Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 22 June 2001
Accepted 23 November 2001

Modeling chemical bonding effects for protein electron crystallography: the transferable fragmental electrostatic potential (TFESP) method

Shijun Zhong,[a]†‡ Voichita M. Dadarlat,[b]† Robert M. Glaeser,[a,c] Teresa Head-Gordon[b,d] and Kenneth H. Downing[a]*

[a]Life Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA, [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA, [c]Department of Molecular and Cell Biology and Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA, and [d]Department of Bioengineering, University of California, Berkeley, CA 94720, USA. Correspondence e-mail: khdowning@lbl.gov

This paper addresses the problem of determining the electrostatic potential of large proteins by the superposition of potentials calculated for small fragments. The use of different atomic and molecular fragments is considered for reproducing the molecular electrostatic potential of different conformations of *N*-acetylalanine methylamide (NAAMA) with an acceptable degree of error as measured by conventional *R* factors used in crystallographic structure refinement. Three different divisions of NAAMA are tested, producing fragments that incorporate increasingly more complete descriptions of molecular bonding with diminishing accuracy in geometric fit to the parent molecule: single atoms in molecules, bonded atoms in molecules and selected functional groups, such as the backbone peptide moiety, or the α-carbon, β-carbon and their associated H atoms. In the resolution range 2.5–25 Å, the fairly straightforward use of single atoms in molecules reduces the calculated *R* factors by 5–15% over a free-atom superposition. No significant further improvement was found at the lowest resolutions with a superposition of single bonds in molecules and *R* factors were found to degrade with larger fragments at higher resolutions because of poor geometry fits to the atoms of the parent molecule. Because the potential distribution even for single atoms depends on the environment, the best accuracy will be obtained by using a library of fragment potentials calculated for each type of atom as a function of important protein conformations.

1. Introduction

Electron crystallography provides an approach complementary to X-ray crystallography for structure determination for proteins in those cases when two-dimensional but not three-dimensional crystals can be obtained, especially for proteins less than 200–500 kDa where single-particle imaging techniques are not anticipated to be applicable at atomic resolution in the near future (Glaeser, 1999). Recent progress in protein electron crystallography has led to the increasing ability to generate atomic resolution structures of biologically significant proteins. The structures of the light-harvesting complex II (Kühlbrandt *et al.*, 1994), tubulin (Nogales *et al.*, 1998) as well as the electron-crystallography prototype bacteriorhodopsin (Henderson *et al.*, 1990; Grigorieff *et al.*, 1996; Kimura *et al.*, 1997) have been solved to resolutions between 3 and 4 Å. The more recent completion of a high-resolution model of the human AQP1 water channel (Ren *et al.*, 2000), and continuing progress on the nicotinic acetylcholine receptor (Miyazawa *et al.*, 1999) suggest that structure determination by electron crystallography will continue to grow in importance.

The validity of a structural model based on either electron or X-ray crystallography data is tested by computing *R* factors, $R(s)$, that relate the observed structure factors, $F_{\rm obs}$, to those calculated from the model, $F_{\rm calc}$:

$$R(s) = \frac{\sum |F_{\rm obs}(s)| - k|F_{\rm calc}(s)|}{\sum |F_{\rm obs}(s)|}, \qquad (1)$$

where the resolution (*i.e.* spatial frequency) is defined by

† These two authors contributed equally to this work.
‡ Present address: Chemistry Department, Wesleyan University, Middletown, CT 06459, USA.

$$s = 2\sin(\theta/2)/\lambda. \quad (2)$$

k is a scale factor and the sums are over a set of diffraction amplitudes in the neighborhood of s. We include the factor k in the definition of the R factor as is typically performed in the X-ray crystallography literature, although for all calculations reported here k is set to one. Typical overall R factors for well refined protein structures in X-ray crystallography are as low as 15%. Partly because electron diffraction amplitudes are not as accurate as those currently obtained with proteins by X-ray diffraction, R factors are typically found to be ~30% or more, but in the best case of refinement can be as low as 25%.

In addition to experimental limitations that currently affect the accuracy of F_{obs}, it is likely that the large values of R factors seen in the electron crystallographic structures are also partly attributable to errors in computing F_{calc} using scattering factors for free (*i.e.* unbonded and neutral) atoms. In the lower-resolution range of data typically obtained with electron crystallography, chemical bonding significantly affects the scattering amplitudes. The electrostatic potential for atoms within molecules is affected by charge redistribution within the molecular environment. In order to quantify the magnitude of the effect of molecular bonding on calculated R factors, we previously compared the Fourier transforms of the electrostatic potentials of a representative collection of protein fragments and guanosine triphosphate (GTP) (a ligand for tubulin and many signaling proteins) based on both free-atom (spherical) scattering factors and on accurate molecular-orbital calculations of the electrostatic potential (Chang *et al.*, 1999). Comparison of these potentials, and their Fourier transforms, showed that errors in amplitudes well over 10% can be expected at resolutions below 5.0 Å ($s < 0.2$ Å^{-1}) when the spherical scattering factors for neutral atoms are used to calculate the molecular structure factors.

Because it is quite clear that better use of low-resolution data can be realized by accounting for chemical bonding effects, we have now investigated how best to incorporate molecular bonding effects into the refinement of the data obtained in electron crystallographic studies. In principle, the best approach would be the calculation of structure factors based on *ab initio* evaluation of the molecular electrostatic potential (MEP) for the entire protein. In the near term, such an approach is just feasible for calculating the electrostatic potential of a small protein with a self-consistent field (SCF) or density functional theory (DFT) level of theory with a reasonable basis-set size, but it is not computationally feasible for the size of proteins of interest and for the necessary number of iterations needed in a refinement calculation. A more tractable approach at present is to modify or replace the atomic form factors themselves, for example by systematic development of form factors based on a database of chemically bonded atomic or molecular fragments. Such an approach would be readily possible using current and standard electronic structure algorithms and computer hardware.

