PROTEINS: Structure, Function, and Bioinformatics 66:279–303 (2007)

Statistical and Conformational Analysis of the Electron Density of Protein Side Chains

Maxim V. Shapovalov and Roland L. Dunbrack Jr.*
Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

ABSTRACT **Protein side chains make most of the specific contacts between proteins and other molecules, and their conformational properties have been studied for many years. These properties have been analyzed primarily in the form of rotamer libraries, which cluster the observed conformations into groups and provide frequencies and average dihedral angles for these groups. In recent years, these libraries have improved with higher resolution structures and using various criteria such as high thermal factors to eliminate side chains that may be misplaced within the crystallographic model coordinates. Many of these side chains have highly non-rotameric dihedral angles. The origin of side chains with high B-factors and/or with non-rotameric dihedral angles is of interest in the determination of protein structures and in assessing the prediction of side chain conformations. In this paper, using a statistical analysis of the electron density of a large set of proteins, it is shown that: (1) most non-rotameric side chains have low electron density compared to rotameric side chains; (2) up to 15% of χ_1 non-rotameric side chains in PDB models can clearly be fit to density at a single rotameric conformation and in some cases multiple rotameric conformations; (3) a further 47% of non-rotameric side chains have highly dispersed electron density, indicating potentially interconverting rotameric conformations; (4) the entropy of these side chains is close to that of side chains annotated as having more than one χ_1 rotamer in the crystallographic model; (5) many rotameric side chains with high entropy clearly show multiple conformations that are not annotated in the crystallographic model. These results indicate that modeling of side chains alternating between rotamers in the electron density is important and needs further improvement, both in structure determination and in structure prediction. Proteins 2007;66:279–303.** © 2006 Wiley-Liss, Inc.

Key words: protein side chains; rotamers; X-ray crystallography

INTRODUCTION

In the past few years, the number of available protein structures has increased dramatically, reaching 37,000 in June 2006. This increase in data allows us to perform large-scale statistical analysis that was not possible even a few years ago. This is especially true for high-resolution structures which are now much more abundant due to the availability of synchrotron X-ray sources. These statistical analyses are the basis for validation of protein structures[1] as well as the derivation of energy functions for prediction and simulation.[2] While the number of unique sequences in the Protein Data Bank is about 24,000, there are more than 3 million sequences available in the non-redundant sequence databases. Structure prediction methods, mostly based on homology, are used to fill this gap.[3] Thus the accurate determination of Cartesian coordinate positions from electron density in X-ray experiments is critical in a number of fields.

Nearly all side-chain prediction methods depend on the concept of side-chain rotamers (reviewed in Ref. 4). From conformational analysis of organic molecules, it was predicted long ago[5,6] that protein side chains should attain a limited number of conformations because of steric and dihedral strain within each side chain and between the side chain and the backbone. As crystal structures of proteins have been solved in increasing numbers, a variety of rotamer libraries have been compiled with increasing amounts of detail and greater statistical soundness; that is, with more structures at higher resolution.[7–17] Lovell et al. proposed methods for selecting structurally well-determined side chains from protein structures, based on a B-factor cutoff and atom–atom contacts (including hydrogens) that might indicate improperly placed atoms. This resulted in lower variance of dihedral angles about average rotamer values, and fewer examples of "impossible" conformations with large steric conflicts. We subsequently used programs from the Richardson group and the same criteria in deriving a version of our backbone-dependent rotamer library.[4]

Although many rotamers with unlikely dihedral angles near the eclipsed positions are removed by the procedures of Lovell et al., it remains an interesting question as to how these so-called non-rotameric side

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NIH; Grant numbers: R01-HG02302, CA06972; Grant sponsors: Pennsylvania Tobacco Settlement; Commonwealth of Pennsylvania.

*Correspondence to: Roland L. Dunbrack Jr., Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111. E-mail: roland.dunbrack@fccc.edu

Received 20 February 2006; Revised 19 June 2006; Accepted 29 June 2006

Published online 1 November 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21150

chains enter into protein structures. The main possibilities are: (1) that they are misfit to the actual electron density, which is rotameric; (2) that they are near the average position of a side chain that is moving between two different rotameric positions; (3) that the backbone is misplaced and therefore the side-chain dihedral angle is not correct; and (4) that they are true positions for the side chain which is held at a strained value near the top of an energy barrier by interactions with the rest of the protein. The values of B-factors alone do not help us to choose among these possibilities, so we have undertaken a study of the electron density, calculated from the deposited structure factors and the model coordinates, of protein side chains in a large sample of proteins. We have examined several features, including the values of side-chain density as a function of dihedral angle, the variation of electron density as a function of χ_1 for single side chains with poor dihedral angle positions, and the entropy of electron density.

We have found that about 15% of non-rotameric side chains (i.e. according to the PDB model) have electron density more consistent with rotameric conformations (sometimes multiple rotameric conformations). About 47% have peaks in density at non-rotameric positions but also have spread-out density consistent with multiple conformations at χ_1. The remaining 38% have electron density consistent with stable conformations at non-rotameric conformations. This occurs for specific side chain types, and visual examination shows that these side chains are fixed in position by a large number of specific interactions with the rest of the protein. Some may be due to misfitting of the local backbone, which would affect the determination of the χ_1 dihedral angle. This is difficult to determine without further refinement of the structure, which is beyond the scope of this paper.

The results are consistent with computational studies using molecular mechanics energy functions by Petrella and Karplus[18] and with the analysis of Lovell et al.[17] In the Petrella–Karplus study, using the CHARMM potential, the authors demonstrate that almost half of so-called non-rotameric side chains are not in a local energy minimum in the context of the crystal environment, while nearly 100% of rotameric side chains are. This indicates that many non-rotameric side chains are poorly refined in X-ray crystal structures. In this paper, we examine electron density rather than using molecular mechanics energy functions to explore the conformations of protein side chains in a statistical manner across large numbers of X-ray structures from the PDB.

MATERIALS AND METHODS

Protein Structure Evaluation Based on X-Ray Diffraction Data: Electron Density as a Measure of Confidence in Atomic Positions

To evaluate a protein structure determined by an X-ray diffraction experiment we need to have two sets of data: (1) an atomic model of the protein, described in terms of Cartesian atom coordinates and (2) structure factors coming from the X-ray experiment. These two data components are downloaded from RCSB Protein Data Bank (ftp://ftp.rcsb.org).[19] The structure factors for ~65% of X-ray structures in the PDB are also deposited with the PDB.

To get a $m \cdot |F_o| \cdot \exp(i \cdot phi_{calc})$ electron density distribution map we use two scripts (*generate.inp* and *model_map.inp*) from the program package CNS (Crystallography and NMR System).[20] The CNS molecular topology file (mtf) required for processing *model_map.inp* was created by running the script *generate.inp*. The atom content in the model was not modified (no addition/deletion of missing/existing atoms in the model). The topology and parameters files were the CNS default, which contain the recommended values for proteins, DNA/RNA, water molecules and carbohydrates. For example, in the CNS_TOPPAR namespace, protein.top, protein.link, and protein_rep.param are respectively the protein topology, linkage, and parameter files.

Using the *model_map.inp* script we generated a $m \cdot |F_o| \cdot \exp(i \cdot phi_{calc})$ electron density map derived from the sigmaA weighted map, $(u \cdot m \cdot |F_o| - v \cdot D \cdot |F_C|) \cdot \exp(i \cdot phi_{calc})$ by setting $u = 1$ and $v = 0$ where m and D are calculated from sigmaA (m is figure of merit and D is estimate of the error in the partial structure from coordinate errors). All reflections within the resolution limits specified by the authors of the atomic model were taken (including the test-set reflections if they were provided). The use of model amplitudes $|F_C|$ for unmeasured data $|F_O|$ was disabled. The anisotropic initial B-factor correction was applied, and the standard bulk solvent correction was used. The map grid size (*grid*) was set to 0.25 for higher accuracy. When we compared atom electron densities in the maps generated using the 0.333 and 0.25 grid values, there was only 1–4% relative difference. The 0.25 *grid* demonstrated satisfactory convergence; decreasing it further would have added significant processing time and memory overhead.

The map covered the whole molecule with the 3Å cushion around non-hydrogen atoms. The following parameter files were used from the CNS_TOPPAR namespace: protein_rep.param, dna-rna_rep.param, water_rep.param, ion.param, carbohydrate.param respectively for proteins, DNA/RNA, water molecules, ions and carbohydrates. If TLS corrections were described in the PDB files, they were not applied. Since a test-set is rarely deposited in structure factor data files in the PDB, we had to use all reflections instead of the test-set reflections for computation of the sigmaA distribution. Therefore, the sigmaA values were overestimated because they were previously used in refinement.

For more details of the parameters used, please refer to the *model_map.inp* and *generate.inp* templates given in the Supplemental Material.

