A new computational microscope for molecules: High resolution MEDLA images of taxol and HIV-1 protease, using additive electron density fragmentation principles and fuzzy set methods

P. Duane Walker and Paul G. Mezey

Mathematical Chemistry Research Unit, Department of Chemistry and Department of Mathematics and Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, SK, Canada S7N 5C9

Received 16 November 1994

Based on the molecular electron density lego assembler (MEDLA) method, a "computational microscope" was developed that generates accurate images of bodies of large molecules at a resolution far exceeding current experimental techniques. The MEDLA "microscope" can be "tuned" to display the high electron density regions of formal chemical bonds; or to show the low density regions of hydrogen bonds and secondary interactions, or to display local shape requirements important in molecular recognition. The power of the method is illustrated by examples of detailed images of taxol, an important anti-cancer agent, and HIV-1 protease, a protein of 1564 atoms. A mathematical framework of the approach, based on fuzzy sets, and the fundamentals of several additional applications of the additive, fuzzy fragmentation principle are presented.

1. Introduction

Understanding chemistry, biochemistry, biotechnology, molecular recognition, drug-receptor interactions, and catalytic processes relies on the knowledge of molecular shapes [1]. The resolution of current X-ray experiments is suitable to determine the location of nuclei in proteins, but is not sufficient yet for a detailed shape analysis of the body of electron density at the chemically important, low density regions of large molecules such as proteins. No current experimental technique and, until the recent introduction of the MEDLA method [2–5], no computational modeling technique could generate reliable, high resolution images of large molecules, and show what they really are: bodies formed by fuzzy electronic charge density clouds of intricate shape features. Fused sphere Van der Waals (VDW) surfaces, and the improved solvent accessible surfaces mimic only the rough features of electron density. Until now, chemists have studied large molecules without

being able to see in reliable detail how these molecules look. Here we report the first accurate, detailed, high resolution images for a molecule beyond the 1500 atom limit, as well as some of the fundamental mathematical principles relevant to the current implementation and additional applications of the approach.

The new *computational microscope*, based on the *molecular electron density lego assembler* method, in short, the MEDLA method [2–5], is suitable to generate and display highly detailed images of large molecular bodies at unprecedented resolution. In reality, molecular bodies are fuzzy electronic charge densities, that until now could be displayed accurately only for small molecules. The new MEDLA computational microscope, developed for macromolecules, can be tuned to view any of the high or low density ranges. If tuned to high density, the pattern of bonding and details of nuclear neighborhoods are displayed, while tuning to intermediate densities reveals space filling aspects, hydrogen bonding, and secondary interactions. At low densities, the space requirements, size and shape features relevant to molecular similarity and recognition are shown. All earlier models, such as wire frame and fused sphere models, fail to represent many of the details of molecular shapes revealed by the MEDLA images. The MEDLA computational microscope improves our "vision" considerably in the microscopic size range of nature; now we can view the molecular world in great detail, including both small and large molecules.

The MEDLA computational microscope is based on the fuzzy electron density fragment additivity principle [2]. An electron density fragment data bank has been generated, based on accurate, high quality *ab initio* quantum chemical calculations for small molecules, and the application of the electron density fragmentation principle [2-4]. *Ab initio* quality electron densities can be constructed for large molecules, for any nuclear arrangement, using experimental or theoretically determined nuclear coordinates, or distorted arrangements assumed to occur along reaction paths or in protein folding processes. The "fuzzy" density fragments also account for the inter-fragment interactions occurring within their molecular neighborhoods. These fuzzy fragments are arranged and combined according to an interpenetration pattern based on the additivity principle [2]. The additive, fuzzy fragmentation-density construction method is *exact* if a molecule is reconstructed from its fragments, and has been shown to be highly accurate when constructing other molecules [2,4], faithfully reproducing the shapes of molecular electron densities [2], including those with hydrogen bonds and nonbonding interactions as calculated at the standard 6-31G** *ab initio* level [4]. In sections 3–9 several aspects of the fragmentation scheme will be discussed, after the power of the method is demonstrated by examples and a summary of test results is given in sections 2 and 3. The MEDLA method can be employed in molecular similarity studies [6], extending the scope of earlier techniques proposed for the local and global shape analysis of molecular fragments [7] and complete molecules [1].

2. High resolution images of taxol and HIV-1 protease

The protein molecule of this study required 21 fragment types, whereas the taxol molecule required several additional fragments from our MEDLA density fragment databank. Each fuzzy density fragment has been previously obtained from a 6-31G** *ab initio* calculation for a smaller molecule, artificially distorted to match the nuclear geometry and local surroundings of the fragment in the target molecule (for more technical detail see later sections of this report). Hence, each fragment type is stored in several versions in the databank, for a range of several possible local nuclear arrangements. By selecting in each case the fragment with matching or nearly matching nuclear geometry, high accuracy can be achieved. Typical fragments are the methyl group, CH_2, NH_2, and the carboxyl group; an account of various tests on hydrogen bonds and other interactions are described elsewhere [4].

Our first example, the taxol molecule of 113 atoms, is an important naturally occurring anti-cancer agent, synthesized recently [8,9]. In fig. 1, the stereochemical structure diagram of taxol, along with a MEDLA image tuned to high density showing the bonding pattern, are displayed. The structure was optimized using the BIOGRAF program [10]. Figure 2 shows four MEDLA images of taxol from the same perspective, "tuned" to electron densities 0.2 a.u. (atomic unit), 0.1, 0.01, and 0.001 a.u., respectively. These are images of 6-31G** *ab initio* quality molecular isodensity contours (MIDCOs). Fine details of the bonding pattern, secondary interactions, the mutual interpenetrations of charge clouds between groups not formally linked by bonds, and the local shapes and space requirements of various functional groups are clearly shown. Conventional space filling models are unable to describe all of these features. The bonding regions between nuclei, nonbonded interactions, and hydrogen bonds, are poorly represented by fused sphere and similar simple models, whereas the computed electron densities provide a far more accurate description. Furthermore, the shapes of aromatic rings at various density thresholds are not suitable for simple, fused-sphere or similar representations: as the π-electron density contribution becomes prominent at lower density thresholds, the isodensity contours show a remarkable "swelling" perpendicular to the approximate plane of the nuclei, whereas the contour changes very little along directions within the plane.

For a molecule of this size, a direct *ab initio* calculation of similar quality would take many hours of CPU time on a Cray supercomputer; by contrast, the MEDLA computations require only 3 minutes of computer time on our KPC Titan 3000 workstation.

The HIV-1 protease monomer, a protein of 1564 atoms in 99 amino acid residues, is by far the largest molecule to date for which *ab initio* quality electron density calculation has ever been attempted. At present, a conventional *ab initio* calculation for proteins is *impossible* due to insurmountable RAM memory requirements. Were these memory problems circumvented, the calculation would still be impossible in practice, as we estimate that a direct, traditional *ab initio* com-

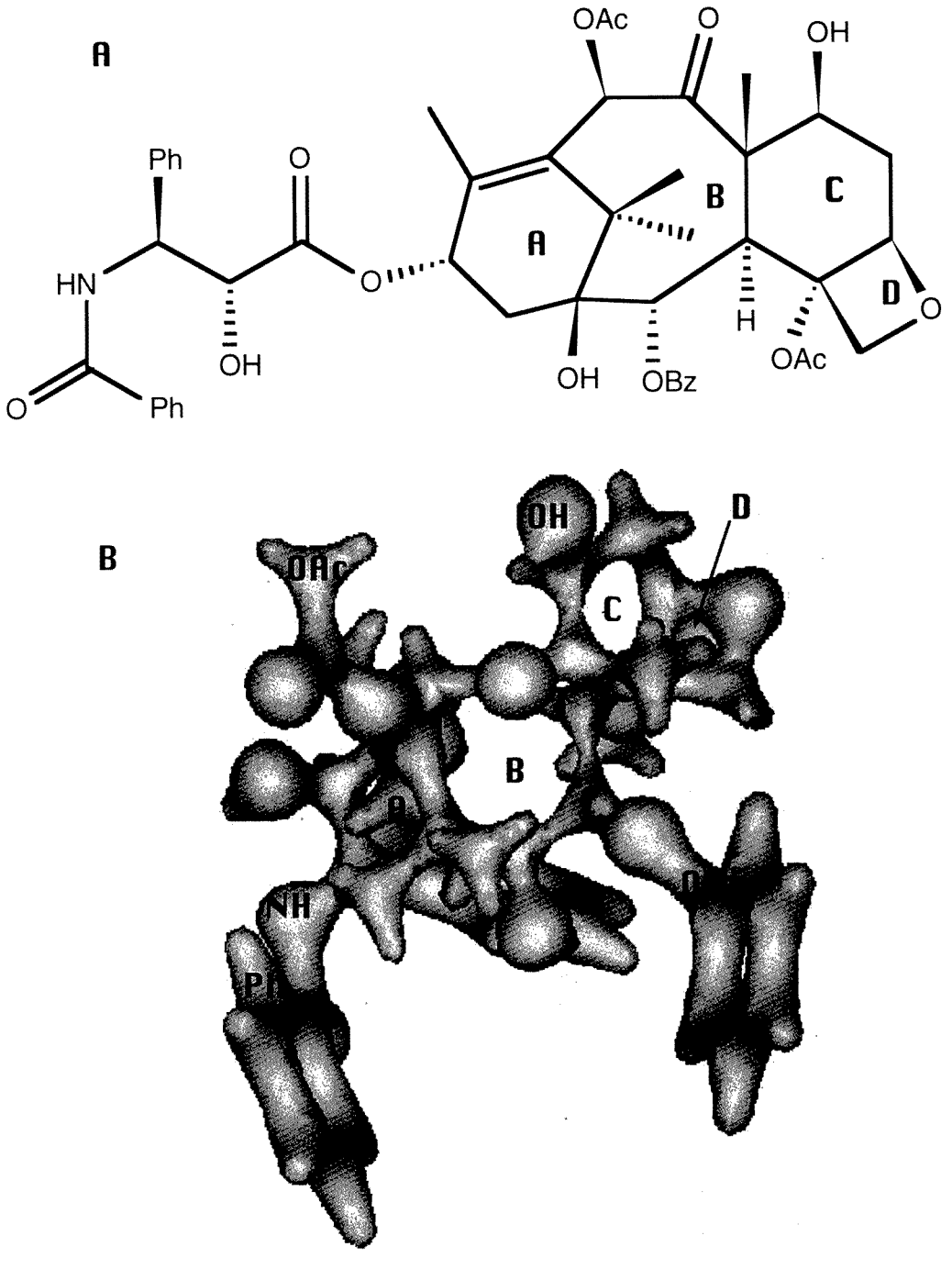

Fig. 1. Stereochemical structure formula of taxol (A) and a MEDLA computational microscope image (B) tuned to high density (0.2 atomic unit), showing the bonding pattern.