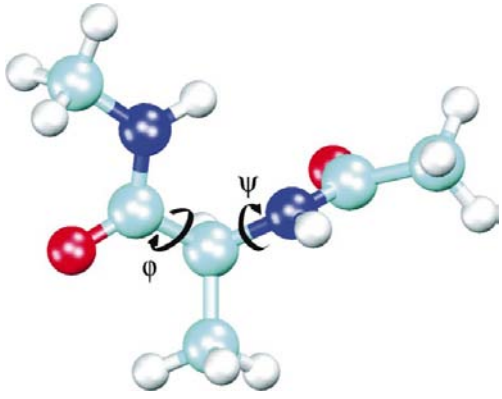

Figure 1
N-acetylalanine methylamide (NAAMA). The prototype small-protein molecule used to explore the best computational approach for removing the error in refinement of electron crystallography data due to chemical bonding effects. Dark blue denotes nitrogen, light blue denotes carbon, red for oxygen and white for hydrogen. The backbone dihedral angles φ and ψ define the different conformations of NAAMA used in this study: *C*1 ($\varphi = -129$, $\psi = 30°$), *C*2 ($\varphi = -57$, $\psi = -47°$) and *C*3 ($\varphi = -70$, $\psi = 70°$).

In this paper, we consider the definition of atomic or molecular fragments and the protocol of superposition of fragment electrostatic potentials that minimizes the errors in reproducing the electrostatic potential of the small dipeptide molecule *N*-acetylalanine methylamide (NAAMA, Fig. 1) in three different conformations described by the backbone dihedrals φ and ψ. Ideally, we would like the sum of the fragmental electrostatic potentials (FESP), $V_i(\mathbf{r})$, to be exactly equal to the electrostatic potential for the whole molecule. However, the superposition of FESPs can only approximate the exact value for the whole molecule owing to differences in the local bonding environments and/or geometry between the molecule and the fragments used to represent it. We have therefore considered three different divisions of the parent NAAMA molecule into fragments that incorporate increasingly more complete descriptions of molecular bonding with diminishing accuracy in geometric fit: single atoms in molecules, bonded atoms in molecules, and functional groups such as the backbone peptide moiety and the α-carbon and amino acid side chain. This paper presents an investigation as to which of these fragment definitions best reproduces the electrostatic potential of the NAAMA molecule as the prototype of real proteins, thereby formulating a future approach for the practical application of molecularly derived form factors and their use in electron crystallography.

2. Methods

There is no rigorous first-principle approach based on physical laws that guides us in the best division of electron density of a molecule to define the optimal atomic or molecular form factor, although various well defined schemes have been proposed (Bader, 1990). We will quantify whether one fragment approach is better than another in incorporating chemical bonding effects using the R factor in (1).

The molecular electrostatic potential can be expressed in the *ab initio* LCAO (linear combination of atomic orbitals) framework (Politzer & Truhler, 1981; Johnson *et al.*, 1993; Szabo & Ostlund, 1996):

$$V(\mathbf{r}) = \sum_a \frac{Z_a}{|\mathbf{R}_a - \mathbf{r}|} - \sum_\mu \sum_\nu P_{\mu\nu} \int \frac{\varphi_\mu(\mathbf{r})\varphi_\nu(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} \mathrm{d}\mathbf{r}'. \quad (3)$$

The first term of $V(\mathbf{r})$ is the nuclear contribution to the molecular electrostatic potential and $\mathbf{R}_a$ represents the atomic position. The second term is the electronic contribution. $\varphi_\mu(\mathbf{r})$ and $\varphi_\nu(\mathbf{r})$ are orbital basis functions, and $P_{\mu\nu}$ is the corresponding element of an appropriate density matrix which is usually produced in the SCF (self-consistent field) process. The double summation in the second term runs over all pairs of orbital basis functions. The density matrix element is defined as

$$P_{\mu\nu} = \sum_j C_\mu^j C_\nu^j, \quad (4)$$

where the sum runs over the occupied molecular orbitals, j, and C_μ^j is the coefficient of basis function μ in the expression of molecular orbital j.

The molecular electrostatic potential can be decomposed into parts, each corresponding to an atomic or molecular fragment of the whole molecule. A molecular fragment is a subset of atoms of a molecule whose electrostatic potential is calculated by partitioning the electronic density matrix in a manner similar to a Mulliken population analysis. The electrostatic potential of the ith fragment at point $\mathbf{r}$ can be defined by

$$V_i(\mathbf{r}) = \sum_a W_a^i \frac{Z_a}{|\mathbf{R}_a - \mathbf{r}|} - \sum_\mu \sum_\nu W_{\mu\nu}^i P_{\mu\nu} \int \frac{\varphi_\mu(\mathbf{r})\varphi_\nu(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} \mathrm{d}\mathbf{r}', \quad (5)$$

where the factor W_a^i is the nuclear decomposition factor and $W_{\mu\nu}^i$ is the electronic decomposition factor. The conditions for additivity are

$$V(\mathbf{r}) = \sum_i V_i(\mathbf{r}) \quad (6)$$

and

$$\sum_i W_{\mu\nu}^i = 1, \quad (7a)$$

$$\sum_a \sum_i W_a^i Z_a = Q, \quad (7b)$$

where Q is the total charge of the nuclei comprising the system. Equation (7a) states that the contribution from the electronic charge distribution corresponding to the basis function pair $\mu\nu$ can be divided and shared by two or more fragments. The contribution from the nuclear charge would be typically confined to one atom, but (7b) offers the possibility that nuclear charge can be redistributed among fragments with the condition that the total charge of the molecule would be preserved.