When we were choosing the type of an ED map and its generation parameters we tried to follow a strategy to decrease the atomic model information component

(model bias) in the output ED map. It is possible that less model-biased ED maps can be generated (1) when experimental phases are available, or (2) a test-set is provided, or (3) by using annealed omit maps. In general, we do not have experimental phases for very many structures or the test-set. Annealed omit maps are computationally expensive. We selected parameters that on the one hand rely only on the data available and on the other hand decrease the model bias. Nevertheless, we emphasize that the maps generated use the atomic model information during calculation of model phases and, therefore are to some extent model biased. However, even with model bias, our results as shown below indicate that many atoms have poor electron density and some atoms are placed improperly.

The resulting electron density map from CNS is a discrete function of the Cartesian coordinates (*x, y, z*) with values defined at node points of a grid put on the unit cell of the protein crystal:

$$\rho(\vec{r}_{i,j,k}) = \rho(x_i, y_j, z_k) \tag{1}$$

Since model atoms and other objects of interest are not mostly located at grid points, we used interpolation to calculate density at other points (see following).

To assess confidence levels of an atom position $\vec{r}_{\text{atom}} = x_{\text{atom}}, y_{\text{atom}}, z_{\text{atom}}$ we calculated two different values from the electron density map: point electron density (PED) and integrated electron density (IED).

Point electron density

$$\begin{aligned}\rho_{\text{point}}(\vec{r}_{\text{atom}}) &= \rho_{\text{point}}(x_{\text{atom}}, y_{\text{atom}}, z_{\text{atom}}) \\ &= \mathit{Quad3DSpline}(x_{\text{atom}}, y_{\text{atom}}, z_{\text{atom}}; \{\rho_{i,j,k}\})\end{aligned} \tag{2}$$

We refer to the point electron density as ρ_{point} with subscript *point* to emphasize that it represents electron density at some point of space. In Eq. (2) $\rho_{\text{point}}(\vec{r}_{\text{atom}})$ designates electron density in the $\vec{r}_{\text{atom}} = x_{\text{atom}}, y_{\text{atom}}, z_{\text{atom}}$ atom position. We use a quadratic three-dimensional spline to get an electron density value in any position. The interpolating function has 10 unknown constants:

$$\begin{aligned}\rho(\vec{r}) = \rho(x,y,z) = {} & A_0 + B_1 \cdot x + B_2 \cdot y + B_3 \cdot z + C_{11} \cdot x^2 \\ & + C_{22} \cdot y^2 + C_{33} \cdot z^2 + 2 \cdot C_{12} \cdot x \cdot y + 2 \cdot C_{23} \cdot y \cdot z \\ & + 2 \cdot C_{13} \cdot x \cdot z\end{aligned} \tag{3}$$

To find a point electron density for each atom we take into account 10 grid points closest to its position and their electron density values and calculate the best fit for the parameters in Eq. (3). We use PED not only for calculating electron density in atom positions of the PDB structures but also in positions with coordinates different from the atom coordinates—for example, as it is used in the integrated electron density calculations shown below and other types of analysis considered later.

Integrated electron density

We calculate an *integrated electron density* (IED) from the following equation:

$$\rho_{\text{integ}}(\vec{r}_{\text{atom}}) = \frac{\int\limits_{|\vec{r}-\vec{r}_{\text{atom}}| \le 1.5\text{Å}} \rho_{\text{point}}(\vec{r}) \cdot \rho_{\text{theoretical\_atom}}(\vec{r} - \vec{r}_{\text{atom}}) \cdot \mathrm{d}\vec{r}}{\int\limits_{|\vec{r}-\vec{r}_{\text{atom}}| \le 1.5\text{Å}} \rho_{\text{theoretical\_atom}}(\vec{r} - \vec{r}_{\text{atom}}) \cdot \mathrm{d}\vec{r}} \tag{4}$$

where $\rho_{\text{theoretical\_atom}}(\vec{r})$ is the theoretical probability density function of electron positions with their atom center at the zero vector $\vec{0} = (0\ 0\ 0)$. We approximate it using the following set of equations:

$$\begin{cases} \rho_{\text{theoretical\_atom}}(\vec{r}) \cong \rho_{\text{theor\_atom\_approx}}(|\vec{r}|) \equiv \rho_{\text{theor\_atom\_approx}}(r) \equiv C \cdot \exp\left(-\frac{r}{a}\right) \\ \int\limits_0^{2\cdot\pi} \mathrm{d}\phi \int\limits_0^{\pi} \sin(\theta)\mathrm{d}\theta \int\limits_0^{\infty} \rho_{\text{theor\_atom\_approx}}(r) \cdot r^2 \cdot \mathrm{d}r \equiv 4 \cdot \pi \cdot \int\limits_0^{\infty} \rho_{\text{theor\_atom\_approx}}(r) \cdot r^2 \cdot \mathrm{d}r = 1 \\ \int\limits_0^{2\cdot\pi} \mathrm{d}\phi \int\limits_0^{\pi} \sin(\theta)\mathrm{d}\theta \int\limits_0^{r_{\text{atom}}} \rho_{\text{theor\_atom\_approx}}(r) \cdot r^2 \cdot \mathrm{d}r \equiv 4 \cdot \pi \cdot \int\limits_0^{r_{\text{atom}}} \rho_{\text{theor\_atom\_approx}}(r) \cdot r^2 \cdot \mathrm{d}r = 0.9 \end{cases} \tag{5}$$

The last equation requires 90% of the "atom" electron density to be in a sphere with radius r_{atom}. Solving these equations we find two constants C and a for each atom type and bond type. The 1.5 Å integration sphere in Eq. (4) is sampled with equidistant points starting from its center with *x*-, *y*-, and *z*-stepsizes equal to the grid spacing in each dimension respectively.

In other words, IED is an average atom electron density calculated based on the theoretical probability density function of electron positions: $\rho_{\text{integ}}(\vec{r}_{\text{atom}}) \cong \langle \rho(\vec{r}) \rangle_{\rho_{\text{theoretical\_atom}}}$. Such an integration procedure "cuts" the atom's electron density from other electron density and averages it. This technique also makes this value more robust and reliable for assessing atom positions and reduces its dependence on the radius of integration. We denote the integrated electron density ρ_{integ} with subscript *integ* to distinguish it from the point electron density ρ_{point}. IED is comparable with other real-space fit statistics,[21–24] expressed as an R value or as a correlation coefficient between "observed" and calculated density.[25]

Atom Confidence Level (PED, IED) Normalization

We want to evaluate electron density across many structures in the PDB in order to perform statistical

analysis of side chains. Because of the variability in water content, dynamics within the crystal, and other features of X-ray crystallography (different crystallographic equipment and software), we need to normalize the density for each structure in a consistent way. To accomplish this, we use the following steps:

1. The μ–3σ electron density level (mean minus three standard deviations of the unit cell electron density distribution) is set to "0" (e/Å^3). We do not use the absolute minimum of the unit cell electron density as a control point for the normalization since it is an unstable value owing to incompleteness of X-ray reflection set, errors in structure factors (amplitudes and phases), etc. The μ–3σ normalization point ($P(\rho < \mu - 3\sigma) \sim 0.15\%$) guarantees a robust estimate of the background electron density level.
2. We use the average electron density of the backbone atoms as a constant across different structures. In general, the backbone is more fixed than the side chains, and we are interested in how mobile the side-chain atoms are relative to the backbone. The backbone atoms (N, C_α, C, O) on average have approximately seven electrons around their nuclei (including the electrons provided by H_N and H_α). The protein typical atom size is about 1.5 Å in radius, so the electron density at the centers of backbone atoms averages about $\frac{7e}{4/3\cdot\pi\cdot(1.5\ \text{Å})^3} = 0.5\frac{e}{\text{Å}^3}$. We decided to include the backbone atom volume constant into the electron density units: we set average backbone density to "$7\frac{e}{14\cdot\text{Å}^3}$".