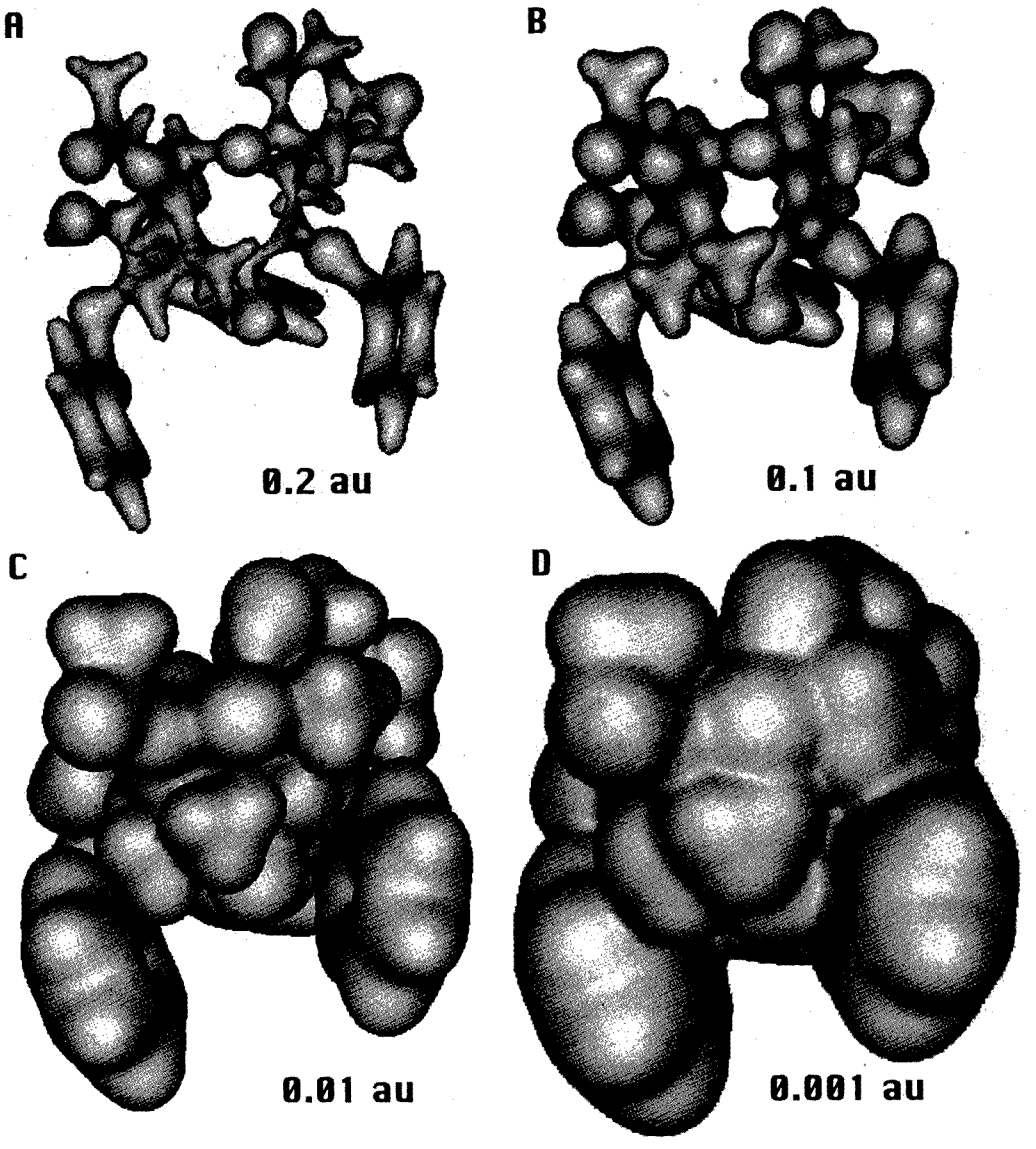

Fig. 2. Four MEDLA computational microscope images (A, B, C, and D) of the taxol molecule, tuned to density thresholds 0.2, 0.1, 0.01, and 0.001 a.u., respectively, shown from the perspective specified in fig. 1 (a.u. = atomic unit).

putation for HIV-1 protease would take *three centuries* of CPU time on a Cray supercomputer. By contrast, the MEDLA computation on our workstation took only 35 minutes, which represents more than a millionfold improvement over the speed of a traditional *ab initio* method of similar, 6-31G** accuracy.

Figure 3 shows a comparison and overview of the wire-frame model of HIV-1 protease, based on the crystallographic coordinates [11], and three, 6-31G** *ab*

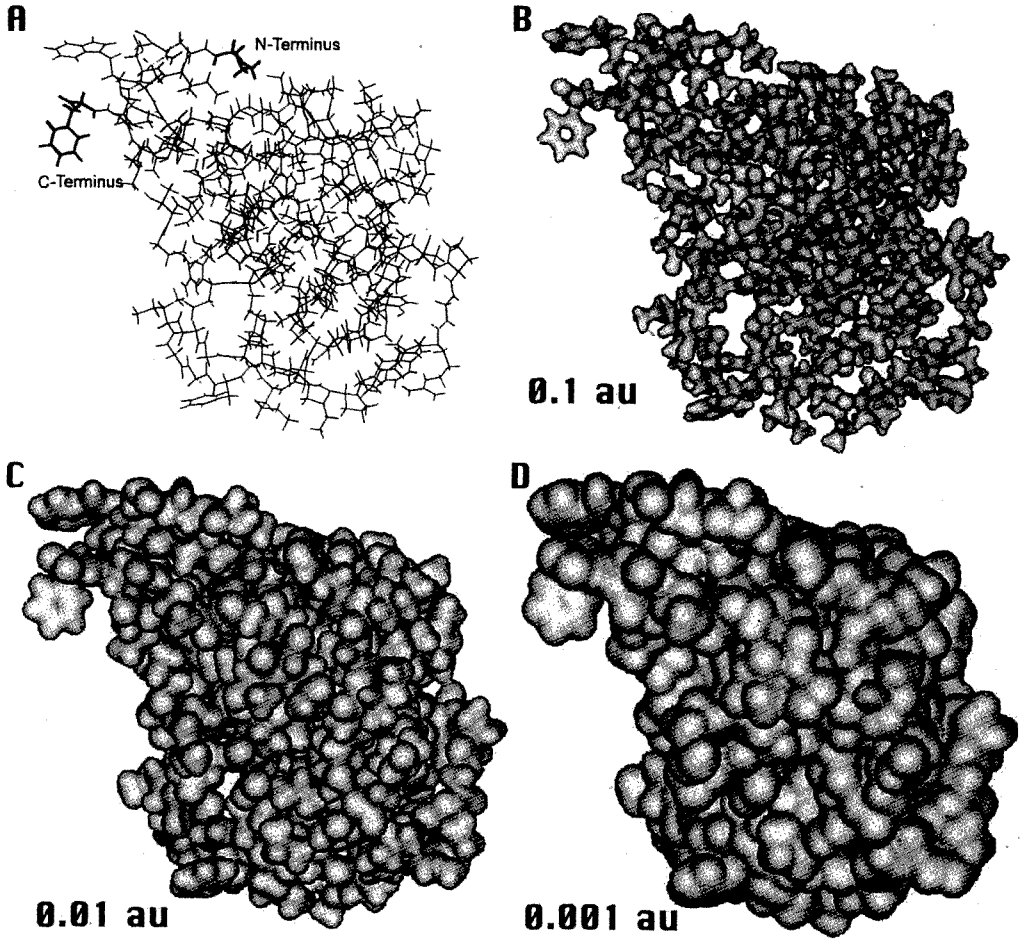

Fig. 3. An overview of the wire-frame model (A) and three MEDLA computational microscope images (B, C, and D), tuned to density thresholds 0.1, 0.01, and 0.001 a.u., respectively, are shown for the HIV-1 protease protein. The monomer contains 1564-atoms in 99 residues; the structure shown is based on crystallographic coordinates.

initio quality MEDLA images from the same perspective, "tuned" to electron densities 0.1, 0.01, and 0.001 a.u., respectively, whereas figs. 4, 5, and 6 show details of the three MEDLA images. The new MEDLA computational microscope generates the entire electronic density for the protein, and can be tuned to any other density thresholds, different from 0.1, 0.01, and 0.001 a.u. There is a wealth of shape and size features revealed at high, intermediate, and low electron densities. Hydrogen bonding, and interactions between molecular regions not formally linked by chemical bonds are also easily recognizable. Conventional wire frame models, ball–and-stick models, and fused-sphere Van der Waals models fail to represent these features. No molecular images of comparable accuracy and resolution have been possible for proteins before the introduction of the MEDLA technique.

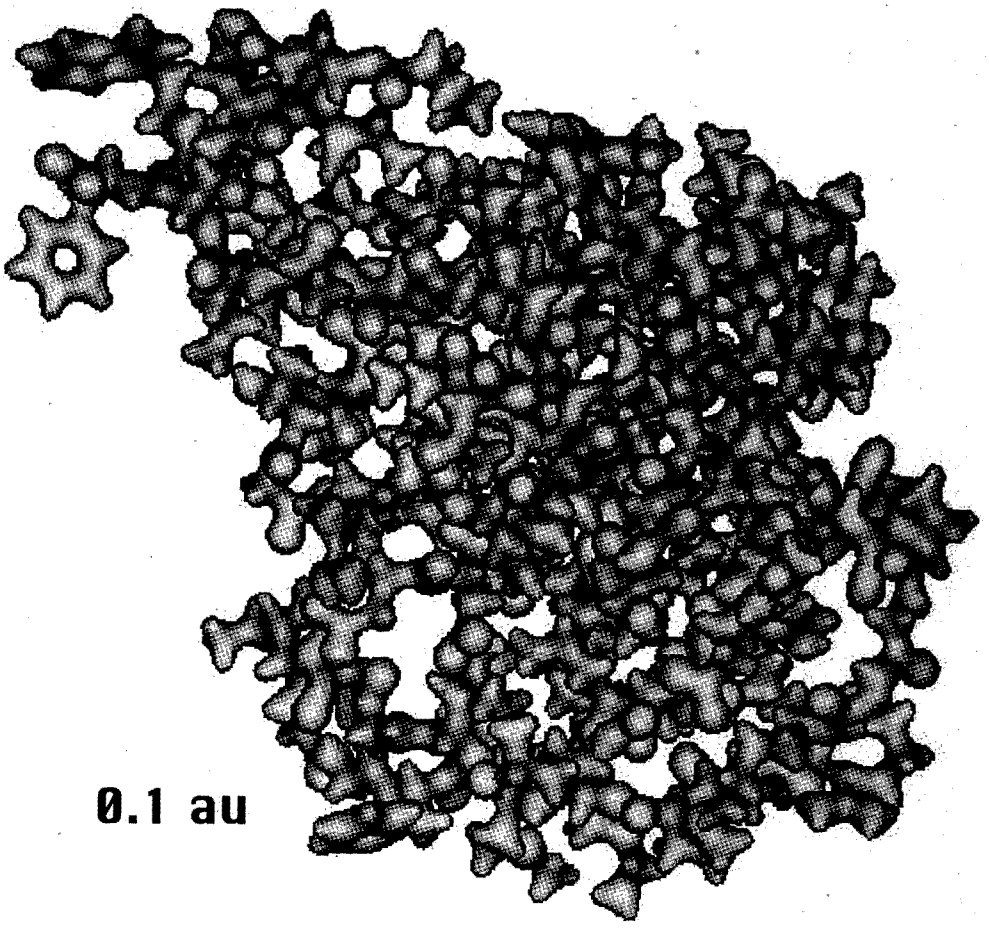

Fig. 4. Detailed MEDLA computational microscope image of the HIV-1 protease protein at the density threshold of 0.1 a.u.