If we compute the FESP using the geometry of the target molecule, it does not matter how we partition the nuclear and electronic charge distributions because the summation over the FESPs will by definition equal the electrostatic potential of the whole molecule. However, we are interested in the chemical bonding effects for large protein molecules whose geometry is still to be determined by crystallographic refinement, and which will never be perfectly reproduced by the superposition of the smaller fragments computed from some reference structure. Therefore our goal is to define the fragments and protocols for their superposition that preserve as much transferable chemical bonding information as possible, and that result in the least amount of error according to (1) in describing the larger target system.

We have used NAAMA as a test system by considering three of its known minimum-energy conformations in the gas phase as defined by backbone dihedral angles φ and ψ. The three conformers are $C1$ ($\varphi = -129$, $\psi = 30^\circ$), $C2$ ($\varphi = -57$, $\psi = -47^\circ$) and $C3$ ($\varphi = -70$, $\psi = 70^\circ$) (Shang & Head-Gordon, 1994). We note that the last conformation exhibits an intramolecular hydrogen bond, while the first two conformations correspond to the β-sheet and α-helical regions of a Ramachandran map.

The definition of the fragments that will eventually comprise the FESP approximation to the entire protein is the first issue. We know that the coefficients C_μ^j in (4) are affected by delocalization over the whole molecular surroundings, and therefore the quality of the fragment electrostatic potential in the context of the target molecule will be sensitive to the fragment definition. In all cases, we use one of the NAAMA conformers as the target molecule whose electrostatic potential we know exactly, and which we approximate by superimposing fragmental electrostatic potentials derived from a different NAAMA conformer. The procedure for defining the electrostatic potentials of the fragments requires a definition of the decomposition factors in (7). We have chosen the following for all work reported here. Within a fragment, the nuclear decomposition factor $W_a^i = 1$ if the ith fragment contains the ath atom, and the nuclear decomposition factor is zero for all atoms involving other fragments. The electronic decomposition factors are defined as

$$\begin{aligned} W_{\mu\nu}^j &= 1.0 \text{ for both } \mu \text{ and } \nu \text{ on fragment } j, \\ &= 0.5 \text{ for only one of } \mu \text{ or } \nu \text{ on fragment } j, \\ &= 0.0 \text{ for neither } \mu \text{ nor } \nu \text{ on fragment } j, \end{aligned} \quad (8)$$

where μ and ν define the basis function of interest (Walker & Mezey, 1993).

When assembling these fragments together to provide a best fit to the electrostatic potential of the target molecule, we first optimized the fit between the target and parent fragment, based on the geometric coordinates. We then performed an electrostatic potential calculation of the properly oriented fragment on the same grid and assembled the fragmental ESPs according to (5).

The geometric fit between the target and parent fragments is accomplished by determining a rotation matrix and a translation matrix that minimizes the root mean square differences between the atomic positions of the parent fragment and the corresponding target fragment (Kabsch, 1976; MacKerell *et al.*, 1998). In the case of atomic fragments, an additional translation step is executed to ensure that the atomic positions are identical in the target and parent molecules. The fragmental ESPs are then calculated from the

parent molecule in the orientation of the target molecule, and the resulting ESPs are assembled by simple superposition according to (6). The resulting potential is Fourier transformed to give F_{calc}, while F_{obs} values are obtained from the straight *ab initio* calculation of the electrostatic potential of the target molecule. *R* factors between the modulus of the structure factor corresponding to the target molecule ESP and the sum of the molecular-fragment structure factors were calculated within resolution zones as in (1).

All *ab initio* calculations were performed with the *Q-Chem* molecular-orbital package (Kong *et al.*, 2000). The geometries of NAAMA in three different conformations and for all fragments were determined from a full geometry optimization using the Hartree–Fock (HF) method and the 6–31+G* basis set. Default convergence criteria defined in *Q-Chem* were used for all optimizations. The electrostatic potential maps for all fragments and NAAMA were generated at the HF/6–31+G* level of theory. We used a cubic grid with length of side equal to 25.4 Å, with data points sampled every 0.2 Å. We also investigated the error introduced by a finite box size and coarseness of the grid mesh by performing two additional calculations with double the box length equal to 50.8 Å and half the mesh grid distance equal to 0.1 Å.