Hence, for each X-ray diffraction structure, we use the following technique to normalize atom electron density (PED, IED) to the same scale:

$$\rho_{\text{norm}} = K\cdot(\rho_{\text{orig}} + a) \tag{6}$$

$$\begin{cases} 0 = K\cdot((\mu-3\sigma)+a) \\ 7 = K\cdot(\langle\rho_{\text{orig}}\rangle_{\text{backbone}}+a) \end{cases} \Rightarrow \begin{cases} a = -(\mu-3\sigma) \\ K = 7/(\langle\rho_{\text{orig}}\rangle_{\text{backbone}} - (\mu-3\sigma)) \end{cases} \tag{7}$$

$$\rho_{\text{point,norm}}(\vec{r}_{\text{atom}}) = K_{\text{point}}\cdot(\rho_{\text{point,orig}}(\vec{r}_{\text{atom}}) + a_{\text{point}}) \tag{8}$$

$$\rho_{\text{integ,norm}}(\vec{r}_{\text{atom}}) = K_{\text{integ}}\cdot(\rho_{\text{integ,orig}}(\vec{r}_{\text{atom}}) + a_{\text{integ}}) \tag{9}$$

where

$$a_{\text{point}} = a_{\text{integ}} \equiv a \equiv \mu - 3\sigma \tag{10a}$$

and

$$K_{\text{point}} \neq K_{\text{integ}} \tag{10b}$$

χ_1 and χ_2 Rotations

To investigate side-chain disorder we rotated an X_γ pseudo-atom by varying its χ_1 dihedral angle with a 5° stepsize (Fig. 1) and calculated the point ED at each position. This was done in two ways: (1) by keeping the original atomic model C_β—X_γ bond length and C_α—C_β—X_γ angle; (2) by substituting the original values with the standard averages from high-resolution peptide structures.[26]

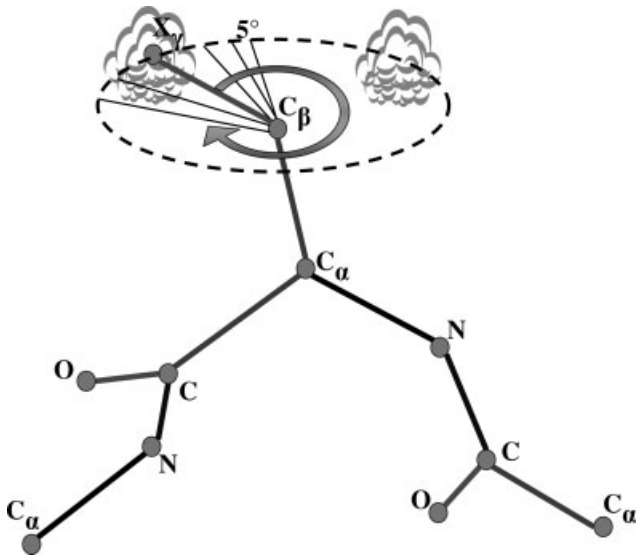

Fig. 1. Rotation of the X_γ pseudo-atom by varying its χ_1 torsion angle with a 5° step and calculating point electron density (PED), $\rho_{\text{point}}(\chi_1)$, at each position. This can be done in two different ways: (1) by keeping the original atomic model C_β—X_γ bond length and C_α—C_β—X_γ angle; or (2) by substituting the original PDB entry values with the standard average values. In this paper, we use the values given in the PDB models, although in cases where C_β is misplaced, the latter may be useful.

$$1)\ \rho^{\text{model}}_{\text{point}} = \rho_{\text{point}}(\chi_1 | C_\beta - X^{\text{model}}_\gamma, C_\alpha - C_\beta - X^{\text{model}}_\gamma) \tag{11}$$

$$2)\ \rho^{\text{stand}}_{\text{point}} = \rho_{\text{point}}(\chi_1 | C_\beta - X^{\text{stand}}_\gamma, C_\alpha - C_\beta - X^{\text{stand}}_\gamma) \tag{12}$$

In some calculations, we added a second variable "bond length" to these functions:

$$1)\ \rho^{\text{model}}_{\text{point}} = \rho_{\text{r}}(\chi_1, r_{\beta\gamma} | C_\alpha - C_\beta - \chi^{\text{model}}_\gamma) \tag{13}$$

$$2)\ \rho^{\text{stand}}_{\text{point}} = \rho_{\text{r}}(\chi_1, r_{\beta\gamma} | C_\alpha - C_\beta - \chi^{\text{stand}}_\gamma) \tag{14}$$

An additional radial variable helps to distinguish ED peaks formed by a single or multiple-conformational X_γ atom from the peaks created by ED noise fluctuation or closely positioned adjacent atoms. For example, if we expect to find a C_γ atom at some χ^0_1 position then $\rho(\chi^0_1, r_{\beta\gamma})$ has to have a maximum at $r_{\beta\gamma} = C_\beta - C_\gamma$ not $C_\beta - H$ or a distance expected for an adjacent water molecule.

The "bond length" $r_{\beta\gamma}$ was varied in the range (0.0–3.0 Å) with a 0.1 Å stepsize. The same technique was applied for the χ_2 rotations of the pseudo X_γ atom

$$\rho_{\text{point}}(\chi_2) = \rho_{\text{point}}(\chi_2 | X_\gamma - X_\delta, C_\beta - X_\gamma - X_\delta) \quad (15)$$

$$\rho_{\text{point}}(\chi_2, r_{\gamma\delta}) = \rho_{\text{point}}(\chi_2, r_{\gamma\delta} | C_\beta - X_\gamma - X_\delta) \quad (16)$$

These calculations were performed with the backbone fixed. It is likely that the backbone adjusts somewhat when the side chain is placed in different rotamers, but we did not account for this. Indeed, recently Davis et al. identified "the backrub motion," a slight adjustment of the backbone for different rotamers of the same side chain.[27]

Side-Chain Conformation Evaluation

We have already introduced *atom* confidence levels (point ED and integrated ED). But to evaluate accuracy of a *backbone* or *side-chain* conformation as a whole, we designed backbone and side-chain confidence levels, defined as (for IED):

$$\rho_{\text{integ}}^{\text{backbone}} = \sqrt[4]{\prod_{k=1}^{4} \rho_{\text{integ}}(\vec{r}_k{}^{\text{atom}})} \Big/ \left\langle \rho_{\text{integ}}^{\text{backbone}} \right\rangle \quad (17)$$

$$\rho_{\text{integ}}^{\text{side\_chain}} = \sqrt[n]{\prod_{j=1}^{n} \rho_{\text{integ}}(\vec{r}_j{}^{\text{atom}})} \Big/ \left\langle \rho_{\text{integ}}^{\text{backbone}} \right\rangle \quad (18)$$

where n is the number of atoms in a side chain and $\langle \rho_{\text{integ}}^{\text{backbone}} \rangle$ is the average protein backbone atoms IED. Following our normalization scheme $\langle \rho_{\text{integ}}^{\text{backbone}} \rangle$ is a constant and equals 7. This formula can be interpreted as a geometric mean of the individual confidence levels of the backbone or side-chain atoms constituting a residue normalized to the average confidence level of the protein backbone atoms. This method is similar to that used in the program *sfcheck*.[21]

Multi-Conformational Side-Chains

Electron density maps built from X-ray data often reveal multiple conformations of some side chains. Occupancy of each conformation is related to the proportion of asymmetric units in the crystal on average in which the conformation is found during X-ray data collection.

Side chains can exhibit multiple conformations starting from X_γ, $X_\delta, \ldots$ side-chain atoms, ignoring multiple C_β positions due to fluctuations in the backbone. The majority of multi-conformational side chains annotated in the PDB (we refer to these as "PDB-declared" or "PDB-multi-conformational") begin with the X_γ atom, where X is *C, O,* or *S*. The two (or more) X_γ positions may belong to the same rotamer or two different rotamers, depending on the dihedral angles or the distances between their positions. If two C_γ atom positions belong to two different rotamers, then the distance between them is usually $d(C_\gamma^A, C_\gamma^B) \sim 1.5$ Å. There are also multi-conformational side chains branching out at the X_γ, $X_\delta, \ldots$ atoms but in this group only parts of the side chains are multi-conformational. In this paper, we focus on side chains with disorder at the X_γ atom (disorder at the χ_1 level) and call them *multi-conformational side chains*.

χ_1 Rotamer Entropy as an Estimate of Side-Chain Disorder

The $\rho_{\text{point}}(\chi_1)$ electron density function can be calculated using Eq. (19) and then normalized [Eq. (20)], and interpreted as a χ_1 probability density function ($\rho_{\text{prob}}(\chi_1)$):

$$\rho_{\text{point}}^*(\chi_1) = \max[0, \rho_{\text{point}}(\chi_1) - \text{mean}(\rho_{\text{point}}(\chi_1))] \quad (19)$$

$$\rho_{\text{prob}}(\chi_1) = \rho_{\text{point}}^*(\chi_1) \Big/ \int_0^{2\pi} \rho_{\text{point}}^*(\alpha) \cdot d\alpha \quad (20)$$

To measure the dispersion of the electron density around χ_1, we calculate an "entropy":

$$S = -\sum_i P_i \cdot \ln(P_i) \cong -\sum_i (\rho_{\text{prob}}(\chi_1^i) \cdot \Delta\chi_1) \times \ln(\rho_{\text{prob}}(\chi_1^i) \cdot \Delta\chi_1) \quad (21)$$

where the superscript i indicates values of χ_1 at each interval. The resulting entropy characterizes how movable the X_γ atom is. Its value is greater when the atom vibrates around its position with a greater amplitude and/or the X_γ atom is multi-conformational (has more than one alternative position). In the entropy calculations, we used a 5° step size in χ_1.

Coordinate-based χ_1 Rotameric, Non-Rotameric and Intermediate Side Chain

An amino acid side chain possessing an X_γ atom in its structure (any residue type except glycine or alanine) can be classified according to its χ_1 torsion angle, as determined from the Cartesian coordinates of the crystallographic model deposited in the PDB. We suggest three categories of χ_1 dihedral angle: rotameric, non-rotameric, and intermediate. The classification is based on the value of χ_1 torsion angle side chains have in the PDB structures as shown in Figure 2.