3. Summary of earlier test results

The MEDLA method, in its simplest form, involves an additive, fuzzy electron density fragmentation scheme analogous to a Mulliken population analysis without integration (see refs. [2–5], and the next section). Even in this simplest form, the MEDLA method has been shown to generate *ab initio* quality electron densities [2,4]. Detailed comparisons of electron densities computed by traditional *ab initio* SCF technique using 3-21G and 6-31G** basis sets, and by the MEDLA method have shown that the *MEDLA result is invariably of better quality* than the standard 3-21G *ab initio* result, and virtually indistinguishable from the standard *ab initio* 6-31G** basis set result. Specific tests included:

(a) detailed comparisons of electron densities obtained by traditional *ab initio*

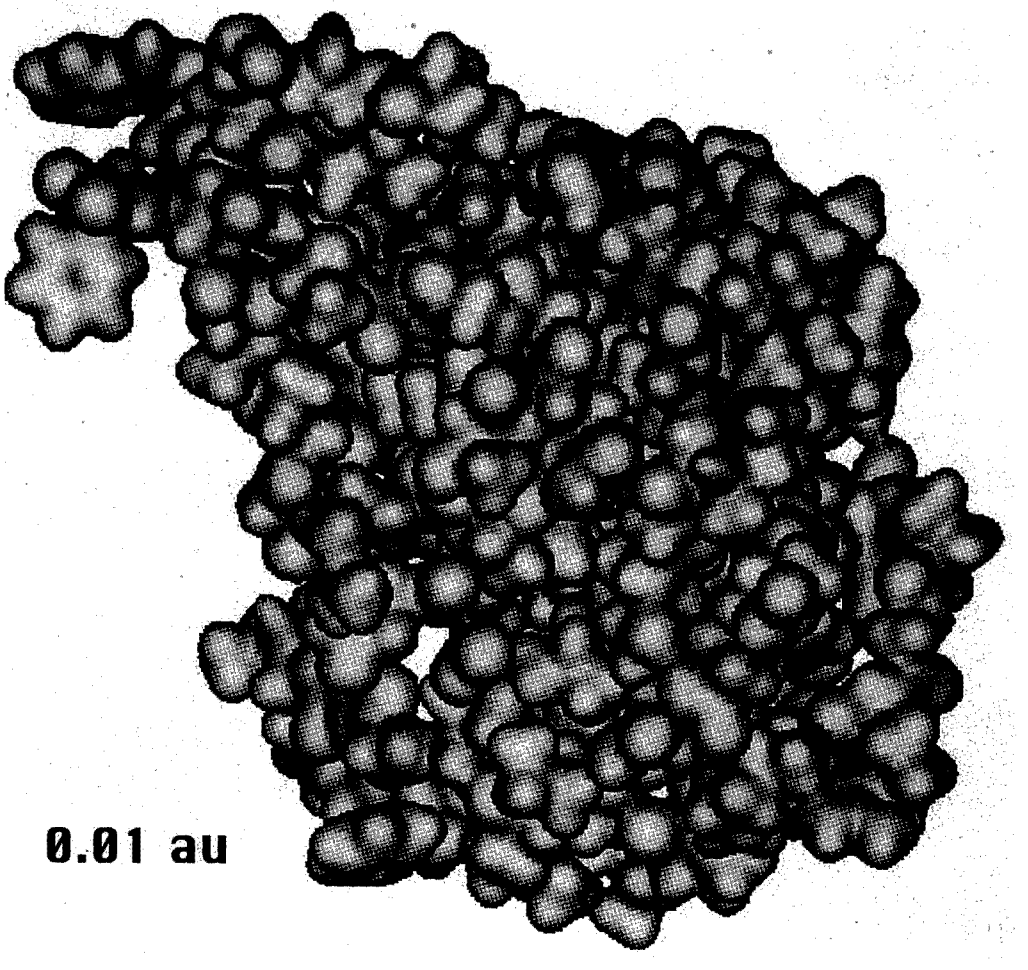

Fig. 5. Detailed MEDLA computational microscope image of the HIV-1 protease protein at the density threshold of 0.01 a.u.

SCF calculations using 3-21G and 6-31G** bases, as well as MEDLA computations for β-alanine [2],
(b) test of a prototype peptide system: traditional *ab initio* SCF calculations using 3-21G and 6-31G** bases, as well as MEDLA computations for glycyl-alanine [4],
(c) test of H-bonding in a helical tetrapeptide, using traditional *ab initio* SCF 3-21G and 6-31G** basis set calculations, as well as the MEDLA method [4],
(d) test of a nonbonded interaction between a sulfur atom and a phenyl ring in a molecular fragment from the pentapeptide metenkephalin, using standard *ab initio* SCF 3-21G and 6-31G** basis set calculations, as well as the MEDLA technique [4].

Since the MEDLA performs consistently better than standard *ab initio* SCF 3-

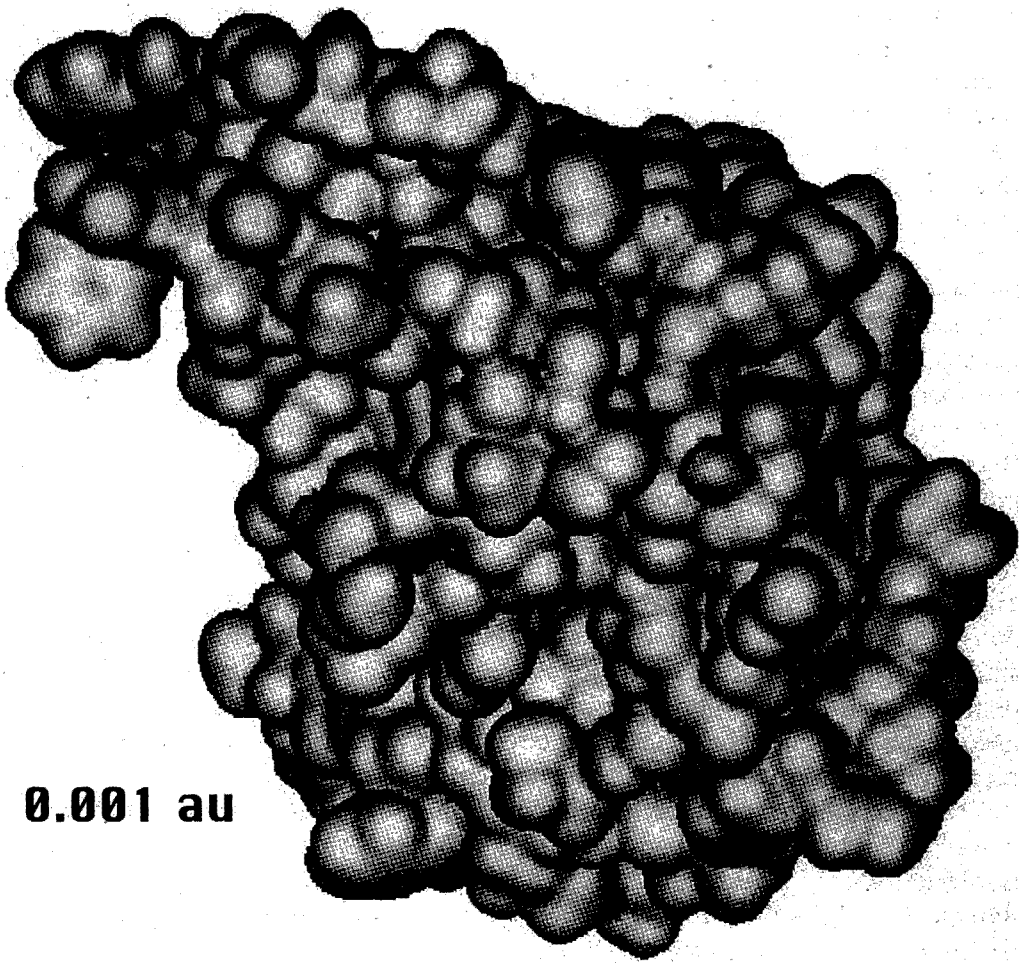

Fig. 6. Detailed MEDLA computational microscope image of the HIV-1 protease protein at the density threshold of 0.001 a.u.

21G basis computations, the claim of "*ab initio* quality" is justified. These tests provided further insight: according to the $a = 0.007$ a.u. (atomic unit) threshold density contours for the hydrogen bonded system (test c), the traditional 3-21G result does not show the hydrogen bond that is indicated by both the traditional 6-31G** and MEDLA techniques, whereas according to the $a = 0.003$ a.u. contours for the S–Phe interaction (test d), the traditional 3-21G result indicates a bridging of the local density contours where still a gap exists according to both the traditional 6-31G** and MEDLA computations. Hence, MEDLA outperforms the standard 3-21G computations in an apparently unbiased way: indicating a feature where the standard 6-31G** result indicates it, and shows the lack of a feature where it is lacking according to the standard 6-31G** result, used as a benchmark.

4. Comparison of additive electron density fragmentation schemes

Electron density fragmentation schemes can be divided into two classes: those with boundaries and those without. Among schemes involving boundaries, the "atoms in molecules" method of Bader [12,13] has the most physical justification: this method generates molecular fragments based on the zero flux surfaces of the gradient of the electron density. Consequently, these fragments have boundary surfaces of various shapes. When such fragments are combined to reconstruct the molecule of their origin, the reconstruction is exact. However, when such density fragments from different molecules are combined to form an approximate electron density of a new molecule, various degrees of discontinuities occur at the boundaries, especially at low densities, where significant gaps can be found [14]. The general problem of density fragments with boundaries is illustrated in fig. 7, where the fragmentation schemes $AB \rightarrow A + B$ and $CD \rightarrow C + D$, as well as the approximate construction of a new molecule, $A + D \rightarrow AD$ are shown schematically. Evidently, the local surroundings of these fragments are different in different molecules, implying that the fragment boundaries of A and D do not precisely align. In general, there are errors of two types: density doubling (100% error) or density gaps (domains of zero density, 100% error). These errors are significant, since they are found at the locations where chemical bonding occurs. This high degree of local incompatibility between the transferred fragments is due to the presence of non-matching boundaries, occurring between the three-dimensional fragments placed within the new molecular system.

The MEDLA technique differs fundamentally from fragmentation approaches involving boundaries. The bodies of single, isolated atoms and molecules are fuzzy, borderless charge clouds, and the same applies for fragments within the MEDLA scheme. The MEDLA technique is based on an additive, fuzzy electron density fragmentation scheme, using boundaryless fragments, hence this scheme avoids the local accumulation of errors of schemes based on boundaries.

The fuzzy fragmentation principle of the MEDLA method [2–4] is illustrated in fig. 8, where the fragmentation schemes $AB \rightarrow A + B$, $CD \rightarrow C + D$, and the approximate construction of the electronic density of a new molecule, $A + D \rightarrow AD$, are shown. Each fuzzy MEDLA electron density fragment (A, B, C, and D) of the parent molecules (AB and CD in the example) is generated by specifying its "share" of the fuzzy molecular charge cloud, assigned to the subset of the nuclei of the fragment. Hence, there is no geometrical division of the electron density of the parent molecule into parts with boundaries; the fragment densities A, B, C, and D of the example are analogous to the fuzzy, boundaryless electron densities of complete molecules. These fuzzy electron density clouds of the fragments are formally "pulled out" from the molecular cloud. When these fuzzy fragments are used to construct the new molecule AD, the mutual interpenetration of the fuzzy fragments A and D prevents any local accumulation of error; there is neither density doubling nor density gap.