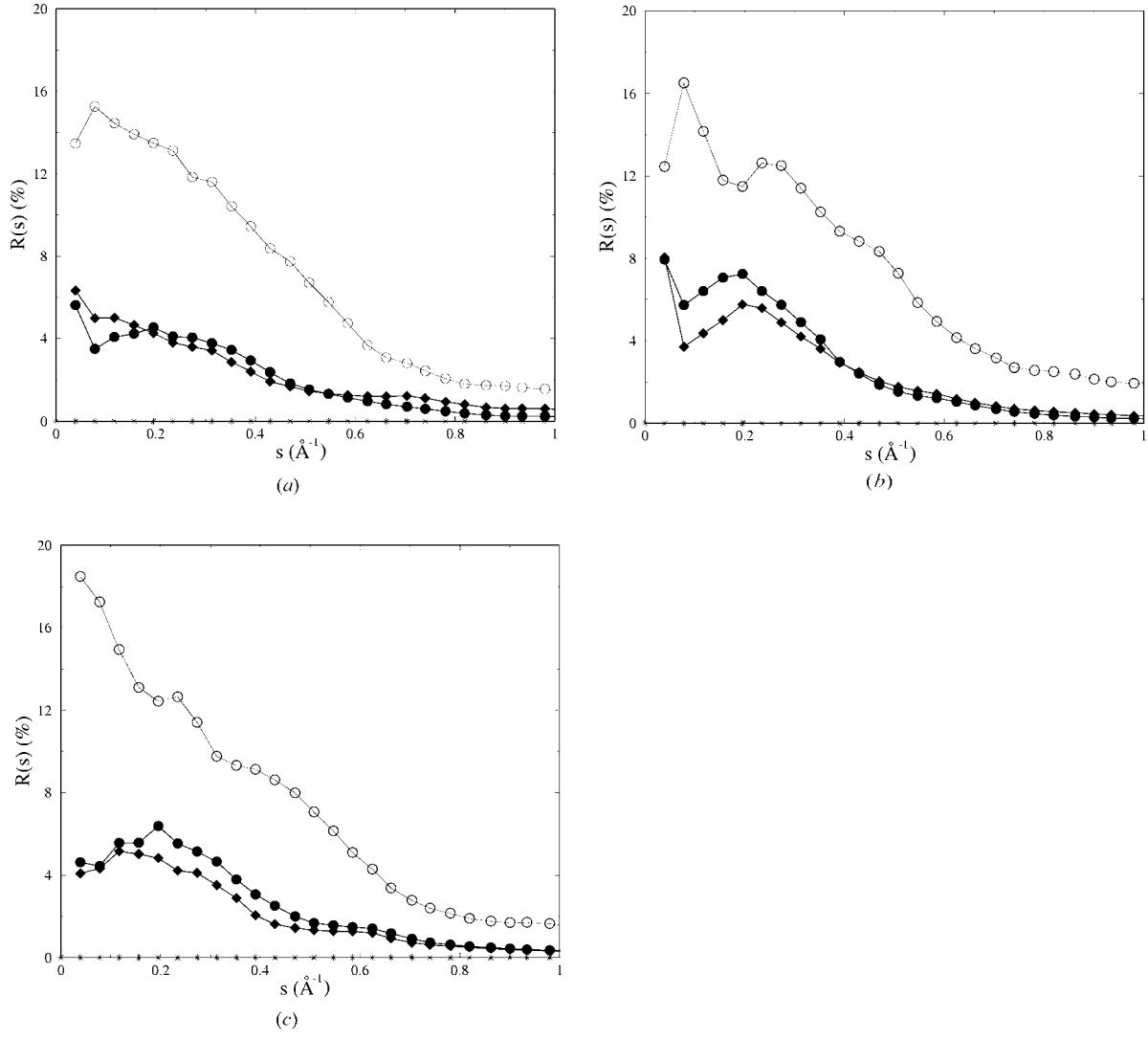

Figure 2
R factor *versus* resolution, *s*. (*a*) The error in reproducing the Fourier transforms of the electrostatic potential of the *C*1 target conformer of NAAMA using a superposition of free-atom potentials (open circles), a superposition of *C*1 atomic (fragmental) ESPs (asterisks, essentially on the baseline), a superposition of atoms derived from the molecular environment of the *C*2 parent conformer (filled diamonds) and a superposition of atoms derived from the molecular environment of the *C*3 parent conformer (filled circles). (*b*) The error in reproducing the Fourier transforms of the electrostatic potential of the *C*2 target conformer of NAAMA using a superposition of free-atom potentials (open circles), a superposition of *C*2 atomic (fragmental) ESPs (asterisks), a superposition of atoms derived from the molecular environment of the *C*1 parent conformer (filled circles) and a superposition of atoms derived from the molecular environment of the *C*3 parent conformer (filled diamonds). (*c*) The error in reproducing the Fourier transforms of the electrostatic potential of the *C*3 conformer using a superposition of free atom potentials (open circles), a superposition of *C*3 atomic (fragmental) ESPs (asterisks), a superposition of atoms derived from the molecular environment of the *C*2 parent conformer (filled circles) and a superposition of atoms derived from the molecular environment of the *C*1 parent conformer (filled diamonds).

3. Results

3.1. Atoms in molecules

First we consider the strategy of dividing the NAAMA conformers into atomic fragments of two types: free atoms and single atoms from molecules. The free-atom scattering factors are traditionally used during refinement in electron and X-ray crystallography [*International Tables for Crystallography*, Vol. C (Cowley, 1992)]. The free-atom superposition approximation provides a benchmark that is free of any influence of molecular bonding. For a free atom, both the electron density and the electrostatic potential are spherically symmetric. The 'single atoms in molecules' electrostatic potentials, on the other hand, carry information regarding charge redistribution and bond directionality due to chemical bonding. In the 'single atoms in molecules' approach, the NAAMA molecule in the parent conformations is divided into 22 different fragment types (*i.e.* the 22 different atoms that make up the molecule). *Ab initio* FESPs are calculated for each parental fragment in the position and orientation that best matches the target fragment position and bond orientations (as described in *Methods*) and then assembled by superposition to become the calculated electrostatic potential of the target molecule. The Fourier transform of the calculated electrostatic potential and the *ab initio* electrostatic potential of the target molecule (*i.e.* the 'observed' ESP) are then calculated and the *R* factors are estimated according to (1). All possible parent–target molecular combinations from the three conformers were considered.