For all side chains, except proline, the rotameric χ_1 are defined as $\chi_1^{\text{PDB}} \in (35°,85°) \cup (155°,205°) \cup (275°,325°)$, non-rotameric χ_1 as $\chi_1^{\text{PDB}} \in (-25°,25°) \cup (95°,145°) \cup (215°,265°)$ and intermediate χ_1 as $\chi_1^{\text{PDB}} \in [25°,35°] \cup [85°,95°] \cup [145°,155°] \cup [205°,215°] \cup [265°,275°] \cup [325°,335°]$. For proline rotameric χ_1 are defined as $\chi_1^{\text{PDB}} \in (25°,45°) \cup (-45°,-25°)$, non-rotameric χ_1 as $\chi_1^{\text{PDB}} \in (-15°,15°)$ and intermediate χ_1 as $\chi_1^{\text{PDB}} \in [15°,25°] \cup [-25°,-15°]$.

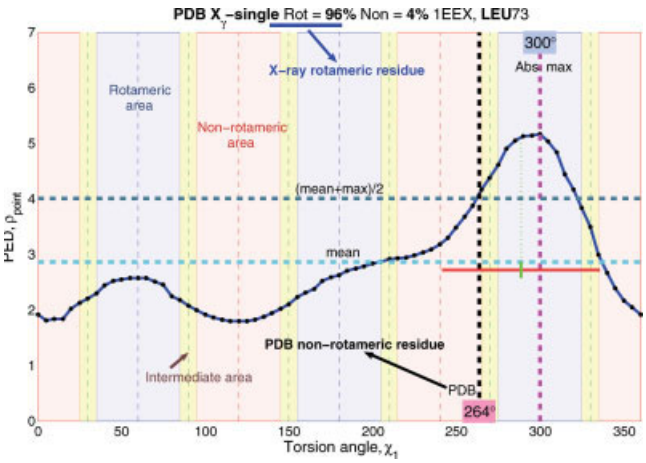

Fig. 2. Point ED-based determination method of χ_1 category: rotameric vs. non-rotameric. Blue, pink and yellow colored areas designate the rotameric, non-rotameric, and intermediate areas respectively. The blue solid line represents X_γ Point ED vs. χ_1. The PDB entry has a non-rotameric χ_1 of 264° (black dashed line); 264° lies just inside in the non-rotameric (215°, 265°) region. The calculation of *Rot* and *Nonrot* ratios (see Methods) indicates that $Rot = 96 \geq 50\%$ (at the top of the plot), and that therefore this leucine side chain is rotameric. In other words, the χ_1 torsion angle could be refined to the rotameric value. In this example the absolute maximum in PED at 300° (purple dash line) is also in the rotameric region. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

χ_1 uncertainty is always present in PDB structures. For the X-ray data it declines with increasing resolution.[28] Since the uncertainty of the position of an atom is roughly one fifth to one tenth of the resolution for high-quality data (R value 0.20 or less), the χ_1 uncertainty for the structures with resolution $\leq$1.7 Å range is $\leq$5–10°. If we used only two categories for side-chain χ_1 classification then the χ_1 uncertainty would allow a number of side chains to belong to both groups: rotameric and non-rotameric and this may bias our results. The proposed 10° gaps eliminate this kind of ambiguity. In this way, residues in the rotameric area are not close to non-rotameric and vice versa.

ED-based Method of χ_1 Category Determination: Rotameric versus Non-Rotameric

The proposed χ_1-rotations (see earlier) allow us to calculate χ_1 based on the observed electron density (in part based on model phases), rather than using the coordinate position as provided in the deposited PDB structure. Here we suggest a way to determine the category of a side chain by analyzing $\rho(\chi_1)$ function.

For any side chain we have point ED as a function of χ_1, $\rho_{\text{point}}(\chi_1)$, where $\chi_1 \in [0°,360°)$. We find the χ_1 range where $\rho_{\text{point}}(\chi_1) > \{\text{mean}(\rho_{\text{point}}(\chi_1)) + \max(\rho_{\text{point}}(\chi_1))\}/2$. This condition is very conservative: it delineates the χ_1 regions where electron density is clearly related to a heavy atom X_γ peak (Fig. 2). For that χ_1 range we calculate the total area (T) between the $\rho_{\text{point}}(\chi_1)$ curve and the $\{\text{mean}(\rho_{\text{point}}(\chi_1)) + \max(\rho_{\text{point}}(\chi_1))\}/2$ cutoff level. Then we split the total area into three components: rotameric (R), non-rotameric (N) and intermediate (I) depending on the value of χ_1 (in accordance with the classification given in the previous section) so that $R + N + I = T$. We define

$$Rot \equiv (R + 0.5 \cdot I)/T \tag{22}$$

to indicate how rotameric the χ_1 conformation is according to the electron density. In contrast,

$$Nonrot \equiv (N + 0.5 \cdot I)/T = 1 - (R + 0.5 \cdot I)/T = 1 - Rot \tag{23}$$

defines how non-rotameric the side chain is.

We say that a side chain is consistent with a rotameric χ_1 conformation according to the electron density if it has $Rot \geq 0.5$, and non-rotameric if $Nonrot \geq 0.5$ (i.e. $Rot < 0.5$). Using the experimental X-ray data (not the model coordinates), this technique allows us to say if some side chains with PDB non-rotameric χ_1 are really closer to a rotameric conformation according to the electron density distribution (see Fig. 2), and vice versa.

While the techniques of finding absolute maximum and calculating *Rot* and *Nonrot* ratios mostly produce identical results, there are occasional cases (not shown) when $\rho_{\text{point}}(\chi_1)$ has a spread-out peak, and the absolute maximum does not accurately designate the χ_1 conformation due to error-level ED fluctuations and incompleteness of the X-ray data. For that reason we chose the proposed, more precise and robust *Rot* and *Nonrot* measures to determine whether a side chain is rotameric or non-rotameric.

Datasets

The protein structure coordinates (atomic models) solved by X-ray diffraction and their corresponding structure factors were taken from the PDB. Protein entries that did not have structure factors stored in the PDB and non-X-ray-crystallographic entries were excluded. The remaining entries were submitted to the web server PISCES[29] to select subsets satisfying resolution, R-factor, and sequence identity criteria. Three datasets were prepared. The first dataset (dataset 1) was derived using the parameters: (0, 1.5] Å resolution range, R-factor $\leq$ 0.15, minimum sequence length of 50 residues, and the maximum sequence identity of any two proteins in the set was 75%. The second dataset (dataset 2) parameters were defined as: (1.5, 3.0] Å resolution range, R-factor $\leq$ 0.25, at least 50 residues length, and 10% maximum sequence identity. The dataset 1 contained 274 entries and dataset 2 gathered 1866 entries. Datasets 1 and 2 were selected in February 2005 and were used primarily for our initial analysis and the development of methods. For the application of the proposed methods, a third high-resolution dataset (dataset 3) was chosen in November 2005, consisting of 1205 structures with a high resolution range of (0, 1.7] Å,

R-factor less than or equal to 0.2, sequence length greater than 50 amino acid residues and mutual percentage identity less than 50%.

CNS has strict requirements on the format of input files. The major format discrepancies are fixed by our programs before passing the input data to CNS-Solve. However, not all errors can be fixed because some input data may be missing. For example a few structure factor files are deficient in both the amplitude and intensity standard deviations that are required for CNS-Solve to build an electron density map. To calculate good ED maps and eliminate any possible input errors, we checked crystallographic R-factors produced by *model_map.inp* with those stated in the PDB files, and skipped any entries having a difference between them greater than 10%. The reasons of the high R-value difference for some of them are mostly due to discrepancies in the input data. This is described in detail by the EDS server developers.[30] So after satisfying the CNS-Solve and R-factor requirements the sizes of datasets 1, 2, and 3 were reduced to 238, 1495, and 1048 structures respectively. The application dataset 3 contains 441,769 amino acid residues in total. The larger high-resolution dataset 3 was used especially for the analysis of non-rotameric side chains, which are rare in very high-resolution structures.

RESULTS

Atom Confidence Levels: Point Electron Densities, Integrated Electron Densities, and B-factors

As an example of comparing high and low electron density side chains, in Figure 3 we show examples of two aspartic acid residues from PDB entry 1GA6.[31] In Figure 3(A1), an aspartic residue (Asp18A) with higher density and lower B-factors at every reported atom position is shown, while in Figure 3(B1), an aspartic acid residue (Asp105A) with lower density and higher B-factors is shown. This latter residue is obviously more mobile than the first one, such that the positions of these atoms are less certain. The general electron density feature of atoms with higher vibration is that their electron clouds are less dense and more spatially dispersed. Thus the IED geometric means for the backbone and side chains of Asp 18A are both 1.08, while those for Asp108A are 0.78 and 0.77 respectively. In Table SI, Supplementary Material, we provide the calculated PED and IED values and other data for these two residues.

