electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

Improving the scattering-factor formalism in protein refinement: application of the University at Buffalo Aspherical-Atom Databank to polypeptide structures

Anatoliy Volkov, Marc Messerschmidt and Philip Coppens

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Improving the scattering-factor formalism in protein refinement: application of the University at Buffalo Aspherical-Atom Databank to polypeptide structures

Anatoliy Volkov,* Marc Messerschmidt and Philip Coppens

Department of Chemistry, University at Buffalo, State University of New York, Buffalo, NY 14260-3000, USA

Correspondence e-mail: volkov@chem.buffalo.edu

Received 18 September 2006
Accepted 24 October 2006

The University at Buffalo theoretical databank of aspherical pseudoatoms has been tested in the refinement of high-resolution (HR; $d_{max} \leq 0.44$ Å) and truncated 'low-resolution' (LR; $d_{max} = 0.83$ Å) X-ray diffraction data sets from the tripeptide Tyr-Gly-Gly monohydrate [Pichon-Pesme *et al.* (2000), *Acta Cryst.* B**56**, 728–737] and hexapeptide cyclo-(D,L-Pro)$_2$-(L-Ala)$_4$ monohydrate [Dittrich *et al.* (2002), *Acta Cryst.* B**58**, 721–727]. Application of the databank to LR data significantly lowers the conventional *R* factor, improves the determination of bonds and angles to within 0.002–0.003 Å and 0.09–0.17° of the values obtained from a complete multipolar refinement of HR data sets, improves the determination of phase angles by 2–6° compared with the standard independent atom refinement (IAM), removes the majority of the bonding features from the residual Fourier difference maps and improves the atomic displacement parameters (ADPs) and the results of the Hirshfeld rigid-bond test. In the description of the aspherical density from experimental X-ray data, theoretical pseudoatoms were found to perform on the same level as the previously reported experimental databank [Pichon-Pesme *et al.* (1995), *J. Phys. Chem.* **99**, 6242–6250; Jelsch *et al.* (1998), *Acta Cryst.* D**54**, 1306–1318], although no direct comparison of the two methods has been performed. The theoretical databank of aspherical pseudoatoms is shown to be a significant aid in the refinement of accurate experimental X-ray data from large molecular systems, in addition to its use in the reconstruction of molecular densities and the determination of electrostatic interaction energies.

1. Introduction

Recent advances in experimental techniques allow the collection of X-ray diffraction data from macromolecular crystals of unprecedented quality and with 'ultrahigh resolution' ($d \simeq 0.60$ Å, $\sin\theta/\lambda \simeq 0.83$ Å^{-1}). In the last few years, several such structures have been reported: Z-DNA CGCGCG with $d_{min} = 0.60$ Å (Tereshko *et al.*, 2001), RNA tetraplex (UGGGGU)$_4$ with $d_{min} = 0.61$ Å (Deng *et al.*, 2001), eel pout type III antifreeze protein RD1 with $d_{min} = 0.62$ Å (Ko *et al.*, 2003), *Pyrococcus abyssi* rubredoxin with $d_{min} = 0.69$ Å (Bönisch *et al.*, 2005), crambin with $d_{min} = 0.54$ Å (Jelsch *et al.*, 2000) and human aldose reductase with $d_{min} = 0.66$ Å (Lamour *et al.*, 1999; Cachau *et al.*, 2000; Howard *et al.*, 2004). At this 'subatomic' resolution, the aspherical bonding features of the electron density (ED) were reported to be visible from the Fourier difference maps (Afonine *et al.*, 2004). This becomes possible because the extension of the data to higher resolution improves the deconvolution of thermal and bonding effects. Meaningful deconvolution of the two effects is not feasible when only LR data are available, because in the

high d, low $\sin\theta/\lambda$ data the thermal motion effect is mixed with the scattering from the aspherical component of the electron density. The independent (spherical) atom model (IAM) used in the conventional refinement of X-ray data introduces a bias in the geometrical and atomic anisotropic displacement parameters (ADPs) which, under the condition of the least-squares fit, partially describe the aspherical bonding density, as realised early in the development of X-ray charge-density analysis (Coppens, 1967).

In the accurate X-ray analysis of small molecules, the deconvolution of thermal motion and bonding density is achieved by (i) using high-order data ($\sin\theta/\lambda > 0.8$ Å^{-1}, $d < 0.63$ Å) and (ii) aspherical modeling of the ED. The most frequently used aspherical model of the ED in molecular crystals is given by the Hansen–Coppens formalism (Hansen & Coppens, 1978; Coppens, 1997), which describes the static ED by a superposition of aspherical pseudoatoms represented by nuclei-centered density expansions,

$$\rho_k(\mathbf{r}) = P_c\rho_c(r) + P_v\kappa^3\rho_v(\kappa r) + \kappa'^3\sum_{l=1}^{4} R_l(\kappa' r)\sum_{m=1}^{l} P_{lm\pm}d_{lm\pm}(\mathbf{r}/r),$$

where ρ_c and ρ_v are spherically averaged free-atom Hartree–Fock core and valence densities (Clementi & Roetti, 1974) normalized to one electron, respectively, $d_{lm\pm}$ are density-normalized real spherical harmonics and R_l are Slater-type radial density functions (Slater, 1932),

$$R_l(\kappa' r) = \kappa'^3\frac{\zeta^{n_l+3}}{(n_l+2)!}(\kappa' r)^{n_l}\exp(-\kappa'\zeta r),$$

with energy-optimized exponents ζ (Clementi & Raimondi, 1963). The dimensionless expansion–contraction parameters κ and κ', along with the populations P_v and $P_{lm\pm}$, are refined in the fitting procedure against experimental structure-factor amplitudes, while the populations P_c of the core shells remains fixed.

The combination of the aspherical model of the electron density and high-resolution data in general successfully deconvolutes the thermal motion effects from the bonding density, which leads to more accurate molecular geometries and ADPs. Although the method works well for small molecules, its application to macromolecules is hampered by several factors: (i) high-resolution data are generally not available, (ii) the number of reflections available is usually not sufficient for a full aspherical atom refinement and (iii) the quality of X-ray data for macromolecules is generally lower than for small molecules. The latter is being addressed by improved crystallization and data-collection techniques. The first two can be circumvented by improvements in the scattering model, which are the subject of this article.

In the mid 1990s, a new approach based on the idea of transferability of aspherical atomic parameters of the ED was developed to aid in the refinement of accurate macromolecular X-ray data. Based on the multipolar analysis of several high-resolution X-ray diffraction data sets collected from relatively small peptides, Lecomte and coworkers (Pichon-Pesme *et al.*, 1995) found that the aspherical parameters in the Hansen–Coppens model are transferable between different molecules for atoms in similar chemical environments. This prompted the construction of the databank of experimental aspherical pseudoatom parameters (Pichon-Pesme *et al.*, 1995, 2004). A number of successful applications of the databank to the refinement of accurate X-ray data of biologically important systems have been reported, including the octapeptide LBZ (Jelsch *et al.*, 1998), NAD$^+$–β-nicotinamide adenine dinucleotide complex (Guillot *et al.*, 2003), human aldose reductase (Muzet *et al.*, 2003), crambin (Fernandez-Serra *et al.*, 2000; Jelsch *et al.*, 2000) and scorpion toxin (Lecomte *et al.*, 2004, 2005).

As an alternative to the experimental approach, molecular electron densities can be obtained from first-principles calculations, which is the method pursued in our work (Koritsanszky *et al.*, 2002; Volkov, Li *et al.*, 2004). In more recent studies, theoretical densities have also been used by Dittrich *et al.* (2005) to generate transferable atoms, labeled invariooms, defined by their hybridization and bond types. Use of the invariooms gives a significant improvement in the refinement of the experimental data from D,L-serine at various resolutions and at different temperatures. The authors conclude that a resolution better than $\sin\theta/\lambda > 0.5$ Å^{-1} ($d <$ 1.0 Å) is needed for a successful refinement. It was also shown (Dittrich *et al.*, 2006) that inclusion of invariooms in the refinement of experimental X-ray diffraction data improves the precision of the Flack parameter and therefore the reliability of deducing molecular chirality in the absolute structure determination. To the best of our knowledge, no further examination of the densities produced by the invariom model is available.