Fig. 2 shows the *R* factor when comparing the structure factors of the NAAMA molecule in various target conformers assembled from the superposition of free atoms, their own atoms, and single atoms in molecules taken from the remaining two parent conformers. Reconstituting the target conformer from its own fragmental atoms leads to values of the *R* factors < 0.03% for all the three conformers over the whole resolution range as shown in Figs. 2(*a*), (*b*) and (*c*) (asterisks). This result constitutes a check that the fragmental 'atoms in molecules' indeed accurately reconstructs the original electrostatic potential. It is apparent from Fig. 2 that inclusion of chemical bonding information through the single atoms in molecules approach leads to a better approximation of the electrostatic potential of the target molecule than the free atoms superposition approach, over the whole resolution range, even when the conformations of the parent and target molecules are substantially different. Over the lowest resolution ranges of 0.04–0.1 Å^{-1}, the *R*-factor error is reduced to almost a third of the free-atom value (*i.e.* from 16% on average to ~6% on average). The reduction in error is mainly due to incorporating local bonding information and accounting for the long tails of the atomic electrostatic potentials.

The residual error found for the superposition of single atoms from a parent molecule to a target molecule could be a function of several factors. At low scattering angles, the electron scattering factors are strongly dependent on the net atomic charge and the mismatch of charge density on each atom in the three conformations must account for some of the residual error. However, for $s > 0.4$ Å^{-1} ($d < 2.5$ Å), the electron scattering factors for charged and neutral species are indistinguishable, so one might have expected the *R* factors to be even smaller at high resolution. We have investigated the effects of the atomic partial charges, based on the crude (but consistent for the comparison here) definition of Mulliken populations (Mulliken, 1955), which differ among the three NAAMA conformations. By matching the Mulliken charges on the parental fragments with the Mulliken charges calculated for each atom in the target conformation, the *R* factor was reduced by about 1% in the range $s < 0.015$ Å^{-1}, but it actually increased by about 1% for $s > 0.4$ Å^{-1}. Thus, it is clear

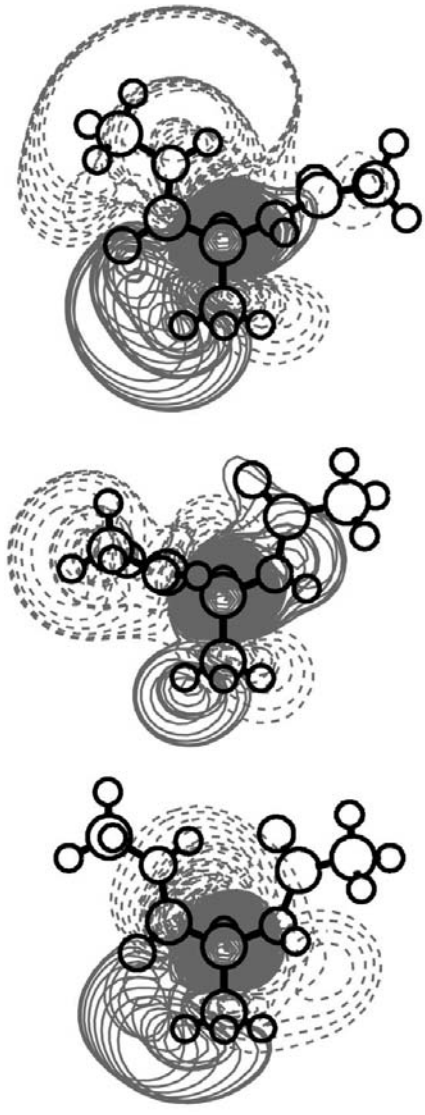

Figure 3
Atomic tail effect in the transferable FESP (TFESP) method: Electrostatic potential of the α-carbon atom in the three conformers, *C*1, *C*2 and *C*3. Each figure shows the molecule superimposed on the central eight sections of the volume contoured at an interval of 0.025 e.s.u., showing the potential distribution at a distance from the atom center. The dashed and solid gray lines represent negative and positive isopotential surfaces, respectively.

that the differences in the shape of the potential have a more significant influence on R factors than the partial charge as analyzed with Mulliken populations.

An additional main source of error in the whole resolution range is the imperfect matching of the long tails of the molecularly derived atomic electrostatic potential of the target molecule by the parental fragment. Fig. 3 illustrates the nature and extent of the tails, showing the electrostatic potential for the α-carbon atom in conformations 1, 2 and 3. Both the shape and orientation of the tails depend strongly on the atomic environment, so it is not possible to match exactly the potential at a distance from an atom in one conformation by using the FESP from a different conformation. In the medium- to low-resolution range, this effect contributes up to 1.5% of the error for atoms that have similar Mulliken charges in the parent and target conformations and ~3.5% for atoms that have different Mulliken charges.

Another source of error might be numerical errors that are introduced by a finite box size (because of the long tails of the electrostatic potential) and coarseness of the grid mesh (because the ESP gradient is very high in regions close to the atomic nuclei). To quantify these errors, we recalculated the R factors for reproducing the electrostatic potentials of the target $C2$ conformation by the approximate structure factors of single atoms derived from the parent $C1$ conformation, but with a finer mesh spacing of 0.1 Å or with a larger box size of 50.8 Å (Fig. 4). It is clear that little error is introduced with our default grid spacing size of 0.2 Å as seen in Fig. 4, where the data appear indistinguishable from the 0.1 Å mesh spacing over all resolution ranges. Doubling the size of the simulation box to 50.8 Å also results in negligible reduction of the error from high resolution to relatively low resolution.