In the Supplementary Material, we provide tables of the mean PED and IED values for dataset 1 for all atom types as well as the side-chain geometric means [Eqs. (17) and (18)]. Carbon atoms have means for both PED and IED close to 6, for nitrogen atoms both PED and IED are $\sim$7, for oxygen atoms both PED and IED are $\sim$8, for sulfur atoms PED is $\sim 15 \pm 4$ and IED is $\sim 11 \pm 3$, and for selenium atoms PED is $\sim 20 \pm 10$ and IED is $\sim 14 \pm 6$. We might expect the selenium values to be higher, relative to the other atom types, and there are several reasons this may not be so: (1) there are not enough data on selenomethionine consisting of only 68 selenomethionine residues with a high standard deviation; (2) the proposed linear normalization is less accurate at the higher ED range; and (3) the radius of integration may be too small for IED.

Lovell et al.[17] have used the Debye–Waller temperature factors (B-factors) to eliminate side chains from a data set that may have inaccurate or uncertain coordinates. The use of maximum B-factor cutoffs resulted in a rotamer library with lower standard deviations for dihedral angles and fewer examples of rare or unfavorable rotamers. In this paper, we are deriving an electron density criterion rather than author-provided B-factors for a similar purpose, and it is fair to ask whether these two methods agree on which side chains coordinates are of questionable quality.

We analyzed the relationship between the PED, IED, and the corresponding Debye–Waller temperature B-factors. Based on our derivation we assumed that both ρ_{point} and ρ_{integ} are proportional to:

$$f(B) = \left[r_{\text{coval}} + \left(\frac{B}{8 \cdot \pi^2} \right)^{\frac{1}{2}} \right]^{-3} \tag{24}$$

(see Appendix A). To check these two relationships and their degree of correlation, the N, C_α, C, O backbone atoms from two datasets were used: (a) the (0, 1.5] Å high-resolution dataset 1; (b) the (1.5, 3.0] Å low-resolution dataset 2. The assumed relationship was confirmed and demonstrated a very strong correlation for both (ρ_{point}; B) and (ρ_{integ}; B) as shown in Table I. As expected, higher values for the B-factor correspond to lower electron density at the atomic coordinate position due to disorder. For some proteins in dataset 1 [Fig. 4(A)] and dataset 2 [Fig. 4(B)] there is a very good correlation between (ρ_{point}; $f(B)$). The regression lines correlate well with the number of electrons for each atom type (8 for O, 7 for N, and 6 for C and C_α).

The question remains whether it is better to use the electron density measures ρ_{point} and ρ_{integ} or the more traditional B-factors. As shown in Table I and Figure 4(A), these two should be strongly correlated via Eq. (24). However, we found many structures where the correlation was much lower, as shown in Figure 4(C,D), and for these structures the B-factors do not give as good a measure of uncertainty in coordinate positions. There are even some low-resolution entries where the B-factors were restrained to a constant value and not used in refinement. For example, PDB entry 2AYU[32] (3.0 Å resolution) has B-factors of 15Å^2 for all atoms. For structures with low correlation of B-factors to electron density, the B-factors do not apparently give a good measure of uncertainty in coordinate positions. In Figure 5, we show the (ρ_{point}; $f(B)$) correlation coefficient dependence on the X-ray resolution. At lower resolution, the correlation is weaker, and it appears that ρ_{point} and ρ_{integ} provide information not contained in the B-factors. While the majority of structures have strong (ρ_{point}; $f(B)$) correlation, in each resolution bin there are some structures that have very low correlation between (ρ_{point}; $f(B)$), even at high resolution.

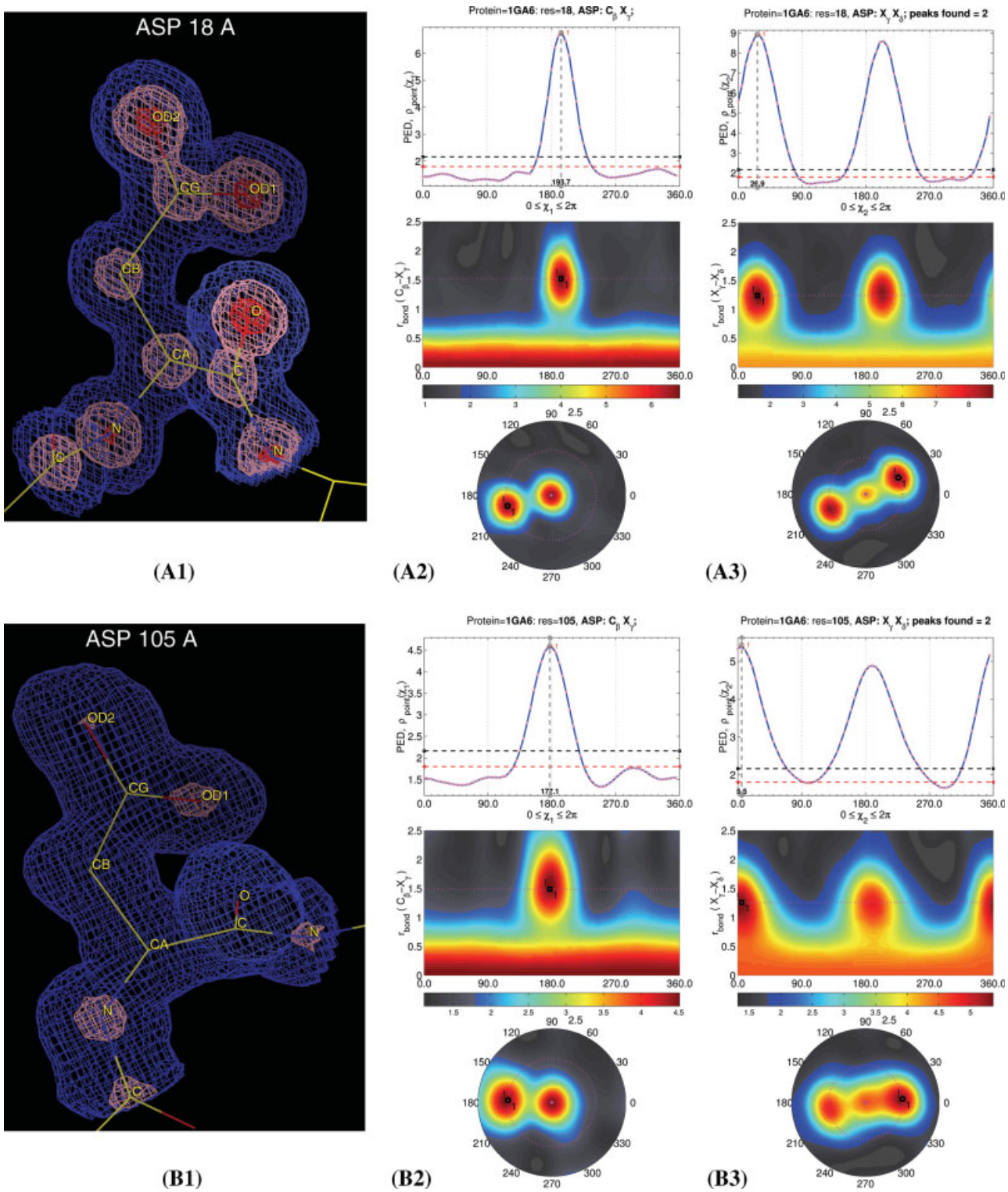

Fig. 3. Two single-conformational aspartic acid residues Asp18A (A) and Asp105A (B) of PDB entry 1GA6 at 1.0 Å resolution.[31] Asp18A B-factors are lower than Asp105A B-factors; Asp18A point and integrated EDs are greater than Asp105A EDs (Table S1 in the Supplemental Material). (**A**1,**B**1) F_O electron-density contours; (1σ, 8σ, 16σ) contour levels (blue, pink, red) for Asp18A and Asp105A respectively; the scales are the same for both residues. The ED is lower and more spread out in case of Asp105A. (**A**2,**B**2) χ_1-rotations of C_γ around $C_\alpha-C_\beta$. a: point ED vs. χ_1; b: point ED vs. χ_1 and $C_\beta-C_\gamma$ in rectangular coordinate system; c: point ED vs χ_1 and $C_\beta-C_\gamma$ in polar coordinate system. The positions of C_β and C_γ are clearly detectable. (**A**3,**B**3): χ_2-rotations of $O_{\delta1,2}$ around $C_\beta-C_\gamma$. a: point ED vs. χ_2; b: point ED vs. χ_2 and $C_\gamma-O_\delta$ in rectangular coordinate system; c: point ED vs. χ_2 and $C_\gamma-O_\delta$ in polar coordinate system. The positions of C_γ, $O_{\delta1}$, and $O_{\delta2}$ are clearly detectable.