In our studies, the parameters of chemically unique pseudoatoms are derived directly from the theoretical densities of a large number of small molecules (Volkov, Li *et al.*, 2004; Dominiak *et al.*, 2007). This procedure parallels that applied in the experimental databank; however, it involves the fitting of theoretical structure factors with known phases obtained *via* Fourier transform of the wavefunction-based density. The method can lead to parameters free of bias, which is practically unachievable for experimental estimates owing to the lack of phase information, often incomplete treatment of thermal smearing and systematic experimental errors. Furthermore, the simulation allows a great variety of atom types and systems to be studied, as the incorporation of new atom types into the theoretical databank requires much less effort than in the experimental case.

Following this approach, we have built an extended database for C, H, N, O and S pseudoatoms applicable to construction of the electron density of proteins (Volkov, Li *et al.*, 2004; Dominiak *et al.*, 2007). Atoms exhibiting the same local structure (as defined by connectivity, bonding and geometry) are considered to be chemically equivalent and their parameters averaged. The latest version of the University at Buffalo databank (UBDB) contains (as of September 2006) 104 atoms types, including all common atom types encountered in peptides, proteins and some other biologically interesting molecules (Dominiak *et al.*, 2007).

In recent studies, we found the level of accuracy in prediction of both local and integrated properties of molecular electron densities constructed with UBDB to be comparable with that from various first-principles calculations at different levels of theory and with different basis-set expansions of the wavefunction (Volkov, Li *et al.*, 2004; Volkov, Koritsanszky, Li *et al.*, 2004). The electrostatic interaction energies between amino-acid molecules obtained from the combination of UBDB and the recently developed EPMM method (Volkov, Koritsanszky & Coppens, 2004) are usually within 4 kJ mol^{-1} of theoretical Density Functional (Hohenberg & Kohn, 1964) calculations at the B3LYP/6-31G** level (Volkov, King *et al.*, 2006). In the latest study, our databank was successfully applied to the calculation of interaction energies between glycopeptide antibiotics and substrates (Li *et al.*, 2006).

As the University at Buffalo databank produces physically meaningful electron densities and related properties, its application to the refinements of experimental X-ray data is a logical extension of the work.

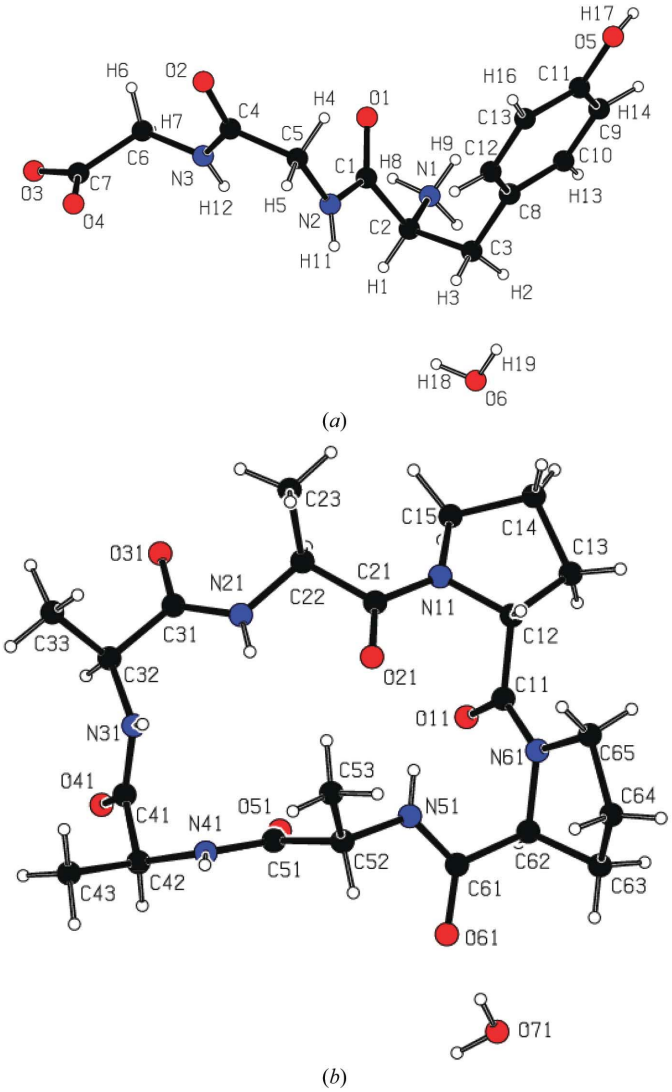

Figure 1
Molecular structures of (*a*) YGG and (*b*) P2A4. Diagrams were created with the program *PLATON* (Spek, 2003).

2. Benchmark systems, data sets and refinements

Two accurate polypeptide data sets were chosen as benchmarks. Point-detector Mo $K\alpha$ X-ray diffraction data on the tripeptide Tyr-Gly-Gly monohydrate (YGG; Fig. 1*a*) were taken from the low-temperature (123 ± 2 K) study of Pichon-Pesme *et al.* (2000). Synchrotron X-ray diffraction data for the hexapeptide cyclo-(D,L-Pro)$_2$-(L-Ala)$_4$ monohydrate (P2A4; Fig. 1*b*), collected at λ = 0.5583 Å (Hasylab/DESY) using a Bruker SMART 1K CCD, were taken from a recent low-temperature (100 ± 1 K) study by Dittrich *et al.* (2002). Two very different data sets (X-ray tube and scintillation counter *versus* synchrotron radiation and CCD camera) collected by two different groups were intentionally chosen in order to avoid possible bias in the data collection and processing.

Two groups of reflections were used for each of the systems: a high-resolution (HR) set including all reflections up to the maximum reported resolution ($\sin\theta/\lambda_{max}$ = 1.15 Å^{-1} for YGG and 1.32 Å^{-1} for P2A4) and a low-resolution[1] (LR) subset truncated at $\sin\theta/\lambda$ = 0.6 Å^{-1} ($d \simeq 0.83$ Å). The HR data sets consisted of 4766 and 21 475 structure factors, while LR data were limited to 1358 and 2513 reflections for YGG and P2A4, respectively.

The following refinements were performed using the *XD*2006 suite of programs (Volkov, Macchi *et al.*, 2006).

Refinement 1. The overall scale factor (OSF), positional parameters for all atoms (including H atoms) and anisotropic and isotropic ADPs for non-H and H atoms, respectively, were varied in a conventional spherical atom refinement.

Refinement 2. Coordinates of all non-H atoms were fixed at those obtained in refinement 1. H atoms were extended along experimental X—H directions to standard neutron distances (*International Tables For Crystallography*, 1992, Vol. *C*, Kluwer Academic Publishers; Allen, 1986). Aspherical pseudoatom parameters from the UBDB were assigned to all atoms using the program *LSDB* (Volkov, Li *et al.*, 2004). Both polypeptide and water molecules were treated as neutral. Populations P_v of atomic valence shells were rescaled using the formula

$$P_{v_i}^{\text{scaled}} = P_{v_i} + \left[\frac{\sum_i Z_i - \sum_i P_{v_i}}{\sum_i \sigma(P_{v_i})}\right]\sigma(P_{v_i})$$

(Faerman & Price, 1990; Volkov, Li *et al.*, 2004), where Z_i is the number of valence electrons in the free ith atom and $\sigma(P_{v_i})$ is the standard deviation of the valence population calculated during construction of the databank. Only the OSF was refined.

Refinement 3. As refinement 2, but with refinement of the positional and U_{ij} parameters of non-H atoms and U_{iso} of H atoms. The riding model was applied to H atoms, *i.e.* shifts in coordinates of the parent atom were applied to the coordinates of all attached H atoms. This refinement type is essentially identical to refinement II applied to the octapeptide

[1] Note that this cutoff is usually referred to as ultrahigh resolution in macromolecular crystallography, but as low resolution in accurate small-molecule studies. The latter description is used in this paper.

(Jelsch *et al.*, 1998) and crambin (Jelsch *et al.*, 2000) data using the experimental databank.