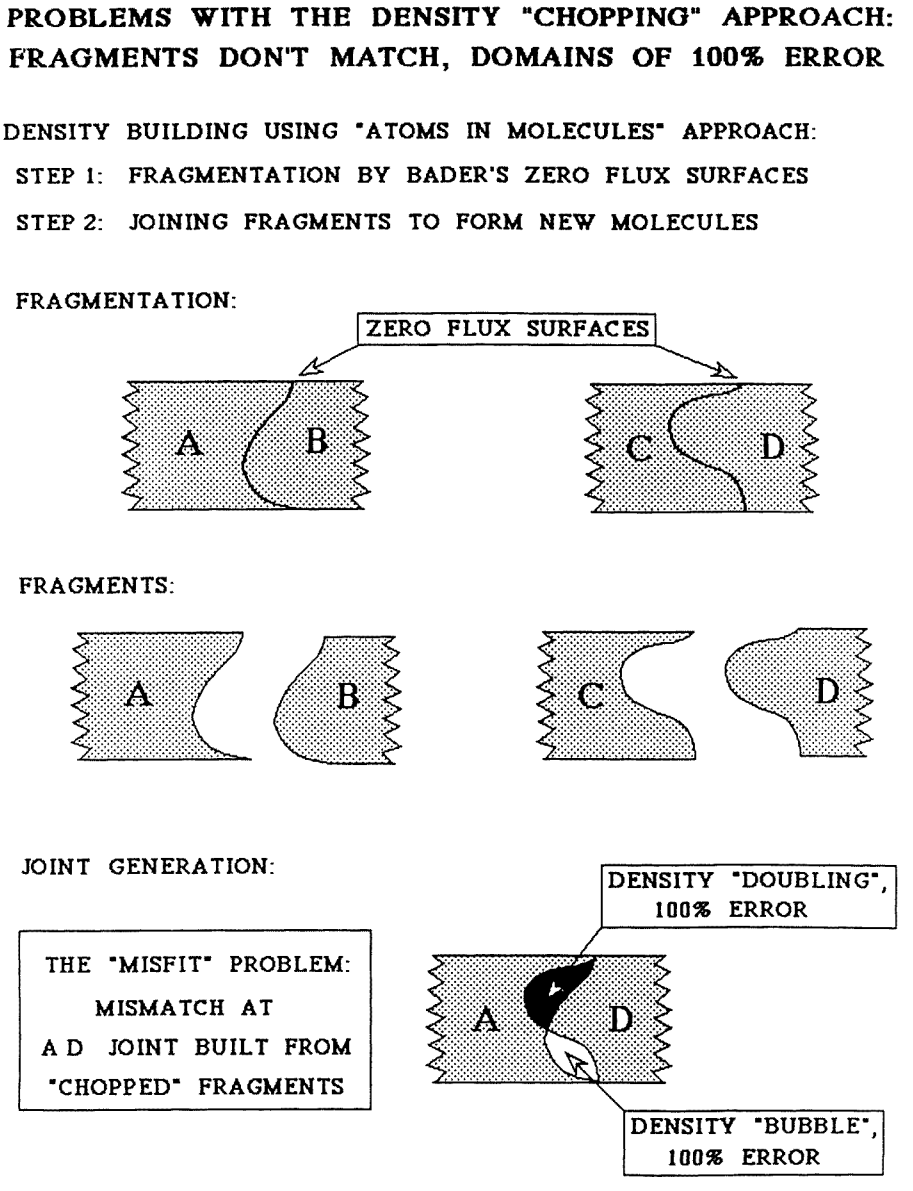

Fig. 7. The two types of errors of electron density construction schemes based on fragments with boundaries: "density doubling" and "density bubbles". Fragments with discrete boundaries, such as fragments with zero density flux boundaries obtained from the "atoms in molecules" approach, lead to 100% local errors of density doubling and density gaps, within the chemically important "bonding range" between the fragments.

Within the conventional SCF LCAO *ab initio* method, using a wavefunction computed for a molecule of some fixed conformation K, the electronic density $\rho(\mathbf{r})$ can be constructed in terms of atomic orbitals $\varphi_i(\mathbf{r})$ $(i = 1, 2, \ldots, n)$, where n is the

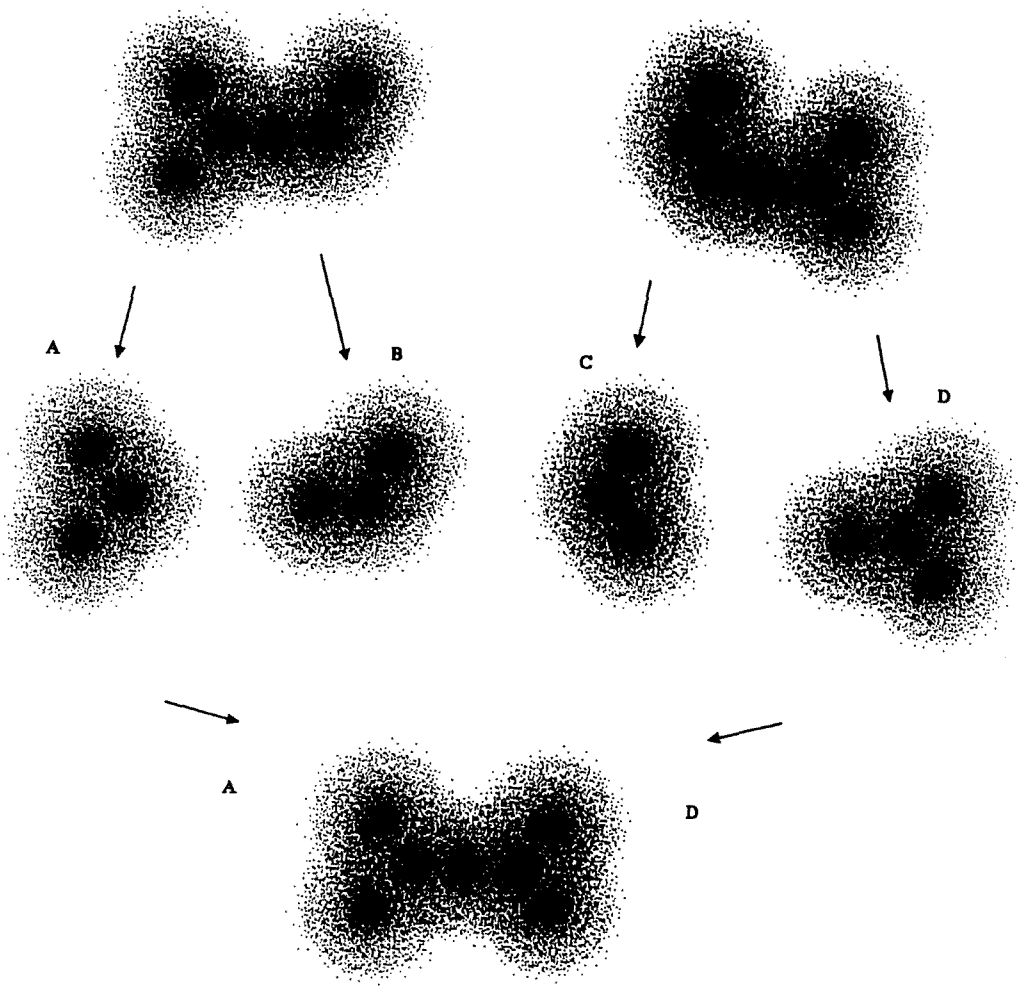

Fig. 8. A fuzzy electron density fragmentation scheme of boundaryless fragments used in MEDLA. This scheme, based on the mutual interpenetration of fuzzy electron density fragments, avoids the local accumulation of errors of schemes based on boundaries.

number of orbitals. If $\mathbf{r}$ is the three-dimensional position vector variable, and $\mathbf{P}$ is the $n \times n$ density matrix, then the electronic density $\rho(\mathbf{r})$ of the molecule is calculated as

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

This electron density $\rho(\mathbf{r})$ corresponds to the fuzzy "body" of the electronic charge cloud, providing a representation for the shape of the molecule.

The simplest of the additive, fuzzy fragmentation schemes of MEDLA can be obtained by a technique analogous to the Mulliken population analysis without

integration. The fuzzy fragments are obtained from relatively small molecules for which traditional *ab initio* calculations are feasible. The kth fuzzy fragment $\rho^k(\mathbf{r})$ of the molecular electronic density $\rho(\mathbf{r})$ can be defined for an arbitrary collection k of the nuclei of the molecule by first defining a fragment density matrix $\mathbf{P}^k$ of the same $n \times n$ dimensions as that of the density matrix $\mathbf{P}$ of the complete molecule.

The elements P_{ij}^k of this $n \times n$ fragment density matrix $\mathbf{P}^k$ for the kth fuzzy fragment $\rho^k(\mathbf{r})$ of the electron density $\rho(\mathbf{r})$ are defined as follows:

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

Based on this fragment density matrix $\mathbf{P}^k$, the electron density of the kth density fragment is defined [2] as

$$\rho^k(\mathbf{r}) = \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij}^k \varphi_i(\mathbf{r}) \varphi_j(\mathbf{r}) . \tag{3}$$

According to the fragmentation scheme, the nuclei of the molecule are distributed into m mutually exclusive families of nuclei, dividing the molecule into m fragments. Consequently, the sum of the fragment density matrices $\mathbf{P}^k$ is equal to the density matrix $\mathbf{P}$ of the molecule, and the sum of the $\rho^k(\mathbf{r})$ fragment densities is equal to the density $\rho(\mathbf{r})$ of the molecule:

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

and

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

The above fuzzy electron density fragment additivity rules (2)–(5) are *exact* on the given *ab initio* LCAO level. That is, the reconstruction of the electronic density $\rho(\mathbf{r})$ of the given small molecule from the corresponding fuzzy fragment electron densities $\rho^k(\mathbf{r})$ is exact. The scheme provides very satisfactory results for other molecules as well; all results reported in refs. [2,4,5], as well as those in the present report, have been obtained using the above simple scheme.

A more convenient form of the above fragmentation scheme is given in terms of membership functions of nuclei within various molecular fragments. As before, we divide the set of nuclei of the molecule into m mutually exclusive groups, denoted by

$$f_1, f_2, \ldots, f_k, \ldots f_m\,, \tag{6}$$

in order to generate m density fragments,

$$\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_k, \ldots \mathrm{F}_m\,, \tag{7}$$

of fragment density functions

$$\rho^1(\mathbf{r}), \rho^2(\mathbf{r}), \ldots, \rho^k(\mathbf{r}), \ldots \rho^m(\mathbf{r})\,, \tag{8}$$

respectively. The membership function $m_k(i)$ of AO $\varphi_i(\mathbf{r})$ in the set of AO's centered on a nucleus of nuclear set f_k of fragment F_k is defined as follows:

$$m_k(i) = \begin{cases} 1 & \text{if } \varphi_i(\mathbf{r}) \text{ is centered on one of the nuclei of set } f_k\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{9}$$

In terms of these membership functions, the elements P_{ij}^k of the $n \times n$ fragment density matrix $\mathbf{P}^k$ of the kth fragment F_k is defined as

$$P_{ij}^k = 0.5[m_k(i) + m_k(j)]P_{ij}\,. \tag{10}$$

The above fuzzy electron density fragment additivity principle [2] can be used for constructing electron densities for other molecules, where the size limitations of conventional *ab initio* computations can be circumvented. The additivity scheme provides a basis for building approximate electron densities for truly large molecules. Pre-calculated electron density fragments stored in a databank can be combined to form an approximate electron density for a different molecule, by selecting and arranging fuzzy fragment densities so that the nuclear positions closely match those in the target molecule. This principle is the basis for the molecular electron density "lego" assembler (MEDLA) technique [2–5], that has been implemented in a computer program, MEDLA 93 [3].

The additive, fuzzy electron density fragmentation scheme described by eq. (2), equivalent to that of eq. (10), is a special case of a more general scheme [15]. The fuzzy molecular electron density fragment additivity principle [2] can be formulated within a more general framework [15]. In the simplest scheme (eq. (2) or eq. (10)) the interfragment electron density is distributed by weighting the relevant interfragment density matrix elements P_{ij} by a factor of 0.5, and by including the resulting quantities in the fragment density matrix $\mathbf{P}^k$. However, the general additivity properties can be maintained by other schemes which distribute the interfragment electron density by some other weights. For example, formal fragment charges calculated in the parent molecules, or simple electronegativity comparisons can serve as the basis of alternative weighting schemes [15].

The more flexible weighting scheme described below leads to a more general electron density fragment additivity principle:

$$P_{ij}^{k} = \begin{cases} P_{ij} & \text{if both } \varphi_i(\mathbf{r}) \text{ and } \varphi_j(r) \text{ are AO's centered on nuclei of fragment } k, \\ w(k,i,j)P_{ij} & \text{if precisely one of the AO's } \varphi_i(\mathbf{r}) \text{ and } \varphi_j(\mathbf{r}) \text{ is centered on a nucleus of fragment } k, \text{ where for the weighting factors the relations } w(k,i,j) \geq 0, \text{ and } w(k,i,j) + w(k',i,j) = 1 \text{ hold, where the fragment } k' \text{ contains the nuclear center of the other AO,} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

For a physically valid additive scheme, the function $w(k,i,j)$ must fulfill some additional conditions. For example, one may take a sign-preserving scalar property $A(i)$ that can be assigned to atomic orbitals. Appropriately scaled electronegativity is such a scalar property. For any such scalar $A(i)$, the choice

$$w(k,i,j) = A(i)/[A(i) + A(j)], \tag{12}$$

where orbital $\varphi_i(\mathbf{r})$ is centered on a nucleus that belongs to the kth fragment, fulfills the required conditions [15].