3.2. Larger fragments

We next consider the use of larger fragments. While greater inclusion of explicit bonding could result in improvement in the approximation to the electrostatic potential of the target conformer, additional imperfections in matching the geometry of a given fragment with its corresponding component in the target will contribute to further error. We consider the case of 'bonded atoms in molecules' obtained by dividing the $C1$ parent conformer into 8 different bonded fragments, as well as the case of defining larger fragments by dividing the $C1$ parent conformer into the three fragments CH_3CONH-, $-CH(CH_3)-$ and $-CONHCH_3$, and superimposing these fragments as a best geometry fit with the $C2$ and $C3$ target conformers of NAAMA.

Fig. 5 shows that the bonded atoms in molecules case provides only negligible improvement over the use of single atoms in the medium- to low-resolution range. Predictably, larger errors arise at higher resolution because atom centers are not perfectly aligned between the bond fragments and target molecule owing to the small differences in bond lengths that occur between similar bonds in different molecular environments. This trend only becomes amplified as we consider larger fragments, where errors are now larger at low resolutions as well, while even more significant degradation in R factors is observed at the highest resolutions. Therefore, we conclude that the better fit to the data in the resolution ranges of 2.5–25.0 Å ($0.04 < s < 0.4$ Å^{-1}) that is most important in electron crystallography is with the use of atomic fragments which allows for the least amount of distortion in geometry between parental fragment and the target molecule.

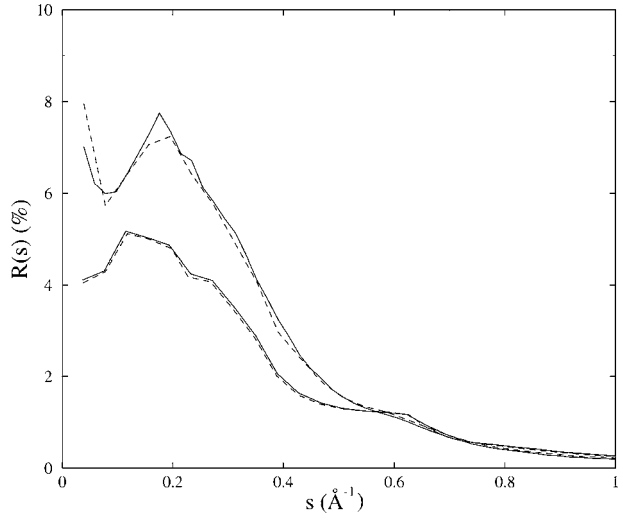

Figure 4
R factor *versus* resolution, s, calculated for the NAAMA molecule in the target $C2$ conformation using the electrostatic potentials of single atoms from the parent $C1$ conformation (top two lines) when the grid size is doubled from 25.4 (solid line) to 50.8 Å (dashed line) grid size; and in the target $C3$ conformation from single atoms in parent $C1$ conformer (bottom two lines) for the original mesh spacing of 0.2 Å (solid line) compared to the 0.1 Å mesh spacing (dashed line).

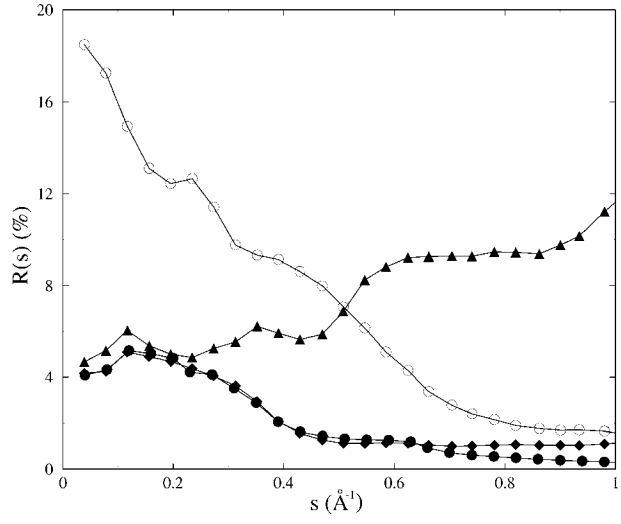

Figure 5
R factor *versus* resolution, s. The error in reproducing the Fourier transforms of the electrostatic potential of the $C3$ target conformer using a superposition of free atoms (open symbols), atoms in molecules (filled circles), bonded atoms in molecules (filled diamonds) and a superposition of larger fragments in molecule potentials (filled triangles), all derived from the parent conformation $C1$.

Table 1
Multipole moments based on electronic structure calculations of electron densities of the three different NAAMA conformers, *C*1, *C*2 and *C*3; φ and ψ angles for these conformers are given in parentheses.

	NAAMA conformer		
Multipole moment	*C*1 (−129, 30°)	*C*2 (−57, −47°)	*C*3 (−70, 70°)
Monopole	0.0	0.0	0.0
Dipole moment (Debye)	4.84	6.88	3.28
Quadrupole trace (Debye Å)	−183.55	−183.07	−184.05

4. Discussion

Calculation of electronic properties is presently feasible only for molecules containing less than a few hundred atoms. A number of methods for extending the calculations to molecules of the size of typical proteins have been proposed based on the superposition of small fragments for which the computational problem is not limiting (Walker & Mezey, 1993). In the case of X-ray crystallography, electron densities can also be obtained experimentally for small molecules when sufficiently high resolution diffraction data are available. Procedures have been developed to superimpose densities from such studies to describe properties of proteins (Pichon-Pesme *et al.*, 1995; Jelsch *et al.*, 2000). We have investigated this approach in the context of computing electrostatic potentials for proteins. Because our work is in an electron crystallographic context, our analysis of the accuracy of fitting is based on calculation of *R* factors between Fourier transforms of the potential, rather than the similarity of an isosurface view of the electron density (Walker & Mezey, 1993) or atomic volume (Bader, 1990).