Refinement 4. As refinement 3, but with refinement of the spherical valence-shell population parameters P_v of all atoms (including H atoms) and κ parameters of non-H atoms. Chemical constraints were applied to chemically similar atoms. An electroneutrality constraint was applied to both the polypeptide and the water molecules, *i.e.* no charge transfer was allowed between the two entities.

Refinement 5. The κ'-restricted multipole refinement (KRMM; Abramov *et al.*, 1999) was applied to the HR data sets only. κ' parameters for both H and non-H atoms, as well as the κ parameter for H atoms, were fixed at theoretically determined values (Volkov *et al.*, 2001). The multipolar expansion was truncated at the hexadecapolar level ($l_{max} = 4$) for the non-H atoms and at the quadrupolar level ($l_{max} = 2$) for H atoms, for which only bond-directed functions with $l, m = 1$, 0 and 2, 0 were refined. In order to reduce the number of least-squares variables, local-symmetry constraints were imposed for some atoms. As in refinement 4, electroneutrality constraints were applied separately to the polypeptide and water molecules.

These refinements are labeled 1–5 throughout the paper. Refinements 1–5 were applied to the HR and 1–4 to the LR sets, respectively. Note that in refinements 2, 3 and 4 deformation density parameters were not refined but fixed at values supplied by the UBDB.

3. Results and discussion

The results of all refinements are summarized in Tables 1–4 and Figs. 2–8 and S1–S14 (supplementary material[2]). In the absence of accurate neutron data, the more comprehensive KRMM refinement 5 is used for reference purposes. Results are analyzed in terms of the conventional R factor, geometrical parameters (bond lengths, angles and torsion angles), ADPs, phases of acentric reflections, residual Fourier maps and the Hirshfeld rigid-bond test (Hirshfeld, 1976). According to the rigid-bond test, refined ADPs are deemed to be physically meaningful if differences in mean-squared displacement amplitudes (DMSDA) along interatomic vectors are $\leq 1 \times 10^{-3}$ Å^2.

3.1. Refinement 1 (IAM)

IAM refinements of HR data sets for both YGG and P2A4 lead to Fourier difference maps which clearly show all the bonding features, including lone pairs of O atoms and bonding features of X—H bonds (Figs. 2, 3 and 4, and S1–S9). However, there are pronounced differences in the residual density at the nuclear positions of non-H atoms between the YGG and P2A4 data sets. The former show positive residual density at these positions, while in the latter the difference densities at the nuclear positions are generally negative. These differences are especially pronounced for the atoms of the peptide bonds (Figs. S1–S3[2] for YGG, and Figs. 4 and S4–S8[2] for P2A4). The agreement of the bond lengths and angles with those from the KRMM refinement is very good, especially for P2A4, which is obviously a consequence of the presence of very high order data (P2A4 and YGG data extend to 1.32 Å^{-1} and 1.15 Å^{-1}, respectively). The unweighted root-mean-

Table 1
R factors (%) and reflections-to-variables ratio (in parentheses) for all refinements.

	YGG		P2A4	
Refinement	HR	LR	HR	LR
1	4.51 (17.3)	2.16 (4.9)	3.44 (46.7)	2.98 (5.5)
2	3.93 (4766)	2.70 (1358)	2.83 (21475)	3.28 (2513)
3	3.66 (21.9)	1.22 (6.2)	2.67 (61.0)	1.84 (7.1)
4	3.57 (19.0)	1.11 (5.4)	2.60 (57.6)	1.80 (6.7)
5	3.42 (10.6)	—	2.53 (43.6)	—

Table 2
R.m.s. differences of bond lengths, bond angles and torsion angles relative to KRMM refinement of HR data.

Note that the parameters from refinements 2 and 1 are identical.

	YGG		P2A4	
Refinement	HR	LR	HR	LR
Bond lengths (Å)				
1	0.003	0.005	0.0008	0.005
3	0.002	0.003	0.0004	0.002
4	0.002	0.003	0.0004	0.002
Bond angles (°)				
1	0.14	0.24	0.04	0.25
3	0.13	0.17	0.01	0.09
4	0.11	0.16	0.01	0.09
Torsion angles (°)				
1	0.19	0.23	0.06	0.23
3	0.12	0.16	0.04	0.09
4	0.11	0.15	0.04	0.09

Table 3
R.m.s. of differences in mean-squared displacement amplitudes (DMSDA) ($\times 10^4$ Å^2) along interatomic vectors for all refinements.

Note that DMSDA from refinements 2 and 1 are identical.

	YGG		P2A4	
Refinement	HR	LR	HR	LR
1	8.77	17.76	3.67	15.64
3	7.38	12.85	2.65	7.09
4	8.04	14.68	2.36	7.00
5	6.38	—	3.09	—

Table 4
R.m.s. differences in phase angles (°) of acentric reflections between refinements.

	YGG		P2A4	
Refinements	HR	LR	HR	LR
1/5	2.33	2.36	4.84	5.24
2/5	1.59	2.42	4.09	5.06
3/5	0.84	1.03	3.94	3.99
4/5	0.79	0.95	3.93	3.98
1/3	2.03	2.06	6.15	6.32
3/4	0.41	0.42	0.31	0.32

[2] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5093). Services for accessing this material are described at the back of the journal.

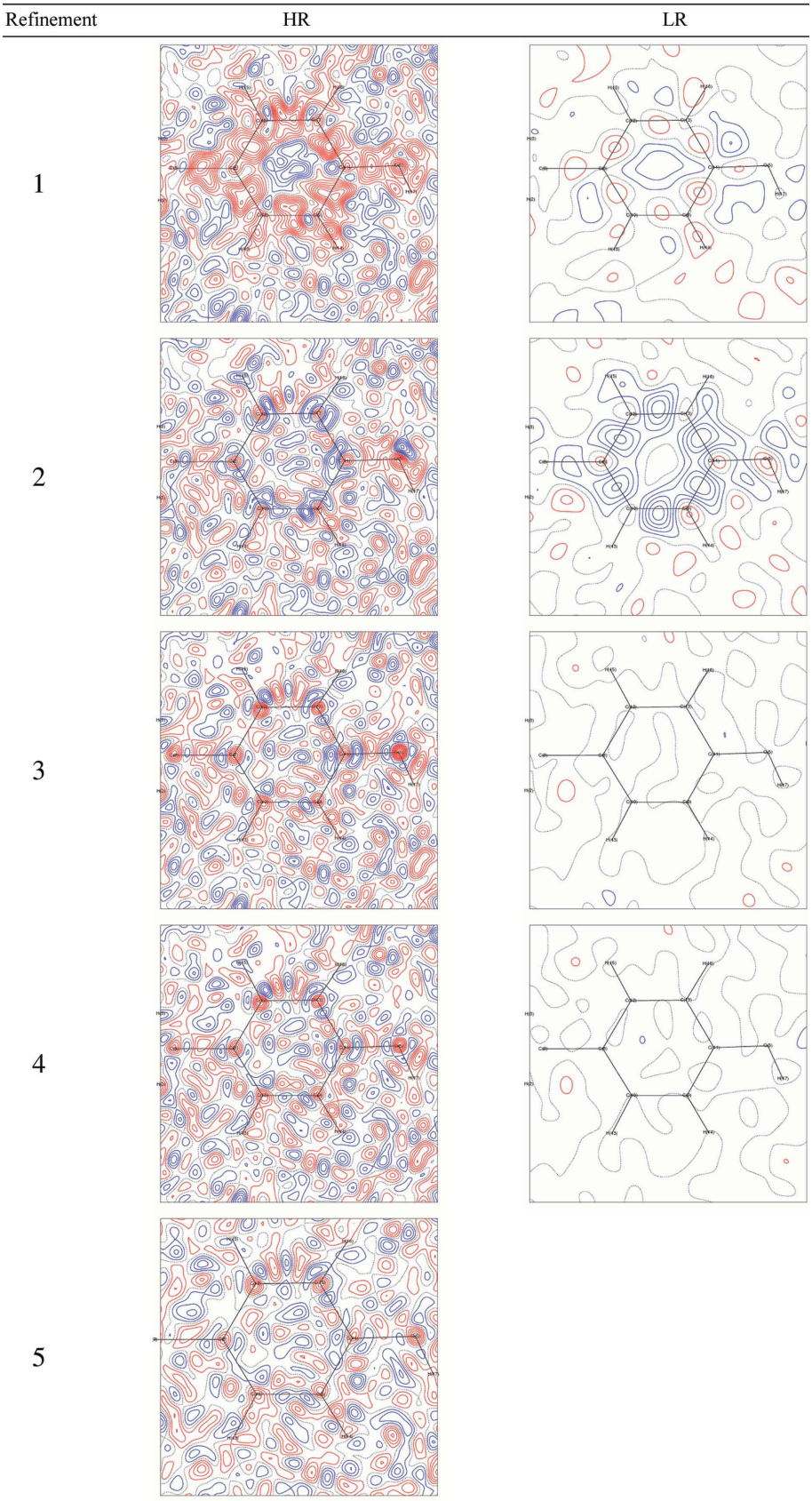

Figure 2
Residual Fourier difference maps in the plane of the phenyl ring in YGG. Contour interval 0.05 e Å^{-3}. Positive, red; negative, blue.