The more general scheme can also be introduced using the membership function formalism of eq. (9), by taking

$$P_{ij}^{k} = [m_k(i)w_{ij} + m_k(j)w_{ji}]P_{ij}, \tag{13}$$

where for the weighting factors $w_{ij}, w_{ji} \geq 0$,

$$w_{ij} + w_{ji} = 1. \tag{14}$$

The simplest fragmentation scheme [2] corresponds to the choice of

$$w_{ij} = w_{ji} = 0.5. \tag{15}$$

Based on the scalar property $A(i)$ discussed above, a simple function w_{ij} can be chosen as

$$w_{ij} = A(i)/[A(i) + A(j)]. \tag{16}$$

Some non-additive schemes of generating pseudo-density matrices are also useful in the study of the local influence of molecular surroundings on the shapes of functional groups. According to one approach, a pseudo-density matrix ${}^{*}\mathbf{P}^{k}$ of a formal molecular fragment for a subset k of nuclei is defined by

$${}^{*}P_{ij}^{k} = \begin{cases} P_{ij} & \text{if AO } \varphi_i(\mathbf{r}) \text{ or } \varphi_j(\mathbf{r}) \text{ is centered on a nucleus of the fragment,} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The resulting pseudo-density ${}^*\rho^k(\mathbf{r})$ of the fragment,

$$ {}^*\rho^k(\mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^*P_{ij}^k \varphi_i(\mathbf{r})\varphi_j(\mathbf{r}) , \tag{18} $$

incorporates an enhanced contribution from the surroundings of the local molecular neighborhood, and can be used as a more sensitive diagnostic tool for the detection of shape differences induced by the placement of functional groups within various molecular neighborhoods. Note, however, that these pseudo-densities ${}^*\rho^k(\mathbf{r})$ of functional groups do not generate an additive scheme. Also note, that further enhancement of the shape-modifying effects of the molecular surroundings can be obtained by a progressive scaling of the interfragment contributions to the pseudo-density matrix ${}^*\mathbf{P}^k$ of a formal molecular fragment. Such progressive scaling can be obtained by taking

$$ {}^*P_{ij}^k = [m_k(i)m_k(j) + (1 - m_k(j))m_k(i)w_{ij} + (1 - m_k(i))m_k(j)w_{ji}]P_{ij} , \tag{19} $$

with weighting factors

$$ w_{ij}, w_{ji} > 1 . \tag{20} $$

By exaggerating the contributions of formal interfragment interactions, the numerical or visual detection and diagnosis of shape-modifying effects, represented numerically or displayed by formal isodensity contours, becomes a simpler task.

The special scheme of eqs. (2)–(10), as well as the general scheme of eqs. (11)–(16) incorporates an additive assignment of the interfragment interactions to the various fragment densities. In fact, the density-modifying effect of the local molecular neighborhood is incorporated within the fragment density in an additive manner. The pattern displayed in fig. 9 illustrates the general principle, whereas fig. 10 shows a chemical example. Assume that the entire molecule, schematically represented by the 48 blocks of fig. 9, is too large for a traditional *ab initio* calculation, however, *ab initio* electron densities for molecules of the size of 18 blocks can be computed directly. In such a case, one may designate the nuclei contained in each block as a nuclear set defining a fragment. For example, the fuzzy electron density fragment for the nuclear set of block F(3, 1, 3) can be calculated from the traditional *ab initio* electron density generated for the formal molecule of the 18 block piece indicated by the heavy line in the fig., including the nuclei of blocks F(i, j, k) for $2 \leqslant i \leqslant 4$, $1 \leqslant j \leqslant 2$, and $2 \leqslant k \leqslant 4$. The peripheral, "dangling" bonds can be tied down by additional H atoms. The local surroundings of fragment F(3, 1, 3) in this smaller molecule is the same as in the large target molecule. Hence, by generating the fuzzy fragment density $\rho^{F(3,1,3)}(\mathbf{r})$ for F(3, 1, 3) within the 18-block molecule, the appropriate "share" of local interactions, as distributed by the fragmentation scheme (2)–(10), are well represented in $\rho^{F(3,1,3)}(\mathbf{r})$, properly approximating these interactions as they are present within the large target molecule. Note that, the fragment electron density for F(2, 1, 4) is taken from a different com-

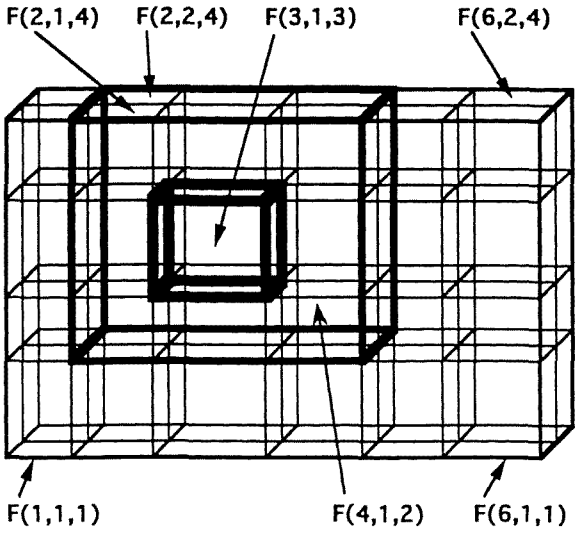

Fig. 9. A schematic illustration of the treatment of fragment-neighborhood interactions.

putation for another 18-block "molecule", the molecule obtained by tying down the "dangling" bonds at the periphery of the set of blocks $F(i,j,k)$ for $1 \leqslant i \leqslant 3$, $1 \leqslant j \leqslant 2$, and $2 \leqslant k \leqslant 4$. In effect, the principle of a "moving scaffold" is applied for each fragment when computing fragment densities. For each fragment density $\rho^{F(i,j,k)}(\mathbf{r})$ the effect of the local surroundings is properly represented, and when all these density fragments are combined in order to construct the electron density of the large molecule, high accuracy can be obtained.

The molecular example of fig. 10 shows three highlighted fragments, F_i, F_j, and F_f, as parts of the target molecule M (top of the figure), as separate entities (at the middle of the figure), and also as parts of smaller molecules $M(F_i)$, $M(F_j)$, and $M(F_f)$, (lower part of the figure), respectively. Within the latter three molecules the local molecular surroundings of these fragments are the same as in the target molecule M, whereas the distant parts of the target molecule are replaced by extra H atoms (denoted by outline font H). Consequently, the fragment densities F_i, F_j, and F_f, obtained by an *ab initio* computation for the smaller molecules $M(F_i)$, $M(F_j)$, and $M(F_f)$, followed by the application of the fragmentation procedure of eqs. (2)–(10), properly and additively represent their density contribution (short range interactions included) to the target molecule, to a good approximation. By applying a similar procedure for all fragments of molecule M, and by properly aligning and adding these fragment densities, the MEDLA electron density of molecule M is obtained.

Note that, the nuclei of fragment F_j are also present within molecule $M(F_i)$, used to generate fragment density F_i. In fact, one could obtain an approximation to the F_j fragment density from the *ab initio* calculation of molecule $M(F_i)$. However, within this molecule $M(F_i)$ the surroundings of nuclei of fragment F_j do not mimic

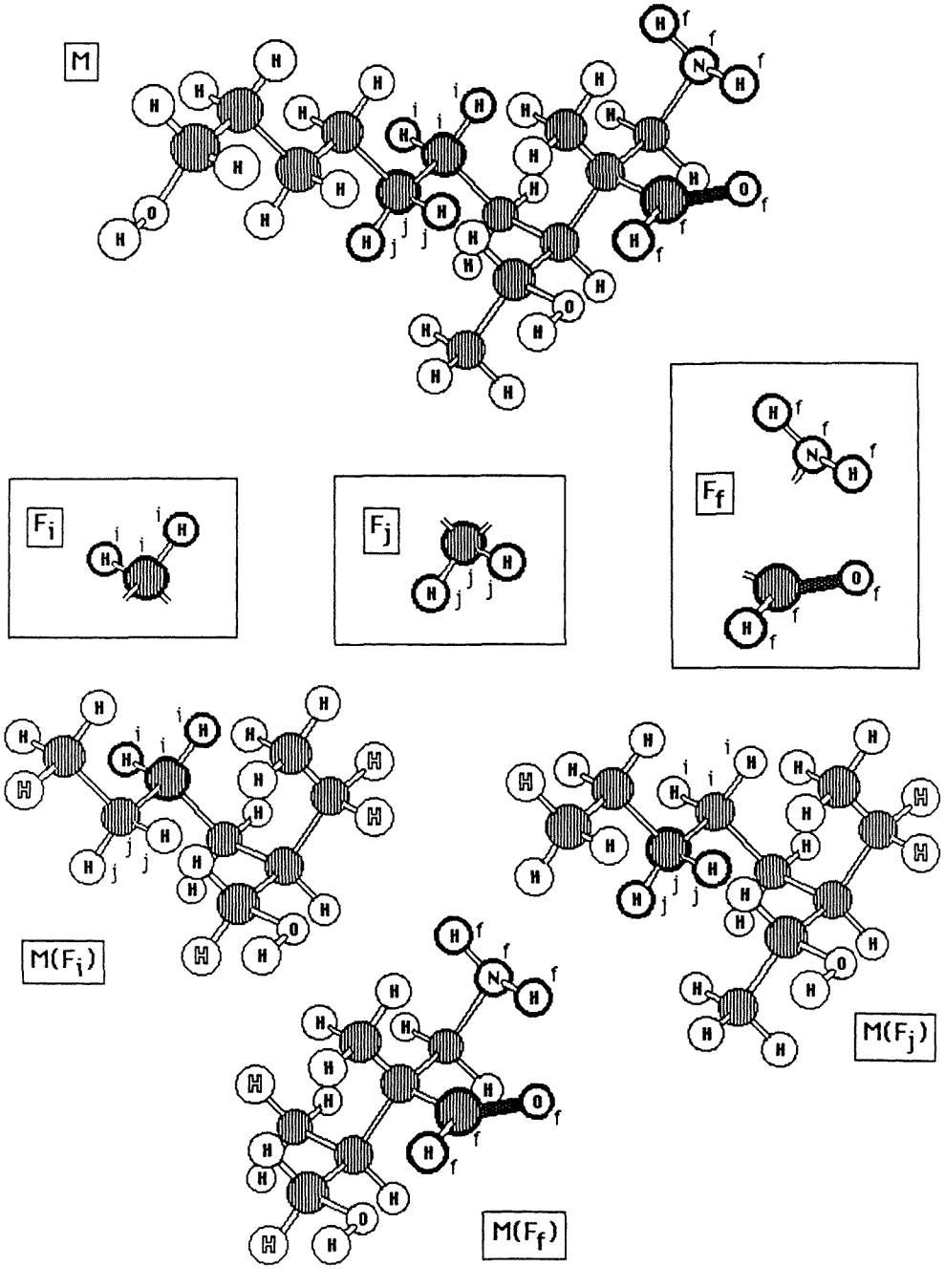

Fig. 10. A chemical illustration of the treatment of fragment-neighborhood interactions.

the actual surroundings in the target molecule M to the same level as these surroundings are reproduced in molecule $M(F_j)$. Consequently, the actual fragment density used for F_j is determined from a calculation for molecule $M(F_j)$, that provides a higher level of accuracy.

The accuracy of the approximation (within the given *ab initio* level of basis) depends on how large surroundings are included in the molecules $M(F_i)$, $M(F_j)$, and $M(F_f)$. Clearly, if these molecules contain a larger portion of the target molecule M, a more accurate final electron density is obtained.