TABLE I. Correlation Between Debye–Waller Temperature B-Factors and Point Electron Density, ρ_{point}, and Integrated Electron Density, ρ_{integ}, for Backbone Nonhydrogen Atoms N, C_α, C, and O for Two Resolution Ranges

		Mean correlation coefficient ± standard deviation									
		Point electron density vs. *f(B)*					Integrated electron density vs. *f(B)*				
Resolution	Prot no. (res no.)	N	C_α	C	O	ALL	N	C_α	C	O	ALL
(0.0, 1.5] Å[a]	238 (73,689)	93 ± 8	89 ± 10	88 ± 10	94 ± 7	91 ± 9	90 ± 9	86 ± 11	84 ± 11	93 ± 8	88 ± 10
(1.5, 3.0] Å[b]	1495 (683,113)	83 ± 13	76 ± 14	82 ± 12	88 ± 12	82 ± 13	82 ± 13	75 ± 14	81 ± 12	87 ± 12	81 ± 14

Correlation coefficients and their standard deviations were calculated for the pairs (1) *f(B)* vs. point ED and (2) *f(B)* vs. integrated ED, where $f(B) = \left[r_{coval} + \left(\frac{B}{8\pi^2}\right)^{\frac{1}{2}}\right]^{-3}$ (see Appendix).

[a]Dataset 1.
[b]Dataset 2.

Finally, B-factors have different scales in different structures in the PDB and it is difficult to compare atom displacements between two protein entries if their B scales are defined in different ways. The different scales and different low and high cutoffs for B-factor values may depend on the refinement package used and how B-factor values are determined.[33] A number of structures have minimum or maximum value cut-offs for B-factors. Because of these considerations, some authors have normalized the B-factors in a protein before comparing different crystal structures.[34–38]

Taking into account these considerations, we believe that a uniform method of calculating normalized atom confidence levels based on an ED map may be used in addition to the B-factors stored in PDB entries, and in some cases they provide higher reliability for evaluation of atom positions. We should point out that the atomic model information has been used in the map calculations as described in Methods. Therefore, the atom confidence levels are to some degree model-biased, and the model includes the B-factors.

It is also necessary to determine whether ρ_{integ} or ρ_{point} provides better information for assessing the quality of protein side chains. By looking through a large number of such plots for proteins with different resolution, we found that in general the scatter plots for (ρ_{point}; *f(B)*) and (ρ_{integ}; *f(B)*) are almost absolutely the same except for very few data points. However, for some atom positions, ρ_{point} and ρ_{integ} differ. In these few cases, ρ_{integ} appears to be a more robust measure than ρ_{point}. The advantage of ρ_{integ} over ρ_{point} can be only demonstrated at very high resolution when ED maps are very detailed and precise. In all other cases the differences are insignificant. In general, it is much faster to calculate ρ_{point} than ρ_{integ} so it is used especially for the two dimensional plots, $\rho(\chi_1, r_{\beta\gamma})$.

Energetically Preferable Side-Chain Configurations Have Higher Electron Density

Side-chain torsion angles are not evenly distributed and instead concentrate in tight clusters (rotamers) around certain values. This division can be explained in physical–chemical terms, in terms of repulsion of bonding molecular orbitals of the 1–2 and 3–4 bonds as well as steric repulsion between atoms 1 and 4.[39] For most of the χ dihedral angles of amino acid side chains, those with rotation about sp^3–sp^3 bonds, there are three minima of the potential energy observed at or near the (60°, 180°, and 300°) χ values (g^+, t, and g^- respectively). Therefore, these staggered conformations are most likely to be populated in the side-chain torsion angle distributions.

We examined electron density versus χ_1 scatter plots for the high-resolution protein structures (dataset 1) and observed that non-rotameric conformations tend to have much lower X_γ electron density than average, as shown in Figure 6(A) for glutamic acid C_γ and Figure 6(B) for serine O_γ. Only a few of the non-rotameric conformations (between clusters) have high confidence levels. In Figure 6(C,D), the average values for the integrated ED are shown in 20° bins, clearly demonstrating the 3-fold periodicity associated with staggered and eclipsed conformations of sp^3–sp^3 bonds. The results for all side chains are given in the Supplementary Material.

With Increasing Atom Confidence Levels (ED) the Variance of g^+, t, and g^- Rotamers Goes Down and Their Means Approach the Canonical Values

The large high-resolution (0,1.7] Å dataset 3 was processed, and for each amino acid residue type with a γ heavy atom, χ_1 and the corresponding X_γ atom confidence level $\rho_{point}(X_\gamma)$ was calculated. Within each residue type χ_1 values were divided into three rotamer groups: g^+ [0°,120°), t [120°,240°), and g^- [240°,360°); proline has only the g^+ [0°,45°) and g^- [315°,360°) rotamers. For each rotamer type, pairs $\chi_1 \leftrightarrow \rho_{point}(X_\gamma)$ were arranged according to their $\rho_{point}(X_\gamma)$. The bin intervals were chosen to cover the whole ED confidence level range with sufficient statistics in each bin. Every bin accommodated a minimum of 50 side chains. The χ_1 means and χ_1 standard deviations were calculated and plotted for each of the bins against the $\rho_{point}(X_\gamma)$ mean. The standard deviations are shown in Figure 7(A), the means in Figure 7(B), and the populations among the three rotamers in Figure 7(C) for selected residues.

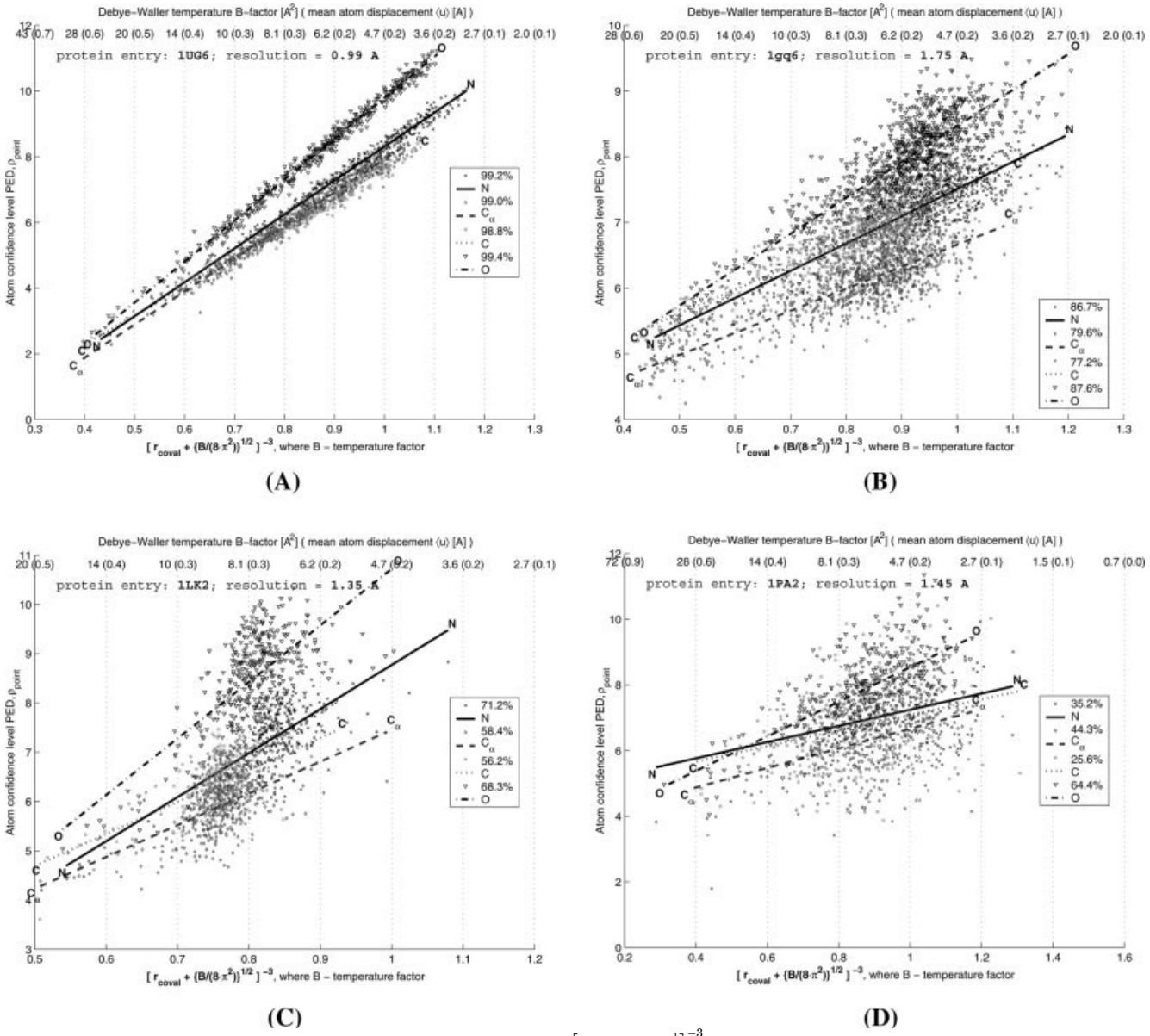

Fig. 4. The point ED ρ_{point} of the backbone atoms N, C_α, C, O vs. $\left[r_{coval} + \left(\frac{B}{8\cdot\pi^2}\right)^{\frac{1}{2}}\right]^{-3} = f(B)$ scatter plots. The regression line for each backbone atom is plotted. The (ρ_{point}; $f(B)$) correlation coefficients are indicated for each backbone atom type in the legend in each plot. At the top of the plots the OX-axis is scaled in the units of B-factor and average atom displacement $\langle u \rangle$. The atom regression lines are shifted according to the number of electrons the atom possesses. (**A**) High correlation: 1UG6 at 0.99 Å resolution; (**B**) high correlation (at the low-resolution range): 1GQ6 at 1.75 Å resolution; (**C**) and (**D**) unusually low correlation for the high-resolution structures: (C) 1LK2 at 1.35 Å resolution and (D) 1PA2 at 1.45 resolution.