square (r.m.s.) values of the DMSDA are also very low: 8.77 × 10^{-4} and 3.67 × 10^{-4} Å^2 for YGG and P2A4, respectively (Table 2). Removal of the high-order data, *i.e.* refinement of LR data sets, has a dramatic impact on the results. The residual Fourier maps become essentially featureless for both YGG and P2A4. Unlike in the HR data sets, no features in the residual density at the nuclear positions are observed with either data set. Relative to the HR refinement, the *R* factor decreases by over 2% in YGG and by about 0.5% in P2A4. However, the r.m.s. of the DMSDA significantly increases: to 17.76 × 10^{-4} and 15.64 × 10^{-4} Å^2 for YGG and P2A4, respectively. The deviations of the geometrical parameters from the KRMM values also become larger. It is interesting to note that, unlike the HR results, the LR results for both YGG and P2A4 show almost identical r.m.s. deviations from the KRMM geometry: 0.005 Å for bond lengths and 0.23–0.25° for bond and torsion angles. Fig. 7 also shows that the absolute values of ADPs from IAM refinements of the LR data deviate much more from KRMM values than those from IAM refinements of HR data (in P2A4 they are almost indistinguishable). While the scattering of the bonding density must still be present in the LR data, the absence of high-order reflections does not allow proper deconvolution of the bonding density and anisotropic displacement effects. In the least-squares refinements of LR data atomic displacement parameters effectively absorb the bonding features, as is evident in the difference maps of refinement 2, discussed in the next section. This also explains the significant difference in the OSF between IAM refinements of HR and LR data sets, with the LR OSF being about 5% larger for both YGG and P2A4, in agreement with earlier experimental measurements of the scale factor for a number of crystals (Stevens & Coppens, 1975). Unlike in refinements of the LR data, the OSF from the IAM refinement of the HR data set is very close to that from the reference KRMM refinement. For YGG the r.m.s. difference in the phase angles of reflections relative to

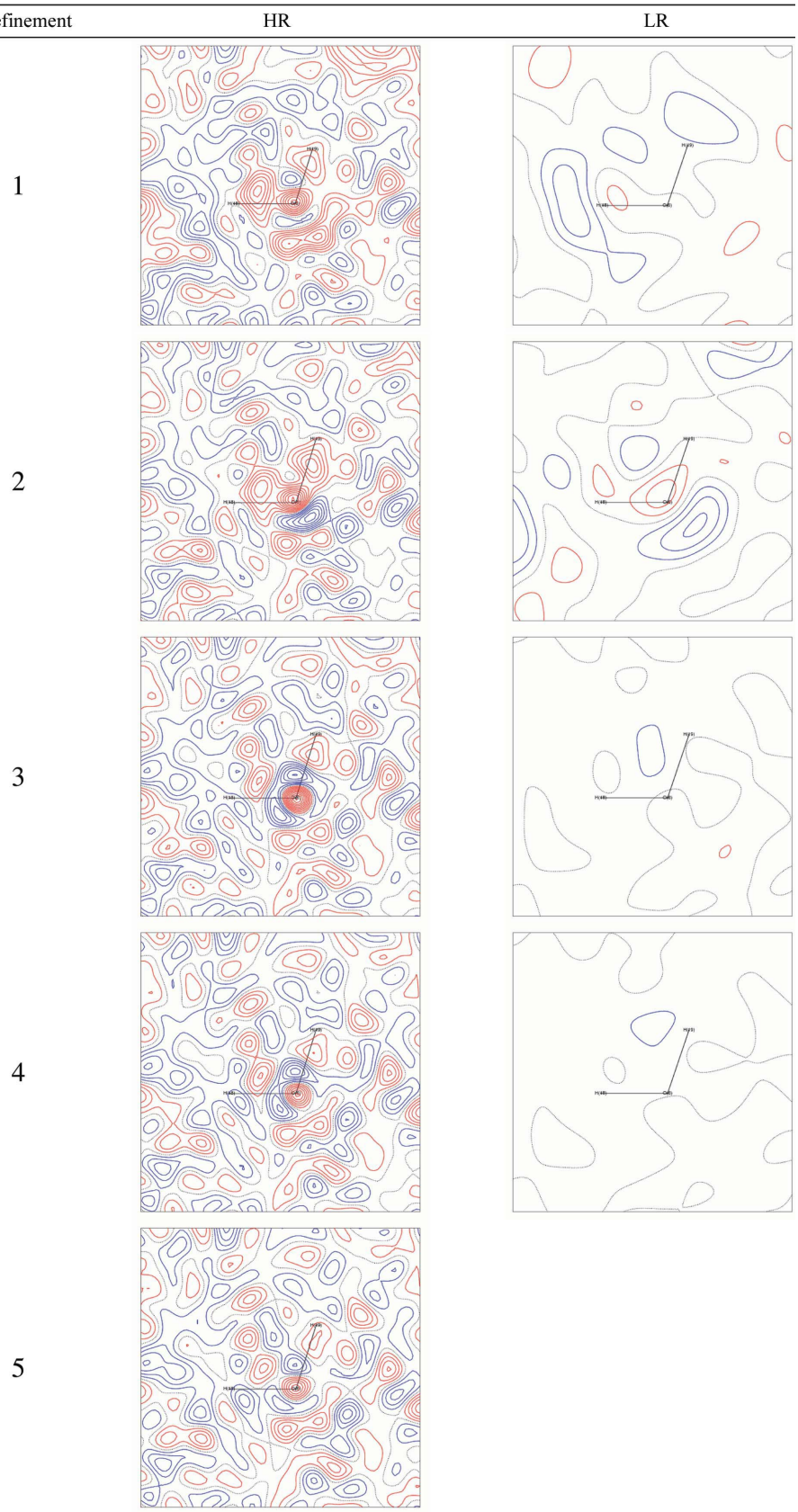

Figure 3
Residual Fourier difference maps in the plane of the water molecule in YGG. Contour interval 0.05 e Å^{-3}. Positive, red; negative, blue.

KRMM values is about the same for both HR and LR data (∼2.3°), while in P2A4 the phases from the HR refinement agree better with the KRMM results than the LR phases.

3.2. Refinement 2

Introduction of UBDB pseudoatom parameters to the structure after IAM refinement (with H atoms extended to neutron distances) lowers the *R* factor for HR data sets by about 0.5%. The residual density in the bonding regions in the Fourier difference maps is significantly reduced. These results are in agreement with previous studies using the experimental databank in which, on application of the databank parameters, the *R* factor decreased from 7.13 to 6.51% for the octapeptide (Jelsch *et al.*, 1998) and to 11.05% from 11.45% for aldose reductase (Muzet *et al.*, 2003). However, with the LR data the *R* factor increases by about 0.5% and residual Fourier maps show significant depletion of density (0.20–0.25 e Å^{-3}) in the bonding regions, which is especially noticeable in the plane of the phenyl ring in YGG (Fig. 2). This 'inverse-bonding' feature and the absence of bonding density in the IAM residual maps illustrate remarkably well that without high-order data the aspherical bonding density is effectively described by the bias in the ADPs. Nevertheless, the scale factor in the LR results improves slightly, with a decrease of about 1–1.5%. As expected, the application of databank parameters changes the phases of the reflections. In general, phases become somewhat closer to those from KRMM refinement for both HR and LR data sets, with the exception of the LR data set of YGG for which the r.m.s. difference becomes slightly larger compared with the IAM refinement.