It is possible to define formal fragments which are composed from non-interacting or weakly interacting parts. Fragment F_f of the example of fig. 10 is special, as it contains a formal hydrogen bond. Such fragments can be used to generate electron densities of hydrogen bonds closely resembling those obtained by standard *ab initio* computations. Note, however, that for the shape description of hydrogen bonds high accuracy has been obtained without invoking this possibility, and it appears that a simple addition of fragment densities of the correct relative nuclear geometry of the hydrogen bond is sufficient to mimic the shapes of hydrogen bond electron densities to a high degree of accuracy. In particular, in ref. [4], the shape of a hydrogen bond of a tetrapeptide has been computed by both traditional *ab initio* 6-31G** and MEDLA techniques, where no actual MEDLA fragment containing a hydrogen bond was used. The results of standard *ab initio* 6-31G** and MEDLA techniques have been found visually indistinguishable.

The pre-calculated, *ab initio* electron densities of fuzzy molecular fragments and functional groups, obtained from traditional *ab initio* calculations on small molecules, are stored in the MEDLA fragment density data bank. For simple molecules of standard or nearly standard nuclear arrangements, a small number of density fragments are sufficient. Typically, saturated, unbranched hydrocarbons in stretched conformations can be described with sufficient accuracy using a few fragments. If, however, crowded conformations or molecules with more varied local moieties are considered, then specialized fragments are required. Such specialized fragments can be obtained from traditional *ab initio* computations of small molecules, where local surroundings and conformation of the fragment resembles those in the target molecule. In such cases, several specialized electron density fragments (called versions) may be required for each nuclear family (fragment type) to account for the special steric arrangements, local polarities and interactions. The MEDLA density fragment database can be continuously updated as needed; there is virtually no limitation on the number of specialized fragments one may include in the database. A CD ROM of approximately 660 megabytes of memory can store 165 different density fragments of sufficient accuracy and variety for the calculation of nearly 6-31G** *ab initio* quality MEDLA electron densities for most proteins. If a much larger density fragment databank is used, then the search for the appropriate fragment can become somewhat time consuming.

Depending on the chemical problem, one may choose large or small fragments within the MEDLA scheme. The smallest fragments contain a single nucleus, such

as the example of the carbon fragment from the methane molecule, discussed in [2]. This actual carbon fragment clearly shows a density distribution characteristic of an sp3 neighborhood. Evidently, different versions of the carbon fragment are required for a slightly distorted methane, and certainly for an sp2-type or sp-type carbon density. If fragments with single nuclei are used, then a variety of fragments is required for each nucleus, to account for the variety of possible molecular neighborhoods where each fragment type may occur.

The computational time required for the construction of MEDLA electron densities from pre-calculated fragment densities grows linearly with the number of fragments, hence it grows (essentially) linearly with molecular size. Consequently, the method is easily applicable to large molecules. This is in contrast with the time requirement of traditional SCF Hartree–Fock methods where for large molecules the dependence is dominated by the fourth power of the number of AO basis functions.

The speed and accuracy of the MEDLA technique suggests new applications in the generation of "molecular level virtual reality". By simulating an excursion descending to the molecular world, for example, following the motion of a water molecule or a potential drug molecule within the interior of a protein cavity, only the locally "visible" portions of the molecules need be constructed, that can be achieved on a near real-time basis. This approach could give important insight for molecular modeling and drug design. The *molecular electron density lego assembler* may then serve as a "*computational microscope*", displaying the dynamic behavior of interacting molecules.

5. Generalized MEDLA schemes

Whereas the MEDLA method, in its simplest form [2], has been shown to generate reliable, *ab initio* quality electron densities for large molecules [4,5], and the accuracy of these densities appears sufficient for many applications, nevertheless, it is possible to modify the scheme and further increase the accuracy of MEDLA electron densities, if needed. There are several alternative choices for fragment selection and possible improvements, at the expense of considerably increasing the required CPU and memory demand of the method.

The need for local shape analysis of molecular fragments [7] has provided the original motivation for selecting the *density domains* of molecules [1] as the basis for defining fragments. A density domain $DD(a)$ is defined as a closed set in 3D space, enclosed by a MIDCO (molecular isodensity contour) $G(a)$ of some density threshold a. An arbitrary selection of nuclei does not necessarily correspond to a density domain, for example, in the ethanol molecule, the nuclei of the terminal CH_3 subset is not separated from the rest of the nuclei by a MIDCO for any threshold value [16]. In this sense, as manifested by electron density, the terminal CH_3 moiety is not a functional group of separate identity; the local charge clouds around the two car-

bon nuclei join one another at a higher density threshold than the clouds around the H nuclei join the cloud around the terminal C nucleus.

The pattern of density domains for the whole range of possible density thresholds *a* shows the gradual buildup of electron density as the bonding pattern of the molecule is established. Consequently, the subsets of nuclei enclosed within density domains at various thresholds are a natural choice as basis for fragmentation, following the schemes discussed in the previous section. Hence, a fragmentation scheme can be based on density domains.

An extreme alternative, mentioned in section 4 and illustrated by the example of a carbon fragment from ref. [2], is based on the smallest possible fragments, each containing a single nucleus.

A useful fragmentation scheme, typically involving four nuclei for the positioning of each fragment, is of special significance. Unless the four nuclei are coplanar, they define a tetrahedron. The electron distribution is dominated by the nuclear geometry, and for a small distortion of the tetrahedron the change of electron density can be approximated by applying the same distortion to the density. Any noncoplanar, tetrahedral arrangement of four nuclei can be obtained by a 3D *linear* transformation from a reference tetrahedron. Note, however, that polyhedra of five or more nuclei do not have the analogous property. Consequently, the case of fragment nuclear positioning fully specified by four nuclei is special. If the required fragment has a nuclear geometry that does not exactly match the nuclear geometry of a corresponding fragment stored in the density databank, but the geometries are similar, then a rapid, approximate, linear scaling of the electron density can be applied using the linear geometrical transformation that interconverts the nuclear arrangements. If the geometry change is small, then this method generates good quality approximate electron densities for the required fragment. A detailed analysis of this approach will be presented elsewhere [17].

The transformation itself can be obtained easily from the coordinates of four reference nuclei A, B, C, and D of the actual fragment in the target molecule and the coordinates of the corresponding four nuclei A′, B′, C′, and D′ in the fragment stored in the database. Three edge-vectors of each of the corresponding two tetrahedra are defined as

$$\mathbf{v}^{(1)} = \mathrm{A} \rightarrow \mathrm{B}\,, \tag{21}$$

$$\mathbf{v}^{(2)} = \mathrm{A} \rightarrow \mathrm{C}\,, \tag{22}$$

$$\mathbf{v}^{(3)} = \mathrm{A} \rightarrow \mathrm{D}\,, \tag{23}$$

$$\mathbf{w}^{(1)} = \mathrm{A}' \rightarrow \mathrm{B}'\,, \tag{24}$$

$$\mathbf{w}^{(2)} = \mathrm{A}' \rightarrow \mathrm{C}'\,, \tag{25}$$

$$\mathbf{w}^{(3)} = \mathrm{A}' \rightarrow \mathrm{D}', \tag{26}$$

respectively. It is useful to collect these column vectors into two matrices, **V** and **W**, with elements

$$V_{ij} = v_j^{(i)} \tag{27}$$

and

$$w_{ij} = w_j^{(i)}, \tag{28}$$

respectively. We assume that nuclei A and A′ are located at the origin of the coordinate system. The linear transformation **T** that converts the points of the ABCD tetrahedron into the corresponding points of the A′B′C′D′ tetrahedron is defined by the relation

$$\mathbf{TV} = \mathbf{W}, \tag{29}$$

that is, by

$$\mathbf{T} = \mathbf{WV}^{-1}. \tag{30}$$

The inverse matrix $\mathbf{V}^{-1}$ exists whenever the tetrahedron ABCD is nondegenerate, that is, whenever the four nuclei are not coplanar.

In the database, reference nucleus A′ is assumed to be at the origin. For the actual fragment ABCD of the target molecule, the translation placing nucleus A to the origin is denoted by **S**. The MEDLA density contribution $\rho_{\mathrm{ABCD}}(\mathbf{r})$ of the ABCD fragment to each point **r** of the target molecule can be obtained as follows:

$$\rho_{\mathrm{ABCD}}(\mathbf{r}) = \rho_{\mathrm{A'B'C'D'}}(\mathbf{TSr}), \tag{31}$$

where $\rho_{\mathrm{A'B'C'D'}}(\mathbf{p})$ is the electron density of fragment A′B′C′D′ at point **p**, stored in the MEDLA database. If the **TSr** transformation generates an out-of-range point **p** not stored in the database, then one sets

$$\rho_{\mathrm{ABCD}}(\mathbf{r}) = 0. \tag{32}$$

If four nuclei, A, B, C, and D are coplanar but not colinear, then D is replaced by a noncoplanar dummy nucleus, and the chemical fragment involves only three actual nuclei, whereas if A, B, and C are colinear, then C is replaced by a noncolinear dummy nucleus, and the fragment involves only two actual nuclei.

Note that the above transformation generates no distortion of the tetrahedron ABCD if the two tetrahedra are congruent, that is, if a fragment with the exact required nuclear geometry is found in the MEDLA database. In such a case, the electron density of the MEDLA database is used without distortion. In the case of exact coincidence of fragment nuclear geometries, an identical technique is also applicable for fragments involving more than four nuclei. If the nuclear geometries of such larger fragments do not coincide, then one may settle for lesser accuracy of the positions of nuclei additional to those defining the tetrahedron, and the

resulting lesser accuracy of electron density; in these cases the **TS** transformation still provides an approximate density transformation from the MEDLA database to the target molecule. If, however, high accuracy is required, then a new fragment density should be calculated with the exact nuclear geometry required, and this density fragment can be added to the database. With the new fragment in the MEDLA database, all distortions due to the transformations are avoided.

The size of fragments and the size of the "coordination shell" around them in the small molecule imitating the actual surroundings within the target molecule are limited only by the feasibility of traditional *ab initio* calculations. Satisfactory accuracy has been achieved in all the test calculations performed. It is possible that for large, conjugated systems, long range interactions may require excessive sizes for the fragments and for their surroundings; nevertheless, this problem can be circumvented. One approach is based on approximate periodicity: a periodic SCF technique can be used to generate a MEDLA fragment density, and this fragment can be built into the target molecule in the usual way.

The simplest implementation of the MEDLA method provides no inherent safeguards for total charge preservation. As the density fragments are superimposed, small deviations from the overall integer charge are possible. Although the overall accuracy of the MEDLA method implies that the error of total charge is likely to be very small, for specific purposes very high accuracy and an inherent condition of *exact* total charge preservation may be required. This can be achieved by a simple scaling of the MEDLA fragment densities during the course of the **TS** transformation. In the simplest implementation of this approach, the actual MEDLA fragments can be selected so that their fragment charges are integers. A scaling factor $f \sim 1.0$ can be determined for each fragment by an integral condition. If the integral (in practice, numerical integral) of the electron density restricted to the actually **TS**-transformed part of the MEDLA fragment is Q', and the actual electronic charge required for the MEDLA fragment is Q, then the factor

$$f = Q/Q' \tag{33}$$

is applied for the electronic density:

$$\rho_{\mathrm{ABCD}}(\mathbf{r}) = f\rho_{\mathrm{A'B'C'D'}}(\mathbf{TSr})\,. \tag{34}$$

This scaled MEDLA approach ensures proper total charge conservation.

It is also possible to scale the density fragment in the database to a prescribed electronic charge. To exploit this approach within a density grid approximation, one must ensure that all grid points and the corresponding densities of the given fragment of the database are transferred to the target molecule. This alternative allows one to store the properly scaled densities in the database, and there is no need for separate scaling in each instance the given database fragment is used.