Figs. 2(*a*), (*b*) and (*c*) show quantitative variations in the ability of the superimposed atomic fragments derived from molecular environments to reproduce the electrostatic potential of the target molecule between particular parent and target conformer pairs. For example, when going from free atoms to single atoms in molecules derived from the same parent *C*1 conformer, the *C*3 target conformer shows a larger improvement in *R* factor (~10–15%) as compared to the smaller (~5–10%) improvement for the *C*2 target conformer (Figs. 2*b* and 2*c*). One possible explanation for why fragments from *C*1 better reproduce the electrostatic potential for *C*3 than they do for *C*2 could be that the two peptide backbone dipoles are opposed in direction for conformations like *C*1 and *C*3, whereas they are more closely aligned in orientation for the α-helical conformer *C*2, as first discussed by Flory (1989), as well as other groups (Shang & Head-Gordon, 1994). This observation provides further insight into the origin of the remaining error as well as the variations in the *R* factor over the whole resolution range of *s* as seen in Fig. 2. It seems clear that parameterized atomic scattering factors for chemically bonded atoms will have to account for the φ and ψ angles of the local backbone residues in order to reduce much of the error that still remains in Fig. 2.

Atom descriptions derived by X-ray crystallography, including partial charges and the aspherical component of the charge distribution, were found to depend on the chemical species but to be largely independent of conformation (Pichon-Pesme *et al.*, 1995; Jelsch *et al.*, 1998). The difference with respect to our work may reflect our focus on the potential, which has a longer range than the electron density.

In a recent report (Yamashita & Kidera, 2001), it was found that a set of Gaussian functions can describe the potential for a small molecule well enough to significantly improve the *R* factor. This approach, while not dealing with the aspherical component, presumably accounts for changes in the net atomic charge but our results suggest that the conformation dependence would still need to be incorporated to allow transferability to proteins.

Our transferable FESP of single atoms in molecules (TFESP) method involves placing atom-based distributions from the parent-molecule conformer at the positions of the atoms in the target conformer. Therefore, the error in reproducing the Fourier transform of the electrostatic potential of the target molecule can be associated with the ability of the superimposed atom charge distributions to reproduce the electrostatic moments of the original charge distribution. Electrostatic multipole-moment vectors and tensors up through hexadecapoles are provided as part of the standard output from calculations using the *Q-Chem* package. Table 1 lists the monopole, the dipole moments and the trace of the quadrupole tensor, for all three conformers.

The substitution of all atoms from a charge-neutral parent molecule to a charge-neutral target conformer automatically matches the charge distribution at the level of the $q = 0$ monopole (if the grid size is large enough) and is a good approximation for the long tails of the molecular electrostatic potential. This is not the case for the free-atom superposition approach, which certainly matches the total charge, 0, but does not reproduce the long tails of the molecular electrostatic potential. In Table 1, the traces of the quadrupole moment tensors are very similar for the three conformers but there are relatively large differences among the magnitudes (as well as orientations) of the dipole moments (in Debye): $\Delta\mu_{12} = 2.04$, $\Delta\mu_{13} = 1.6$ and $\Delta\mu_{23} = 3.6$. Differences in the orientations of both dipole and quadrupole moments among the three conformations will contribute to differences in the Fourier transforms of the electrostatic potential. The relatively high values of the *R* factor at low to medium resolutions (the absolute value of which depends on the parent/target conformer pair) could be related to the differences in the total dipole and quadrupole moments between the parent/target pair but no analytical relationship is derived to quantitate these differences.

5. Conclusions

In this paper, we have considered the use of different atomic and molecular fragments to reproduce the molecular electrostatic potential of different conformations of NAAMA with an acceptable degree of error as measured by conventional *R*

factors used in the refinement procedure common in crystallography. This partition scheme for FESP is similar in spirit to the LEGO method used for the electron density (Walker & Mezey, 1993). We have evaluated three different ways of dividing NAAMA into fragments that incorporate increasingly more complete descriptions of molecular bonding with diminishing accuracy in geometric fit to the parent molecule: single atoms in molecules, bonded pairs or clusters of atoms in molecules, and functional groups such as the backbone peptide moiety and the α-carbon and amino acid side chain. Unlike the LEGO method, we find unacceptably large errors using large fragments for electron crystallography. Over the entire resolution range examined, we find that the fairly straightforward use of transferred electrostatic potentials for single atoms in molecules provides approximately 5–15% improvement in calculated R factors over the free-atom superposition approach even with a substantial mismatch between the environment of the parental and target atom fragments. No significant further improvement was found at the lowest resolutions with bonds in molecules or larger fragment descriptions, and R factors were found to degrade at higher resolutions with the use of these larger fragments because of poor geometric fits to the positions of atoms in the target molecule.