3.3. Refinement 3

Refinement 3 is the most important type of refinement performed in this study. All UBDB pseudoatom parameters are fixed, while the OSF, ADPs and positional parameters (of non-H atoms only) are refined. A drop in *R* factor of about 1% relative to the

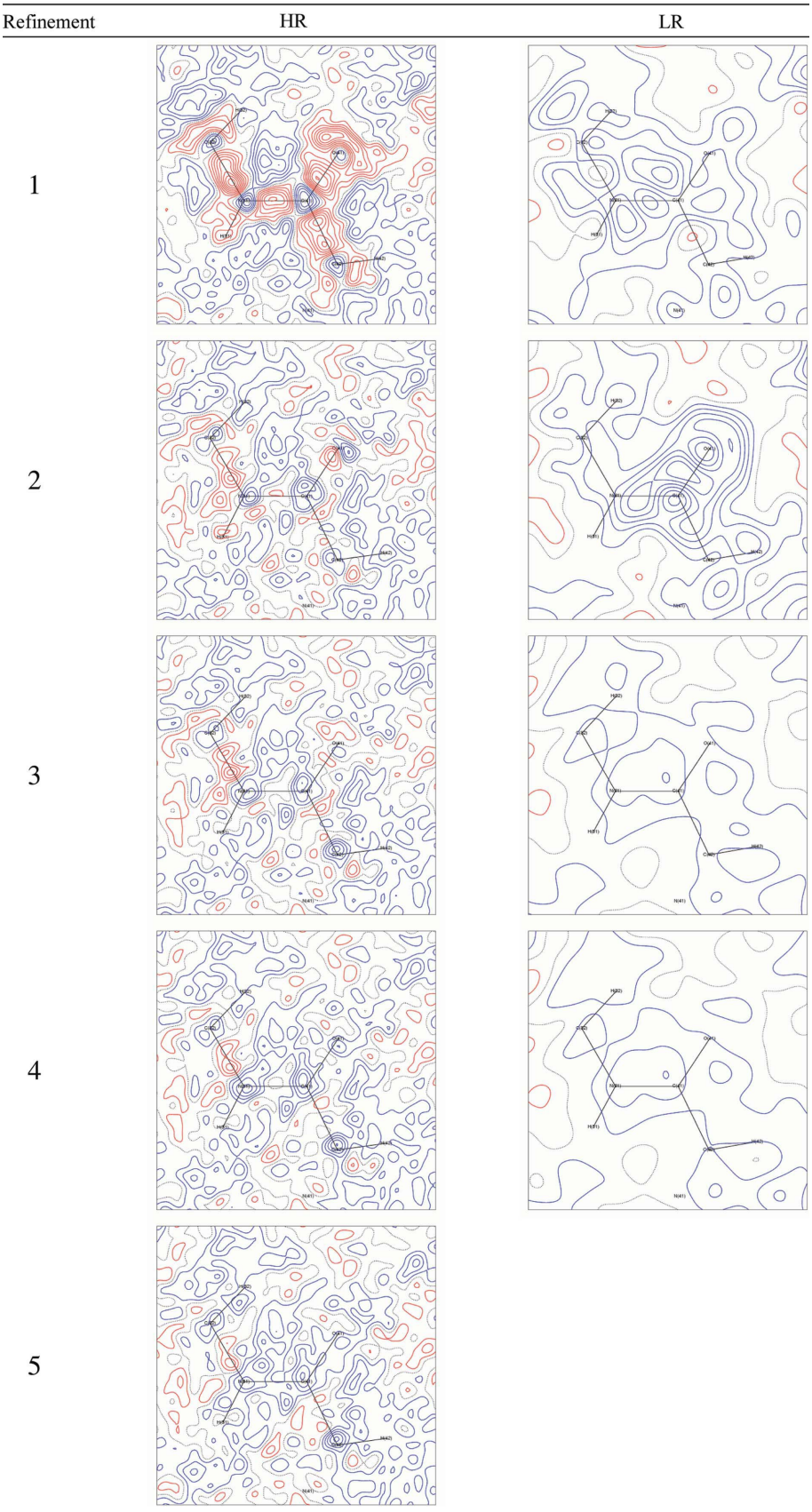

Figure 4
Residual Fourier difference maps in the plane of the peptide bond N31—C41—O41 in P2A4. Contour interval 0.05 e Å^{-3}. Positive, red; negative, blue.

conventional IAM refinement is observed for both HR and LR data sets. This is in an excellent agreement with previous studies using both the experimental databank and theoretical invarioms. For example, the *R* factor decreased by 0.92% in the octapeptide (Jelsch *et al.*, 1998), by ∼1% in (*E*)-2,2′-dimethylstilbene (Jelsch *et al.*, 2005), by 0.5% in aldose reductase (Jelsch *et al.*, 2005) and by 0.8% in HR and 1.4–1.5% in LR data sets of D,L-serine (Dittrich *et al.*, 2005). Note that both our study and the work by Dittrich *et al.* (2005) show more pronounced improvement in the *R* factor for LR data than for HR data. As the small-molecule LR cutoff is typical for the better macromolecular data sets, this is an important conclusion. Residual density maps from LR refinements become very clean (the most significant features are about 0.05 e Å^{-3}) for both YGG and P2A4, which is again in accord with the results for D,L-serine (Dittrich *et al.*, 2005). Residual maps from the HR refinement of YGG still show some significant positive spherical features at the positions of the nuclei of non-H atoms (about 0.3 e Å^{-3} and larger), while these features are not present in the P2A4 maps. This is especially evident from the comparison of maps in the plane of the water molecule (Figs. 3 and S9). Overall, the residual density maps from the HR refinement 3 are much cleaner for P2A4 than for YGG, while the opposite is observed with the LR data. Note that these spherical features in the YGG maps remain present at a reduced magnitude in refinements 4 and 5 in which P_v and κ parameters of the spherical valence shell are refined. Overall, the residual maps from this type of refinement are of the same quality as those obtained using the experimental databank (Jelsch *et al.*, 1998) and theoretical invarioms (Dittrich *et al.*, 2005), and are close to those from a complete multipolar refinement 5 (KRMM). As expected, significant changes are observed for the OSF in the LR refinement. Its reduced value is within 1–2% of the HR KRMM scale factor, compared with a difference of about 5% for the standard LR IAM refinement. Geometry bias is reduced as

inclusion of aspherical atoms improves the agreement with KRMM results in both bond lengths and angles. For P2A4, the ADPs from refinements 3 and 5 of HR data sets are almost indistinguishable, while some small ($\leq$0.001 Å^2) differences are observed for YGG. Inclusion of the aspherical density significantly improves the ADPs for LR data sets, yet they are still 0.001–0.003 Å^2 larger than those from the KRMM refinement. This is in accord with the LR refinement of the octapeptide (Jelsch *et al.*, 1998), in which the r.m.s. improvement in U_{ij} values compared with IAM results was found to be about 0.005 Å^2. Significant improvements over the standard IAM refinement are also shown in the rigid-bond analysis, especially for the LR data. For example, in P2A4 the r.m.s. of the DMSDA is reduced by a factor of two, *i.e.* 7.09 $\times$ 10^{-4} Å^2 compared with the IAM value of 15.64 $\times$ 10^{-4} Å^2. Even for HR data sets an r.m.s. improvement of about 1 $\times$ 10^{-4} Å^2 is observed for both compounds. As expected, improvements are also observed for the phase angles of reflections. Thus, for the HR YGG data, the r.m.s. difference from KRMM phases is reduced from 2.33° after IAM refinement to only 0.84° after refinement 3 (from 4.84 to 3.94° in P2A4), while for LR data the changes are from 2.36 to 1.03° for YGG and from 5.24 to 3.99° in P2A4. Absolute changes in the phases of reflections increase with increasing $\sin\theta/\lambda$ (Figure S14). However, the direct comparison of phases from refinement 3 and the IAM refinement (1) show r.m.s. improvements of about 2.0 and 6.2° for YGG and P2A4, respectively (see Table 4). In previous studies using the experimental databank, the r.m.s. phase difference between refinements analogous to 1 and 3 in the octapeptide was 2.6° (Jelsch *et al.*, 1998), while the average phase difference between the same types of refinements in crambin was 3.8° (Jelsch *et al.*, 2000), in good agreement with the present result. Overall, the advantages of this refinement, which uses aspherical pseudoatom parameters from the databank, compared with the conventional IAM refinements are obvious: improvements in geometry, ADPs, phases and residual Fourier maps. The number of refined parameters in these refinements is actually smaller than in the standard IAM, because the riding model is used for positional parameters of H atoms.