A third alternative involves a single scaling, carried out on the calculated electron density as the final step. An integration (in practice, numerical integration) of the MEDLA density of the target molecule is performed, and a scaling factor f' is

determined that converts the approximate electronic charge to the required integer value (usually, the nearest integer). This scaling factor f' is then applied to the entire MEDLA electronic density of the target molecule, resulting in a *scaled MEDLA density*:

$$\rho_{\mathrm{SCMEDLA}}(\mathbf{r}) = f' \rho_{\mathrm{MEDLA}}(\mathbf{r}) . \tag{35}$$

The principle of additive, fuzzy electron density fragmentation is applicable within computational methods based on correlated wavefunctions and density functional theory. These approaches will be explored elsewhere.

6. Three-dimensional tiling approach for enhanced MEDLA densities

One enhancement of the accuracy of MEDLA electron densities is based on a method involving multiple tilings of the 3D space domain containing the target molecule. In the process of partially overlapping the fuzzy fragment densities in the target molecule, the electron densities are rather accurate near the centers of the fragments and the largest errors are expected at points where the density contributions from neighboring fragments are approximately the same at the peripheral ranges of each fragment. With reference to the scheme of fig. 9, showing a rectangular compartmentalization of the nuclei of the target molecule, the errors of the MEDLA densities are expected to be the largest near the boundaries of these compartments. Note that these compartment boundaries refer the partitioning of the nuclei into boxes, and the electronic clouds of the fuzzy density fragments themselves have no boundaries and extend beyond the nuclear compartment boundaries. It is possible to further increase the accuracy of the MEDLA densities by taking several, different compartmentalizations of the nuclei, ensuring that each boundary point of each compartment becomes an interior point of another compartmentalization. By properly weighting and combining the MEDLA densities obtained for each compartmentalization, a more accurate electron density can be computed for the target molecule.

The principle of this approach is illustrated by a two-dimensional example of $2^2 = 4$ different compartmentalizations (tilings) of the plane, shown in fig. 11. The edge length of each square is taken as π. The tiles indicated by the heavy solid lines correspond to the reference tiling A, and the tilings C, E and G are derived from A by translations, using the translation vectors $(0, \pi/2)$, $(\pi/2, 0)$, and $(\pi/2, \pi/2)$, respectively. Each boundary point of each tile is an interior point of another tile in some other tiling scheme.

The same principle is applied in the actual three-dimensional tiling approach, involving $2^3 = 8$ different tiling schemes, A, B, C, D, E, F, G, and H, as shown in fig. 12. The edge length of each cube is taken as π. The centers of cubes for the eight different tilings are given by the vectors

$$\mathbf{r}'_{u,v,w}(i,j,k) = [(i + 0.5u)\pi, (j + 0.5v)\pi, (k + 0.5w)\pi] , \tag{36}$$

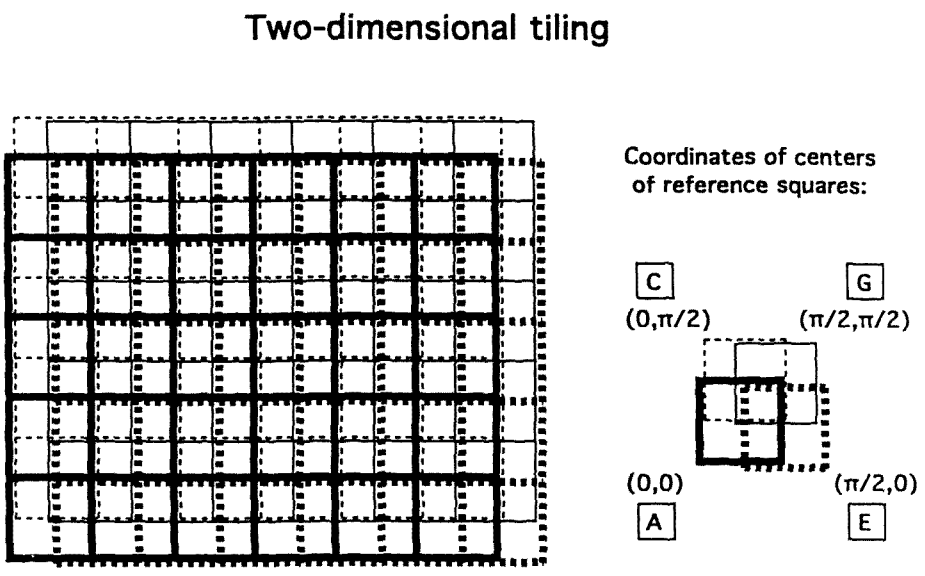

Fig. 11. A four-fold, two-dimensional tiling scheme.

where these A, B, C, D, E, F, G, and H tilings follow the lexicographic order of triples (u, v, w), with the integers u, v, and w fulfilling

$$0 \leqslant u, v, w \leqslant 1 . \tag{37}$$

In order to distinguish the MEDLA electron densities $\rho_{\mathrm{MEDLA}}(\mathbf{r})$ obtained by the original technique from those generated by a tiling technique, the notation $d(\mathbf{r}) = d(x, y, z)$ will be used for the latter.

The electron density $d(x, y, z)$ is calculated by a trigonometric weighting of the $d_{uvw}(x, y, z)$ MEDLA electron densities of the individual tilings:

$$d(x,y,z) = \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{w=0}^{1} d_{uvw}(x,y,z)[u + (1-2u)\cos^2 x] \times [v + (1-2v)\cos^2 y][w + (1-2w)\cos^2 z] , \tag{38}$$

where for $u = 0$

$$[u + (1-2u)\cos^2 x] = \cos^2 x , \tag{39}$$

and for $u = 1$

$$[u + (1-2u)\cos^2 x] = 1 - \cos^2 x = \sin^2 x . \tag{40}$$

The above concise formula for $d(x, y, z)$ is equivalent to the lengthier formula given in fig. 12. Similarly to the two-dimensional case, each boundary point of each tile is an interior point of another tile in some other tiling scheme. Furthermore, the trigonometric interpolation formula (38) between the individual MEDLA densities

Three-dimensional tiling

Coordinates of centers of cubes of tilings A, B, C, D, E, F, G, H:

	A	B	C	D	E	F	G	H
x	$i\pi$	$i\pi$	$i\pi$	$i\pi$	$(i+0.5)\pi$	$(i+0.5)\pi$	$(i+0.5)\pi$	$(i+0.5)\pi$
y	$j\pi$	$j\pi$	$(j+0.5)\pi$	$(j+0.5)\pi$	$j\pi$	$j\pi$	$(j+0.5)\pi$	$(j+0.5)\pi$
z	$k\pi$	$(k+0.5)\pi$	$k\pi$	$(k+0.5)\pi$	$k\pi$	$(k+0.5)\pi$	$k\pi$	$(k+0.5)\pi$

The eight direct MEDLA electron densities by fragment generation using tilings A, B, C, D, E, F, G, and H are:

$d_A(x,y,z)$, $d_B(x,y,z)$, $d_C(x,y,z)$, $d_D(x,y,z)$,

$d_E(x,y,z)$, $d_F(x,y,z)$, $d_G(x,y,z)$, $d_H(x,y,z)$.

Enhanced MEDLA electron density by trigonometric interpolation:

$$\begin{aligned} d(x,y,z) = {} & d_A(x,y,z)\cos^2 x \cos^2 y \cos^2 z + d_B(x,y,z)\cos^2 x \cos^2 y \sin^2 z \\ & + d_C(x,y,z)\cos^2 x \sin^2 y \cos^2 z + d_D(x,y,z)\cos^2 x \sin^2 y \sin^2 z \\ & + d_E(x,y,z)\sin^2 x \cos^2 y \cos^2 z + d_F(x,y,z)\sin^2 x \cos^2 y \sin^2 z \\ & + d_G(x,y,z)\sin^2 x \sin^2 y \cos^2 z + d_H(x,y,z)\sin^2 x \sin^2 y \sin^2 z \end{aligned}$$

Fig. 12. An eight-fold, three-dimensional tiling scheme and trigonometric averaging for enhanced MEDLA electron densities.

$d_{uvw}(x, y, z)$ obtained in each of the eight tiling schemes ensures that for each tiling (u, v, w) the individual MEDLA contribution $d_{uvw}(x, y, z)$ is weighted by zero for each boundary point of each tile, it is weighted by 1.00 for each midpoint of each tile, yet the sum of all weights of contributions $d_{uvw}(x, y, z)$ adds up to a total weight of 1.00. This scheme is designed to eliminate most of the (already small) errors of the fuzzy density contributions at the boundaries of nuclear compartments, and to improve the overall accuracy of the approach.

The calculation of the $d(x, y, z)$ interpolated MEDLA electron density requires approximately eight times more CPU time than the individual MEDLA densities $d_{uvw}(x, y, z)$. Some care must be taken in order to avoid nuclear positions falling on tile boundaries. This problem can always be avoided by a small translation of the tiling schemes or by selecting a different edge length for the tiles (in fact, by selecting a different unit for length, since the edge length has a numerical measure of π within the scheme).

7. Fuzzy set approach to additive electron density fragmentation

The fuzzy electron density contribution of each fragment to the points **r** of the three-dimensional space can be described by the methods of fuzzy set theory. One may ask the question: to which molecular fragment (density fragment) does the given point **r** belong? This problem can be phrased in terms of the following simple scheme.

Consider the MEDLA electron density constructed according to the method described by eqs. (3) and (5), using any of the *additive* fragmentation schemes, such as those described by eq. (2), or by eqs. (11), (12), (13), (14), (16). In principle, each of the m density fragments,

$$\rho^1(\mathbf{r}), \rho^2(\mathbf{r}), \ldots, \rho^k(\mathbf{r}), \ldots \rho^m(\mathbf{r}),$$

contribute to the overall MEDLA electron density at point **r**. With reference to the total MEDLA electron density

$$\rho(\mathbf{r}) = \sum_{k=1}^{m} \rho^k(\mathbf{r}),$$

a fuzzy membership function of point **r** within each molecular fragment

$$\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_k, \ldots \mathrm{F}_m$$

can be defined as

$$\mu_k(\mathbf{r}) = \rho^k(\mathbf{r})/\rho(\mathbf{r}). \tag{41}$$

This membership function for point **r** expresses the degree of belonging of point **r** to fragment F_k, as expressed by the relative contribution of fragment density $\rho^k(\mathbf{r})$ to the total MEDLA electron density $\rho(\mathbf{r})$ in the given molecule. The sum of these membership functions is unity,

$$\sum_{k=1}^{m} \mu_k(\mathbf{r}) = 1. \tag{42}$$

For each point **r** the density fragment with the largest membership function value is regarded as the molecular moiety primarily occupying the given location **r** of

the three-dimensional space. If the density fragments are selected on the basis of a *density domain* criterion, then this fuzzy set approach provides a density-based criterion for deciding which functional group F_d exerts the dominant influence over a given region of space within a molecule. For the given region R_d,

$$R_d = \{\mathbf{r} : \mu_d(\mathbf{r}) = \max\{\mu_1(\mathbf{r}), \mu_2(\mathbf{r}), \ldots \mu_d(\mathbf{r}), \ldots \mu_m(\mathbf{r})\}\} . \tag{43}$$

Usually, there is only a single maximum connected component for each index d, that is, for each functional group F_d. If there are two or more maximum connected components of R_d, this is an indication that the fragment F_d is subject to strong interfragment interactions within the target molecule, and F_d is likely to lack the property of possessing an isodensity contour within the target molecule that separates the nuclei of the fragment from all other nuclei of the target molecule. If this is the case, fragment F_d no longer qualifies as a density domain functional group, as it ceases to be one within the target molecule, even though it is a functional group within its parent molecule.

