quality electron density distribution for a protein or any nonperiodic molecule of this size was not feasible.

Selected MIDCOs from the computed charge distribution of the protein crambin are presented in Figure 11. This conformation of the protein has an elongated globular shape, as shown by the top views in the figure. Many fine details of the shape of the protein can be explored visually using the displays of MEDLA MIDCOs. The gradual buildup of electron density can be followed easily by considering the sequence of MIDCOs for different thresholds. The merging of electronic density clouds between parts of the protein not linked directly by formal bonds is an important feature not well represented by earlier models. These mergers start to occur at about the same density threshold at many locations within the protein. If this trend is found general for favored conformations of other proteins, this could give a tool for partially justifying favored mutual side chain arrangements. The actual space filling aspects of internal domains of the globular protein are particularly well represented by a sequence of MIDCOs for different density thresholds.

Our last example demonstrates the range of the new possibilities for electron density modeling of proteins: MEDLA 6-31G** electron density has been calculated for the gene 5 protein (g5P) from bacteriophage M13, a molecule of >1000 atoms. This protein has 87 amino acid residues[21] and a total of 1384 atoms. Although the MEDLA technique itself has no inherent size limitation, this molecule is presently near the limit of what the visualization software of our workstation can handle. We estimate that a direct *ab initio* calculation of the same quality would take more than a century on a CRAY supercomputer. The MEDLA calculation took 21 min on our workstation.

Figure 12 shows two views of the gene 5 protein MIDCOs for three threshold values. Both the N and C termini are located in the upper left corner of the first view in Figure 12. Fine details of the fuzzy electronic charge cloud are clearly distinguishable and can be viewed at a variety of density thresholds. By abandoning the constraint of conventional protein models of a single surface of specific boundary, the MEDLA technique describes realistically the gradual, continuously emerging space filling aspects of the fuzzy, interpenetrating, three-dimensional electron densities, as the MIDCO thresholds are gradually decreased. The results provide detailed information on bonding characteristics as well as nonbonded interactions.

Whereas the displays of MEDLA MIDCOs can provide a wealth of important shape information and valuable clues to protein behavior, it is clear that due to the congestion of detailed shape features seen in the MIDCOs an *exhaustive visual shape analysis* of these surfaces is impractical for most purposes.

(21) Coleman, J. E.; Williams, K. R.; King, G. C.; Prigodich, R. V.; Shamoo, Y.; Konigsberg, W. H. In *Protein Engineering*; Oxender, D. L., Fox, C. F., Eds.; Alan R. Liss, Inc.: New York, 1987.

However, earlier nonvisual methods for molecular surface analysis and shape comparisons[2-4,14,15,18-20] can be adopted to MEDLA MIDCOs, and for objects of this high level of complexity, *automated, nonvisual, computer-based shape analysis techniques* offer a much more reliable approach.

5. Closing Remarks

We have presented the results of a novel application of the Molecular Electron Density Lego Assembler method to large molecules of biological interest. The MEDLA approach provides the first method capable of calculating realistic electron densities, faithful representations of fuzzy molecular bodies, and a range of molecular surfaces for proteins and large polypeptides.

Future work will concentrate on developing new and modifying present techniques for the automated, nonvisual shape analysis of the MEDLA molecular bodies and molecular (MIDCO) surfaces. Of special concern is developing a new method for the rapid computation of shape similarity measures for a series of large molecules based on their MEDLA electron densities. Energy relations based on density fragments will be used for large-scale structure optimization problems. A related, important area of MEDLA development is dynamic modeling of conformational changes, such as protein folding processes, and the study of molecular interactions. The MEDLA approach will also be adapted to calculate molecular electrostatic potentials from the electron densities, replacing earlier point charge models with realistic charge densities. In addition, we play to investigate the MEDLA technique using alternative electron density fragment databases, for example, those computed with superior basis sets and correlated wave functions, or based on density functional methods.

Acknowledgment. This work was supported by the Natural Sciences and Engineering Research Council of Canada.