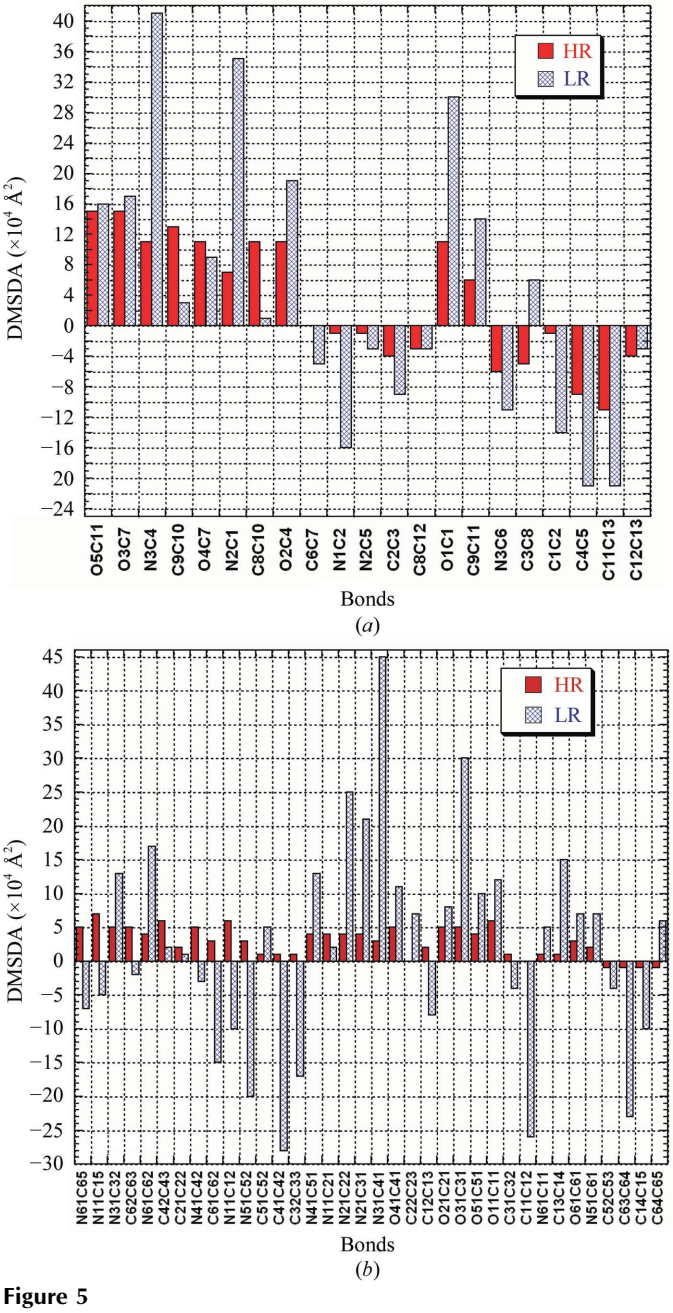

Figure 5
Differences in mean-squared displacement amplitudes (DMSDA) along interatomic vectors ($\times$10^4 Å^2) from IAM refinements of HR and LR data sets in YGG (*a*) and P2A4 (*b*).

3.4. Refinement 4

Results of the limited charge-density refinement 4 are expected to agree somewhat better with those from the

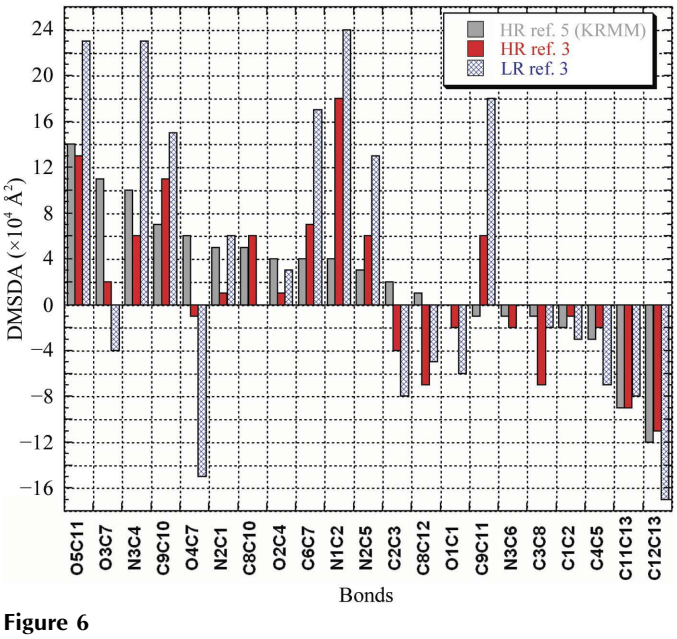

Figure 6
Differences in mean-squared displacement amplitudes (DMSDA) along interatomic vectors ($\times$10^4 Å^2) in YGG from LR refinement 3 and refinements 3 and 5 of HR data.

KRMM refinement than refinement 3. Indeed, for HR data sets the R factor further decreases by about 0.1–0.15% to values that are halfway between those from refinements 3 and 5. Spherical features in the residual maps of YGG are reduced in amplitude compared with refinement 3 and are much closer to those from the KRMM refinement. However, aspherical features in these maps do not change because aspherical pseudoatom parameters are not refined. Bond lengths and angles do not improve much, which is likely to be because only nuclei-centered spherical features in the residual maps can be refined (which should affect ADPs much more than the positional parameters). For P2A4 the r.m.s. deviations of the geometrical parameters from the KRMM results are essentially the same for both HR and LR data sets. In YGG small improvements of 0.01–0.02° are found for bond angles and torsion angles. ADPs in HR data sets are essentially the same as in refinement 3. As expected, further (small) improvements in ADPs are observed in LR refinements, yet they are still slightly higher than KRMM values. It is interesting to note that results of the rigid-bond test from refinement 4 are slightly improved in P2A4 for both HR and LR data sets, while the opposite is observed in YGG. Similarly, the OSF from refinement 3 of P2A4 becomes much closer to that from