These considerations would suggest that refinement of electron crystallography data could be further enhanced by replacing free-atom atomic form factors by atom-based expansion centers that describe local chemical-bonding effects. One significant advantage in further developing the atoms in molecules approach for use in refinement for both electron and X-ray protein crystallography is that much of the refinement software will be at least partially transferable when using these suitably modified atomic form factors. Quantitative variations in the ability of the superimposed parental atomic fragments to reproduce the electrostatic potentials of the target conformer as well as conclusions drawn from the qualitative multipole analysis suggest that we can further control the R-factor error in practice by not only taking into account local chemical bonding effects but incorporating nonlocal effects of the overall charge distribution of the molecule as well. In this case, we would define an atom-centered molecular form factor that would not only include the chemical identity and the local bonding environment or valency but also information pertaining to its greater molecular environment as well. One possible way to make this feasible in practice is to compute atom-centered information for a range of conformational variables such as the energetically accessible regions of φ and ψ as represented in Ramachandran plots, and which may also depend on side-chain torsional angles.

An important issue for the near future is the appropriate mathematical basis for describing these atom-centered form factors that incorporate molecular bonding effects. These might include something as obvious as a multipole expansion, as has been used in high-resolution X-ray crystallography (Coppens, 1997), or investigations of so-called Stewart atoms (Stewart *et al.*, 1965), a way of recovering to a good approximation the atomic identity from a molecular density, and which has been tested for reproducing the molecular electrostatic potentials as well (Gill, 1996; Gilbert *et al.*, 2000).

This work has been supported by the National Institutes of Health, GM51487, and by the Office of Health and Environmental Research, US Department of Energy, under Contract DE-AC03-76F00098.

References

Bader, R. F. W. (1990). *Atoms in Molecules – a Quantum Theory*. Oxford University Press.

Chang, S., Head-Gordon, T., Glaeser, R. M. & Downing, K. H. (1999). *Acta Cryst.* A**55**, 305–313.

Coppens, P. (1997). *X-ray Charge Densities and Chemical Bonding*. New York: Oxford University Press.

Cowley, J. M. (1992). *International Tables for Crystallography*, Vol. C, edited by A. J. C. Wilson, pp. 223–245. Dordrecht: Kluwer Academic Publishers.

Flory, P. J. (1989). *Statistical Mechanics of Chain Molecules*. Munich: Hanser.

Gilbert, A. T. B., Lee, A. M. & Gill, P. M. W. (2000). *J. Mol. Struct. Theochem.* **500**, 363–374.

Gill, P. M. W. (1996). *J. Phys. Chem.* **100**, 15421–15427.

Glaeser, R. M. (1999). *J. Struct. Biol.* **128**, 3–14.

Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. (1996). *J. Mol. Biol.* **259**, 393–421.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckman, E. & Downing, K. H. (1990). *J. Mol. Biol.* **213**, 899–929.

Jelsch, C., Pichon-Pesme, V., Lecomte, C. & Aubry, A. (1998). *Acta Cryst.* D**54**, 1306–1318.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 3171–3176.

Johnson, B. G., Gill, P. M. W. & Pople, J. A. (1993). *Chem. Phys. Lett.* **206**, 239–246.

Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.

Kimura, Y., Vassylyev, D. G., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T. & Fujiyoshi, Y. (1997). *Nature (London)*, **389**, 206–211.

Kong, J., White, C. A., Krylov, A. I., Sherrill, D., Adamson, R. D., Furlani, T. R., Lee, M. S., Gwaltney, S. R., Adams, T. R., Ochsenfeld, C., Gilbert, A. T. B., Kedziora, G. S., Rassolov, V. A., Maurice, D. R., Nair, N., Shao, Y. H., Besley, N. A., Maslen, P. E., Dombroski, J. P., Daschel, H., Zhang, W. M., Korambath, P. P., Baker, J., Byrd, E. F. C., Van Voorhis, T., Oumi, M., Hirata, S., Hsu, C. P., Ishikawa, N., Florian, J., Warshel, A., Johnson, B. G., Gill, P. M. W., Head-Gordon, M. & Pople, J. A. (2000). *J. Comput. Chem.* **21**, 1532–1548.

Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. (1994). *Nature (London)*, **367**, 614–621.

MacKerell, A. D. Jr, Brooks, B., Brooks, C. L. III, Nilsson, L., Roux, B., Won, Y. & Karplus, M. (1998). *The Encyclopedia of Computational Chemistry 1*, edited by P. v. R. Schleyer, pp. 271–277. Chichester: John Wiley.

Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. (1999). *J. Mol. Biol.* **288**, 765–786.

Mulliken, R. S. (1955). *J. Chem. Phys.* **23**, 1833–1841.

Nogales, E., Wolf, S. G. & Downing, K. H. (1998). *Nature (London)*, **391**, 199–203.

Pichon-Pesme, V., Lecomte, C. & Lachekar, H. (1995). *J. Phys. Chem.* **99**, 6242–6250.

Politzer, P. & Truhler, D. G. (1981). Editors. *Chemical Applications of Atomic and Molecular Electrostatic Potentials*. New York: Plenum Press.

Ren, G., Cheng, A., Reddy, V., Melnyk, P. & Mitra, A. K. (2000). *J. Mol. Biol.* **301**, 369–387.
Shang, H. S. & Head-Gordon, T. (1994). *J. Am. Chem. Soc.* **116**, 1528–1532.
Stewart, R. F., Davidson, E. R. & Simpson, W. T. (1965). *J. Chem. Phys.* **42**, 3175.
Szabo, A. & Ostlund, N. S. (1996). *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover: Dover Press.
Walker, P. D. & Mezey, P. G. (1993). *J. Am. Chem. Soc.* **115**, 12423–12430.
Yamashita, H. & Kidera, A. (2001). *Acta Cryst.* A**57**, 518–525.