The results for all side chain types are given in the Supplementary Material. In Table II, the decrease in standard deviation from the lowest ED bin to the highest ED bin is given for each amino acid type and rotamer.

For all analyzed residues, as shown in Table II, each of the three χ_1 rotamers (two for proline) has decreasing standard deviation of χ_1 with increasing electron density atom confidence level $\rho_{point}(X_\gamma)$. The largest decreases in standard deviations belong to arginine (16.6°), glutamic acid (16.5°), and methionine (15.7°). Conversely, tryptophan, proline, tyrosine, phenylalanine, and histidine have the smallest decreases in standard deviation of 3.6°, 4.5°, 7.5°, 7.6°, and 9.3° respectively. Over all of the amino acids, the largest decreases belong to long flexible side chains such as arginine and lysine and small side chains such as serine, while the smallest decreases belong to proline and the aromatic residues. For the latter, the large electron density in the ring presumably makes locating the γ atoms fairly straightforward.

The same type of analysis was done for the χ_1 means of the g^+, t, and g^- rotamers [Fig. 7(B)]. We found that the g^+, t, and g^- means of χ_1 move closer to their canon-

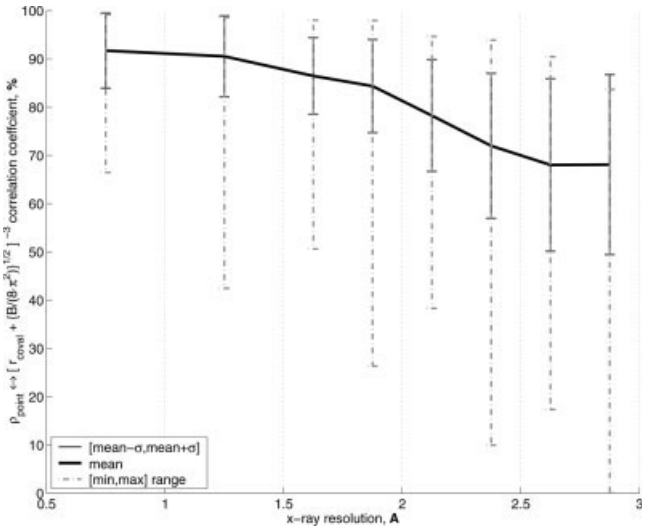

Fig. 5. Mean correlation coefficient (ρ_{point}; $f(B)$) vs. mean X-ray diffraction resolution. The solid and dashed vertical lines represent the standard deviation and min/max of the correlation coefficient respectively for each resolution range bin. The correlation between the atom PED confidence level ρ_{point} and Debye–Waller temperature B-factor steadily increases with increasing resolution. The correlation coefficient standard deviation slightly reduces with higher resolution. In each resolution bin there are some PDB entries having extremely low correlation coefficients.

ical values with increasing X_γ ED confidence level. The strength of the effect varies with amino acid type. The leucine residues have a change in the mean dihedral angles of 15.2°, glutamic acid: 6.9°, proline: 6.3°, arginine: 5.9°, lysine: 5.2°, all others: in the (2°, 5°) range. With a few exceptions the average angles change in the same direction for a particular rotamer across side-chain types. For both the g^+ and g^- rotamers, the angles increase toward their canonical values with increasing density, while the angles decrease for the t rotamers. Our results are consistent with those of MacArthur and Thornton.[40] They found that variance in dihedral angle about rotameric positions continues to decrease as a function of higher resolution, and they interpreted this as the result of more side chains being modeled in multiple conformations, rather than single, non-rotameric conformations.

g^+, t, and g^- Rotamer Populations versus $\rho_{point}(X_\gamma)$: Convergence of Their Proportions

The large, high-resolution dataset 3 was used to track how the χ_1 rotamer populations vary with the X_γ point ED confidence level. The ED range was split into bins with a more sophisticated sample size technique as described in Appendix B. For each ED bin, the g^+, t, and g^- rotamer populations were calculated. All residue types experience rotamer population fluctuations in the lowest ED region where χ_1 uncertainty is highest. After passing an atom confidence level threshold, the rotamer population trends stabilize and the g^+, t, and g^- proportions converge [Fig. 7(C)].

Entropies Calculated from χ_1 ED Probability Distributions as an Indication of Disorder

Side chains with low electron density at the model position also tend to have electron density that is significantly spread out rather than localized at the X_γ position. We calculated entropy as described in Methods, by summing over χ_1 in a 5° step. To obtain more reliable statistics, the larger dataset 3 was used

The side chains in dataset 3 were divided into three groups, depending on the χ_1 dihedral angle calculated from the model coordinates from the PDB and whether there was more than one conformation annotated in the PDB entry: (1) all single-conformational side chains, (2) single-conformational side-chains with non-rotameric χ_1 (as defined earlier) and (3) multi-conformational side chains having alternative X_γ atoms at least 60° apart (for proline—at least 30° apart).

For the single-conformational side chains with any χ_1 (group 1) we observe a strong entropy decrease with increasing $\rho_{point}(X_\gamma)$, as shown in green in Figure 8(A1,B1,C1). The χ_1 non-rotameric side-chains (group 2, shown in blue) exhibit significantly higher entropy although they are much less represented (1.6% of all side chains in data set 3). The multi-conformational side chains (group 3, in red) demonstrate even higher entropy than the non-rotameric side-chains. These features of the three groups occur for all side-chain types as shown in Figure 9. Note that the calculated entropy [Eq. (21)] for Val, Ile, and Thr is artificially high because of the presence of two γ heavy atoms.

The results indicate that many of the non-rotameric side chains have electron density distributions that are potentially consistent with multi-conformational side chains, and in many cases these side chains could probably be refined to two or more rotameric positions.

Some Non-Rotameric Side Chains Are Incorrectly Modeled and Exhibit Rotameric χ_1 in Their Electron Density Maps

We examined if non-rotameric side chains in PDB models really have non-rotameric χ_1 in their ED maps. The calculations were done for all residue types except proline—owing to its unique χ_1 rotamer nature and valine, isoleucine and threonine—residues having more than one X_γ atom and requiring more complicated analysis.

The PDB single-conformational side chains were split into three subgroups: rotameric, non-rotameric and intermediate (as described earlier; Fig. 2) according to their χ_1 values calculated from the deposited coordinates. The percentages in each category are shown in Table III. For each of those three subgroups we applied the proposed *Rot* and *Nonrot* measures as calculated from the electron density (see earlier) to determine how many of the side chains with non-rotameric and intermediate χ_1 might be refined to rotameric χ_1, and how many side chains with rotameric χ_1 might have ED consistent with non-rotameric χ_1 as a control point. For