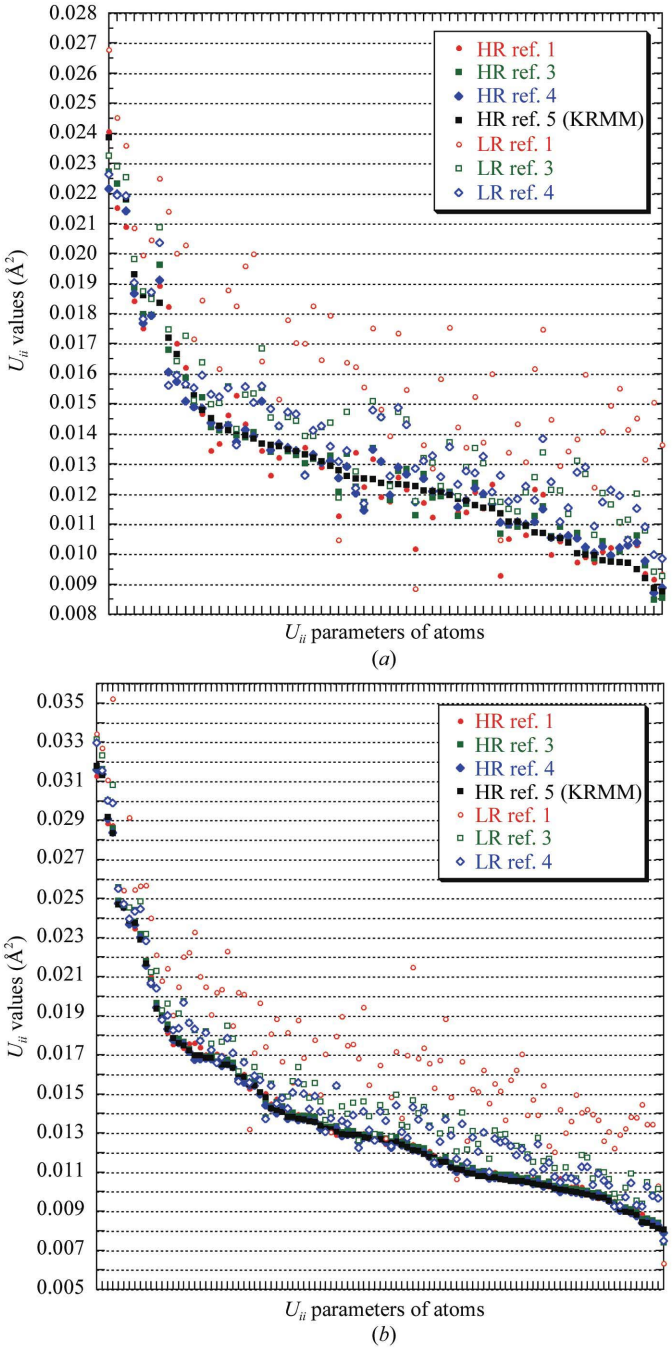

Figure 7
U_{ii} parameters (Å^2) of YGG (*a*) and P2A4 (*b*) from all refinements sorted in descending order of KRMM values (individual values are given in Tables S1 and S2 of the supplementary material).

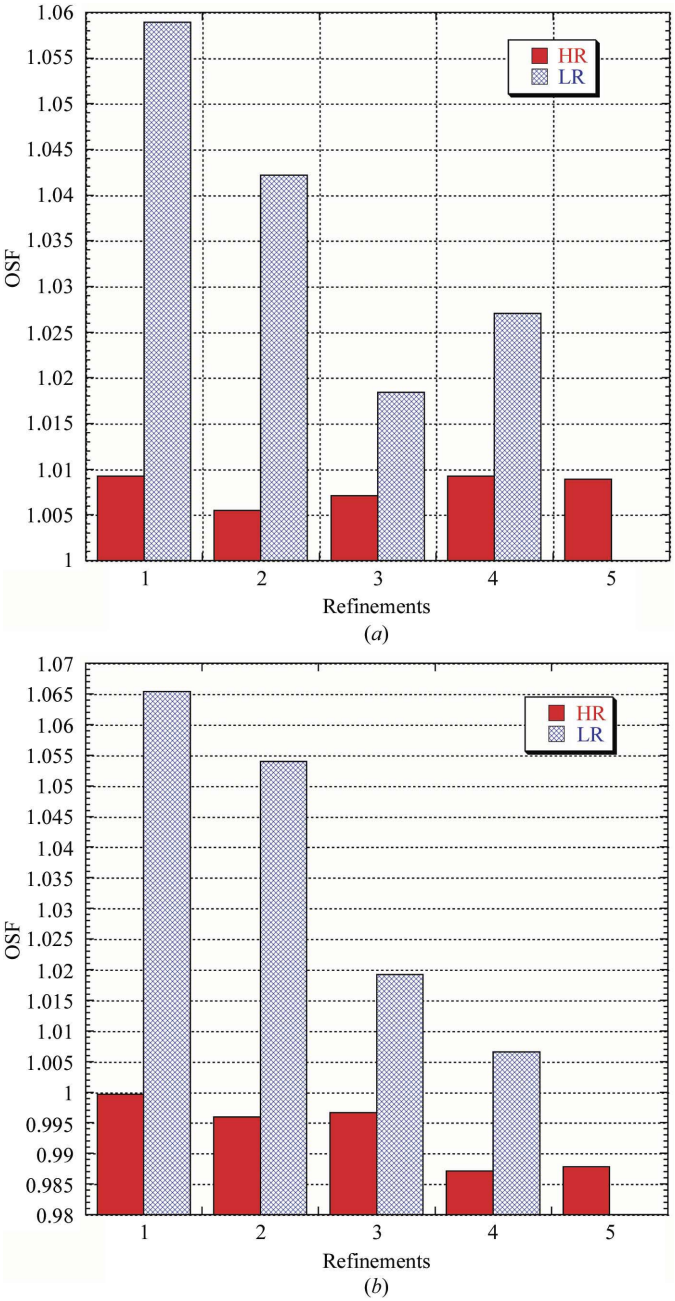

Figure 8
Scale factors in refinements of YGG (*a*) and P2A4 (*b*).

KRMM refinement for both HR and LR data sets, while in the YGG LR refinement it actually deviates more from the KRMM value. Phases of reflections from refinement 4 change by only 0.3–0.4° compared with refinement 3 for all data sets (see Table 4). While refinement 4 does show some improvements over refinement 3, it may not have a practical application in the refinement of large macromolecular structures as the number of refined parameters increases (even when using chemical constraints), which leads to a decrease in the number of reflections to number of refined parameters ratio. Nevertheless, this type of refinement represents an important intermediate step between databank and full multipolar refinements. Note that the latter are only feasible with data sets extending further in reciprocal space than what is commonly referred to as ultrahigh resolution.

3.5. Refinement 5

It is beyond the scope of this paper to discuss the results of full multipolar refinements of given data sets, which have been reported in the original studies by Pichon-Pesme *et al.* (2000) and Dittrich *et al.* (2002). However, we note that our κ'-restricted multipole model results (taken as a reference for UBDB testing) are in a very good agreement with those presented in the original papers. *R* factors are well within 0.1% for both YGG and P2A4 and residual Fourier maps and rigid-bond test in the case of P2A4 are as good as in the original study.

4. Summary

In the present study, we have applied the University at Buffalo theoretical databank of aspherical pseudoatoms to the refinement of two experimental data sets from polypeptides which originate from different sources. Both high-resolution (HR; $\sin\theta/\lambda_{\max} > 1.1$ Å^{-1}) and low-resolution (LR; $\sin\theta/\lambda_{\max} = 0.6$ Å^{-1}) refinements were performed.

No bonding features are visible in the residual density maps after IAM refinement of even excellent quality LR data because they are accounted for by the bias in the ADPs and positional parameters. In the absence of high-order data, application of an *a priori* determined aspherical electron-density model is essential if accurate ADPs and positional parameters are to be obtained. The aspherical parameters may originate from either experimental studies (Pichon-Pesme *et al.*, 1995, 2004) or, as in our case, from theoretical calculations. Theoretically derived pseudoatom parameters, in addition to being highly transferable and able to faithfully reproduce electrostatic properties of theoretical densities at the level of the theory on which they are based, are well suited for refinements of experimental data. The effect of use of aspherical scattering factors is reflected in basically all statistical and physical descriptors of the LR refinements.

(i) The *R* factor is lowered by about 1% on an absolute scale (or between 22 and 50% of the initial value).

(ii) The molecular geometry is improved; for example, bond lengths and angles are determined within 0.002–0.003 Å and 0.09–0.17°, respectively, of values from multipolar refinements of HR data.

(iii) ADP parameters become much closer to those determined using multipolar refinements of HR data sets.

(iv) The rigid-bond test is significantly improved (by as much as a factor of two in P2A4) and in most cases satisfies the 0.001 Å^2 criterion.

(v) Overall scale factors become significantly closer (within 1–2%) to values determined using HR data sets, while the OSF from IAM refinements shows a bias of +5%.

(vi) The phases of reflections become much closer to those determined using multipolar refinements of HR data sets.