8. A MEDLA-based enhancement of the X-ray structure refinement process for better experimental electron densities

Diffraction intensities of X-ray structure determination serve as the basis of direct determination of atomic coordinates of molecular species forming a crystal [18]. In more conventional X-ray structure determination, the so-called phases of the diffracted waves are used. Exploiting the three-dimensional periodicity of the crystal, the electron density distribution $\rho(\mathbf{r})$ is described by a Fourier series,

$$\rho(\mathbf{r}) = V^{-1} \sum_{\mathbf{h}} \mathbf{F}_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}) , \tag{44}$$

where V is the volume of the unit cell of the crystal, $\mathbf{h}$ is a vector of integral components h, k, and l, whose values are inversely proportional to the intercepts of the axes defining the unit cell with the imaginary plane $P_{\mathbf{h}}$, cutting through the crystal, and the $\mathbf{F}_{\mathbf{h}}$ structure factors are real or complex numbers representing the characteristics of the X-ray scattering associated with the planes $P_{\mathbf{h}}$. The structure factors can be expressed as

$$\mathbf{F}_{\mathbf{h}} = |\mathbf{F}_{\mathbf{h}}| \exp(i\phi_{\mathbf{h}}) , \tag{45}$$

where the angle $\phi_{\mathbf{h}}$ is the phase associated with $\mathbf{F}_{\mathbf{h}}$.

In most of X-ray structure determination approaches based on phases, the local electron density distributions are assumed to be spherical or elliptical, represented, for example, by spherical or elliptical gaussian functions. Based on an initial analysis of the X-ray scattering results, and on accumulated information on the structure of known molecular moieties, estimated nuclear coordinates and the gaussian electron densities associated with the assumed nuclear locations are used to inter-

pret the diffraction data. Usually, this is an iterative process, called structure refinement, where by gradually readjusting the assumed nuclear positions, and the associated gaussian electron density distributions, an improved agreement is obtained between the actual diffraction data and the nuclear geometry.

The MEDLA method provides an improved representation of the electron density distribution in the structure refinement process. For each assumed nuclear arrangement of the iterative process, a MEDLA electron density distribution can be computed, that can replace the gaussian density representations of the conventional technique by more realistic electron densities. In the process of X-ray structure refinement, the MEDLA electron densities are updated (recalculated) in each step within the iterative scheme.

In the context of the structure refinement process of X-ray crystallography, the advantages of more realistic densities obtained by MEDLA are twofold:

(i) using a more faithful electronic charge density for each assumed nuclear geometry in the course of the iterative structure refinement process, the comparison with the experimental diffraction pattern in each iterative step becomes a more sensitive and more reliable criterion for accepting or rejecting an assumed structure,

(ii) the more accurate density representation allows a more exhaustive interpretation and utilization of the structural information contained in the experimental diffraction data.

9. MEDLA-based conformation analysis approaches

The electron densities computed by the MEDLA method form the basis of approximate energy relations. As follows from the fundamental theorems of density functional theory, for the ground electronic state, the electron density determines the energy. Whereas the actual construction of such energy functions from first principles is a problem that has not been solved yet satisfactorily for large molecules, the MEDLA scheme allows one to introduce a practical, approximate representation of molecular energy of large systems. Interfragment interactions can be modeled by semiclassical potentials, analogous to a *molecular mechanics* approach, where for the short range interactions quantum mechanical effects dominate, whereas for the long range interactions electrostatic and other steric effects are dominant. Such a scheme, described in detail elsewhere [17], is useful for generating approximate energy functions and geometry optimization algorithms for macromolecular conformation analysis, protein folding problems, and intermolecular interactions. For an improved energy representation, the atomic core regions require a higher resolution for density grid points, where a technique, analogous with the FSGH fused spheres guided homotopy method [1] of MIDCO modeling can be used.

Within the interior of a globular protein, the actual space filling aspects are parti-

cularly well represented by a sequence of MEDLA MIDCOs $G(a)$ for different density thresholds a. These space filling characteristics are determined by a merging of electronic density clouds between parts of the protein not linked directly by formal bonds. This is an important feature not well represented by earlier models, such as fused spheres VDW surfaces. The experience with protein MEDLA MIDCOs indicates that these mergers start to occur simultaneously at about the same density threshold a_m, at many locations within the protein. As suggested in ref. [4], this trend, observed for the favored conformations of proteins, can give a tool for partially justifying favored mutual side chain arrangements and folding patterns.

The simplest utilization of this idea is the basis for the *Self-Avoiding MIDCO* approach to macromolecular conformation analysis. By selecting a suitable critical threshold value a_m characterizing the onset of mergers, the corresponding MIDCOs $G(K, a_m)$ can be generated for a range of nuclear configurations K. A suitable threshold value for a_m is likely to fall within the range [0.003 a.u., 0.005 a.u.]. In addition to a_m, a suitable small tolerance criterion $\Delta a < 0.001$ a.u. is chosen. For hydrogen bonds, regarded as a special case, a threshold $a_H \sim 0.007$ a.u. appears appropriate. Alternatively, each interaction for all nuclear pairs, not represented by conventional bonds, can be assigned a specific critical threshold value or a range of threshold values, forming an interval including values for hydrogen bonds.

A simple *contact principle* can be used for accepting and rejecting nuclear configurations:

> A given configuration K is accepted if all nonbonded mergers which appear for MEDLA MIDCO $G(K, a_m - \Delta a)$ and all hydrogen bonds which appear for MEDLA MIDCO $G(K, a_h - \Delta a)$ are not yet merged in MEDLA MIDCOs $G(K, a_m + \Delta a)$ and $G(K, a_h + \Delta a)$, respectively.

If the nonbonded mergers, including hydrogen bonds, occur within the specified ranges, this conforms with the experience obtained with favored conformations of proteins. Hence, the above criterion is expected to serve as a valid basis for configuration selection. By scanning a range of nuclear configurations K, one may maximize the number of mergers fulfilling this criterion. The nuclear arrangement K_{mm} with the maximum number of proper mergers is expected to provide a good approximation to a preferred nuclear arrangement.

Note that this approach cannot be regarded as a MIDCO version of hard surface contact models, such as those obtained using fused sphere VDW surfaces. The various parts of a macromolecular MIDCO folding back upon itself readjust not only their local conformation but also their size. In contrast to fused sphere models, in the vicinity of a merger, the actual local shape of the MIDCO undergoes a dramatic change. The mutual interpenetration of electron densities of molecular parts placed side by side increases the electron density of both parts, hence the MIDCO $G(K, a)$ for the given threshold a shows a significant "swelling" directed towards the site of the eventual merger evident at some lower density threshold. This feature

of the Self-Avoiding MIDCO method is responsible for the incorporation of interactions into the conformation analysis approach, without actually relying on energy considerations.

A simple, approximate conformational energy function can be based on a reward-and-penalty function associated with these mergers. A *density threshold potential function*, analogous in shape to a Morse potential where the distance variable is replaced by the value of the density threshold a where the merger occurs, can be associated with each merger. The (negative) minimum of each of these density threshold potentials is at the a_m or a_h value, for nonbonded interactions and hydrogen bonds, respectively, the potentials are zero for zero density threshold $a = 0$, and have high positive values for high density thresholds. A minimization of the sum of these potentials can be used in the search for favored nuclear arrangements K. Details of this technique will be presented elsewhere [17].

10. Closing remarks

During the past decade computational quantum chemistry has become competitive with experimental structure determination methods for small molecules. The MEDLA method further enhances the quantum chemical approach, and the applications of theoretically sound, physical approaches to biochemistry. The new MEDLA computational microscope, as applied to macromolecules such as proteins and other natural products, surpasses both the accuracy and scope of current experimental techniques for the generation of detailed, realistic molecular images; the new technique gives a detailed view of the molecular world. For the first time, reliable shape and size studies can be carried out for macromolecules of thousand atoms or more, which up till now were not amenable to high resolution experimental or theoretical electron density determination. Using the MEDLA computational microscope, biochemists will be able to visualize faithful, detailed images of the macromolecules they study.

The numerical shape similarity measures, as applied to MEDLA electron densities, are expected to become useful tools in computer-aided molecular engineering and pharmaceutical drug design. The MEDLA electron densities and associated molecular shape analysis may contribute important approaches to the solution of the most puzzling aspects of the formation of life-forming biopolymers [19]; in particular, the natural selection of a relatively small number of definite-sequence polymers from the myriads of diverse possibilities. The MEDLA electron densities provide a shape-based selection criterion augmenting the tools used in a current approach to the problem of definite-sequence biopolymers [20,21].

Acknowledgement

This work was supported by NSERC of Canada.

References

[1] P.G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology* (VCH Publ., New York, 1993).
[2] P.D. Walker and P.G. Mezey, J. Amer. Chem. Soc. 115 (1993) 12423.
[3] P.D. Walker and P.G. Mezey, *Program MEDLA 93* (Mathematical Chemistry Research Unit, University of Saskatchewan, Saskatoon, Canada, 1993).
[4] P.D. Walker and P.G. Mezey, J. Amer. Chem. Soc. 116 (1994) 12022.
[5] P.D. Walker and P.G. Mezey, Canad. J. Chem. 72 (1994) 2531.
[6] P.D. Walker and P.G. Mezey, J. Comput. Chem., in press.
[7] P.G. Mezey, Int. J. Quant. Chem. Quant. Biol. Symp. 14 (1987) 127.
[8] K.C. Nicolau, Z. Yang, J.J. Liu, H. Ueno, P.G. Nantermet, R.K. Guy, C.F. Claiborne, J. Renaud, E.A. Couladouros, K. Paulvannan and E.J. Sorensen, Nature 367 (1994) 630.
[9] (a) R.A. Holton, C. Somoza, H.-B. Kim, F. Liang, R.J. Biediger, P.D. Boatman, M. Shindo, C.C. Smith, S. Kim, H. Nadizadeh, Y. Suzuki, C. Tao, P. Vu, S. Tang, P. Zhang, K.K. Murthi, L.N. Gentile and J.H. Liu, J. Amer. Chem. Soc. 116 (1994) 1597.
(b) R.A. Holton, H.-B. Kim, C. Somoza, F. Liang, R.J. Biediger, P.D. Boatman, M. Shindo, C.C. Smith, S. Kim, H. Nadizadeh, Y. Suzuki, C. Tao, P. Vu, S. Tang, P. Zhang, K.K. Murthi, L.N. Gentile and J.H. Liu, J. Amer. Chem. Soc. 116 (1994) 1599.
[10] *Program BIOGRAF* (Biodesign, Inc., 199 S. Los Robles Ave., Pasadena, CA 91101, 1988).
[11] S. Spinelli, Q.Z. Liu, P.M. Alzari, P.H. Hirel and R.J. Poljak, Biochimie 73 (1991) 1391 (BPDB 15-OCT-92 1HHP).
[12] R.F.W. Bader and T.T. Nguyen-Dang, Adv. Quant. Chem. 14 (1981) 63.
[13] R.F.W. Bader, Acc. Chem. Res. 9 (1985) 18.
[14] C. Chang and R.F.W. Bader, J. Phys. Chem. 96 (1992) 1654.
[15] P.G. Mezey, Methods of molecular shape-similarity analysis and topological shape design, in: *Molecular Similarity in Drug Design*, ed. P.M. Dean (Chapman & Hall – Blackie Publ., Glasgow, UK, 1995).
[16] P.G. Mezey, Canad. J. Chem. 72 (1994) 928.
[17] P.D. Walker and P.G. Mezey, to be published.
[18] J. Karle, Proc. Natl. Acad. Sci. USA 88 (1991) 10099.
[19] K. Fukui, *Proc. 5th IFC Symp.*, Institute for Fundamental Chemistry, Kyoto, 1989.
[20] S. Arimoto, K. Fukui, K.F. Taylor and P.G. Mezey, Int. J. Quant. Chem. 53 (1995) 375.
[21] S. Arimoto, K. Fukui, K.F. Taylor and P.G. Mezey, Int. J. Quant. Chem. 53 (1995) 387.