These findings are consistent with previous studies using both the experimental databank (Jelsch *et al.*, 1998, 2000, 2005; Muzet *et al.*, 2003) and theoretical invarioms (Dittrich *et al.*, 2005). However, the UBDB allows a much greater flexibility in the types of molecular systems that can be studied than the experimental DB, as the incorporation of new atom types into the theoretical databank is rapid and essentially unlimited.

The limited multipolar refinement 4 (*i.e.* refinement of valence-population and expansion-contraction parameters) does not lead to any significant improvements compared with refinement 3. For example, in YGG the DMSDA from refinement 4 are even larger than those after refinement 3. This might be related to the presence of spherical residual density features at atomic positions (Figs. 2 and 3) which are not observed in P2A4. In such cases refinement 3 is highly preferable as it may provide the best possible description of the aspherical electron density, including charge-transfer and expansion–contraction parameters of the atomic valence shells.

Further application of the UBDB to large biologically important systems, as well as a thorough comparison with experimental databanks, are currently under way.

The University at Buffalo aspherical-atom databank and executables of the program *LSDB* for various platforms are available from http://harker.chem.buffalo.edu.

5. Implications for refinement of accurate macromolecular data

It is clear from the above analysis that the refinement procedure routinely applied to protein structures leads to a considerable bias in the temperature parameters, while the positional parameters are affected to a lesser extent. The phases as determined by such a refinement differ from those with the improved scattering-factor model by r.m.s. values of 2–6°. When accurate data at reasonable resolution (∼0.8 Å) are available, the use of the theoretical databank leads to significantly improved results without an increase in the number of refined parameters. In order for this procedure to be applicable, the spherical atom refinement should yield residual Fourier maps of the same quality as found in the benchmark systems described in this paper. These maps in general should show low noise ($|\Delta\rho_{\max}| \simeq 0.2$ e Å^{-3}) and an indication of aspherical electron density in lone-pair and bonding regions. A full aspherical atom refinement is in

general not warranted, unless exceptionally high-resolution data of ∼0.6 Å or better, with satisfactory completeness and statistics in the higher order data, are available.

Financial support of this work from the National Institutes of Health (GM56829) and the National Science Foundation (CHE0236317) is gratefully acknowledged.

References

Abramov, Y. A., Volkov, A. V. & Coppens, P. (1999). *Chem. Phys. Lett.* **311**, 81–86.

Afonine, P. V., Lunin, V. Y., Muzet, N. & Urzhumtsev, A. (2004). *Acta Cryst.* D**60**, 260–274.

Allen, F. H. (1986). *Acta Cryst.* B**42**, 515–522.

Bönisch, H., Schmidt, C. L., Bianco, P. & Ladenstein, R. (2005). *Acta Cryst.* D**61**, 990–1004.

Cachau, R., Howard, E., Barth, P., Mitschler, A., Chevrier, B., Lamour, V., Joachimiak, A., Sanishvili, R., Van Zandt, M., Sibley, E., Moras, D. & Podjarny, A. (2000). *J. Phys.* **10**, 3–13.

Clementi, E. & Raimondi, D. L. (1963). *J. Chem. Phys.* **38**, 2686–2689.

Clementi, E. & Roetti, C. (1974). *At. Data Nucl. Data Tables*, **14**, 177–478.

Coppens, P. (1967). *Science*, **158**, 1577–1579.

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

Deng, J., Xiong, Y. & Sundaralingam, M. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 13665–13670.

Dittrich, B., Hübschle, C. B., Messerschmidt, M., Kalinowski, R., Girnt, D. & Luger, P. (2005). *Acta Cryst.* A**61**, 314–320.

Dittrich, B., Koritsánszky, T., Grosche, M., Scherer, W., Flaig, R., Wagner, A., Krane, H. G., Kessler, H., Riemer, C., Schreurs, A. M. M. & Luger, P. (2002). *Acta Cryst.* B**58**, 721–727.

Dittrich, B., Strumpel, M., Schäfer, M., Spackman, M. A. & Koritsánszky, T. (2006). *Acta Cryst.* A**62**, 217–223.

Dominiak, P. M., Volkov, A., Li, X., Messerschmidt, M. & Coppens, P. (2007). In the press.

Faerman, C. H. & Price, S. L. (1980). *J. Am. Chem. Soc.* **112**, 4915.

Fernandez-Serra, M. V., Junquera, J., Jelsch, C., Lecomte, C. & Artacho, E. (2000). *Solid State Commun.* **116**, 395–400.

Guillot, B., Muzet, N., Artacho, E., Lecomte, C. & Jelsch, C. (2003). *J. Phys. Chem. B*, **107**, 9109–9121.

Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.

Hirshfeld, F. L. (1976). *Acta Cryst.* A**32**, 239–244.

Hohenberg, P. & Kohn, W. (1964). *Phys. Rev.* **136**, B864–B871.

Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., Moras, D., Schneider, T. R., Joachimiak, A. & Podjarny, A. (2004). *Proteins*, **55**, 792–804.

Jelsch, C., Guillot, B., Lagoutte, A. & Lecomte, C. (2005). *J. Appl. Cryst.* **38**, 38–54.

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.

Ko, T.-P., Robinson, H., Gao, Y.-G., Cheng, C.-H. C., DeVries, A. L. & Wang, A. H.-J. (2003). *Biophys. J.* **84**, 1228–1237.

Koritsanszky, T. S., Volkov, A. & Coppens, P. (2002). *Acta Cryst.* A**58**, 464–472.

Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Van Dorsselaer, A., Podjarny, A. & Moras, D. (1999). *Acta Cryst.* D**55**, 721–723.

Lecomte, C., Guillot, B., Jelsch, C. & Podjarny, A. (2005). *Int. J. Quant. Chem.* **101**, 624–634.

Lecomte, C., Guillot, B., Muzet, N., Pichon-Pesme, V. & Jelsch, C. (2004). *Cell. Mol. Life Sci.* **61**, 774–782.

Li, X., Volkov, A. V., Szalewicz, K. & Coppens, P. (2006). *Acta Cryst.* D**62**, 639–647.

Muzet, N., Guillot, B., Jelsch, C., Howard, E. & Lecomte, C. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 8742–8747.

Pichon-Pesme, V., Jelsch, C., Guillot, B. & Lecomte, C. (2004). *Acta Cryst.* A**60**, 204–208.

Pichon-Pesme, V., Lachekar, H., Souhassou, M. & Lecomte, C. (2000). *Acta Cryst.* B**56**, 728–737.

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

Slater, J. C. (1932). *Phys. Rev.* **42**, 33–43.

Spek, A. L. (2003). *J. Appl. Cryst.* **36**, 7–13.

Stevens, E. D. & Coppens, P. (1975). *Acta Cryst.* A**31**, 612–619.

Tereshko, V., Wilds, C. J., Minasov, G., Prakash, T. P., Maier, M. A., Howard, A., Wawrzak, Z., Manoharan, M. & Egli, M. (2001). *Nucleic Acids Res.* **29**, 1208–1215.

Volkov, A., Abramov, Y. A. & Coppens, P. (2001). *Acta Cryst.* A**57**, 272–282.

Volkov, A., King, H. F. & Coppens, P. (2006). *J. Chem. Theory Comput.* **2**, 81–89.

Volkov, A., Koritsanszky, T. S. & Coppens, P. (2004). *Chem. Phys. Lett.* **391**, 170–175.

Volkov, A., Koritsanszky, T., Li, X. & Coppens, P. (2004). *Acta Cryst.* A**60**, 638–639.

Volkov, A., Li, X., Koritsanszky, T. S. & Coppens, P. (2004). *J. Phys. Chem. A*, **108**, 4283–4300.

Volkov, A., Macchi, P., Farrugia, L. J., Gatti, C., Mallinson, P., Richter, T. & Koritsanszky, T. (2006). *XD2006. A Computer Program Package for Multipole Refinement, Topological Analysis of Charge Densities and Evaluation of Intermolecular Energies from Experimental or Theoretical Structure Factors.* University at Buffalo, NY, USA; University of Milano, Italy; University of Glasgow, UK; CNRISTM, Milano, Italy; Middle Tennessee State University, TN, USA.