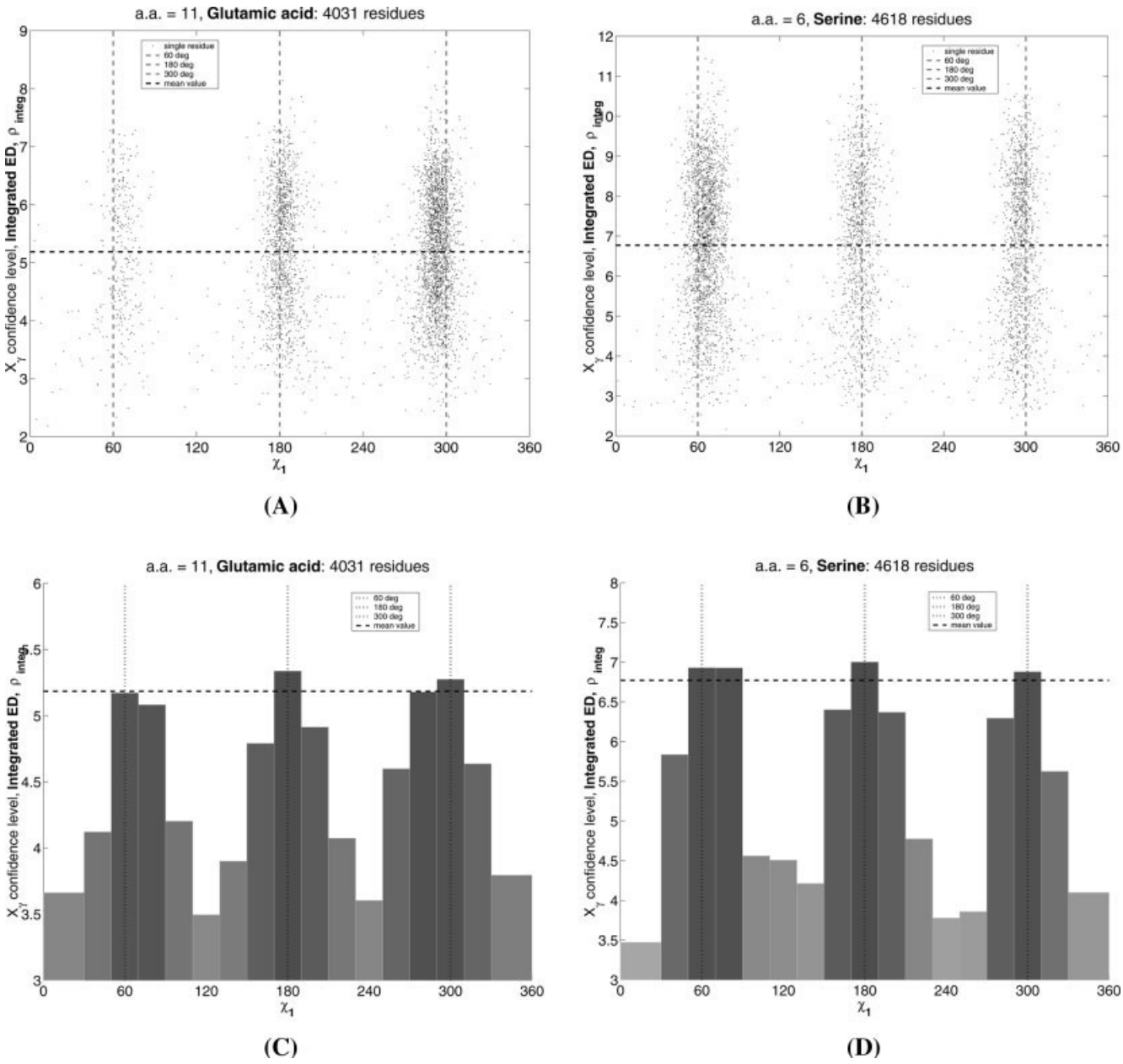

Fig. 6. Electron density levels vary with χ_1 angles. (**A**,**B**) X_γ integrated ED vs. χ_1 scatter plots for the high-resolution protein structures (dataset 1) for glutamic acid and serine respectively. (**C**,**D**) mean X_γ integrated ED vs. mean χ_1 for the 20° bins centered on the canonical values (60°, 180°, 300°). The vertical lines designate the idealized *gauche*$^+$, *trans*, *gauche*$^-$ χ_1 values. The horizontal lines show the mean electron density for the whole [0°, 360°) χ_1 range. The averaged atom confidence levels have maxima approximately at the idealized rotameric positions and minima in the middle between them.

those potentially misclassified residues (e.g., non-rotameric residues in the PDB coordinates that have rotameric density), we further divided them into those with entropy above and below the mean plus one standard deviation (discussed in next section). These calculations were performed for the comprehensive dataset 3, and the results are shown in Table IV.

We found that 15% of all investigated χ_1 PDB-non-rotameric residues are actually more consistent with rotameric conformations, according to their electron density distributions. Leucine, arginine, glutamic acid, lysine, and glutamine have the highest percentages of incorrectly modeled non-rotameric side chains—21, 20, 19, 19, 18% respectively. The lowest percentages belong to tyrosine (2%), phenylalanine (3%), histidine (5%), asparagine (6%), and tryptophan (7%). The aromatic residue types are less likely to have incorrectly modeled χ_1 because of the large size of the aromatic rings, which are easy to identify in electron density maps.

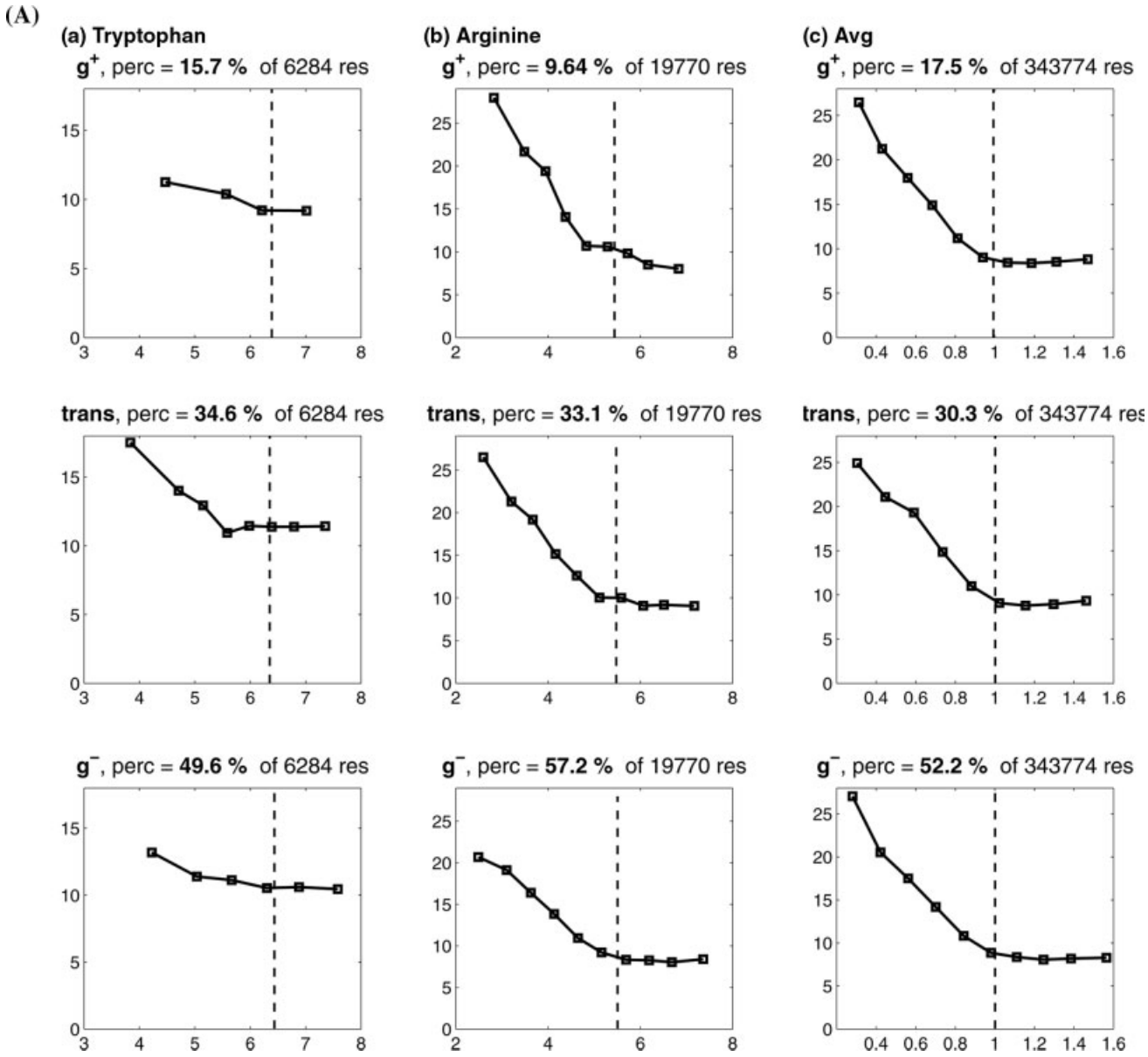

Fig. 7. (**A**) g^+, *trans*, and g^- χ_1 standard deviations measured in degrees vs. X_γ point ED for (**a**) tryptophan, (**b**) arginine and (**c**) averaged over all residue types. Tryptophan experiences the smallest drop in standard deviation (Table II) while arginine the largest. The vertical lines designate the average ED of the g^+, *trans*, and g^- rotamers. Each data point represents at least 50 side chains. The total number of residues of each type and their relative g^+, *trans*, and g^- percentages are indicated above the plots. In (c), X_γ electron density was divided by the whole-PED-range X_γ mean specific for each residue type in order to scale all residue types to the same PED scale. (**B**) g^+, *trans*, and g^- mean χ_1 dihedral angles measured in degrees vs. X_γ point ED for (**a**) serine, (**b**) glutamic acid, (**c**) leucine and (**d**) proline. The vertical lines designate the average ED of the g^+, *trans*, and g^- rotamers. The horizontal lines indicate the average χ_1 of the g^+, *trans*, and g^- rotamers calculated for the whole ED range. Each data point represents at least 50 side chains. The total number of residues of each type and their relative g^+, *trans*, and g^- percentages are indicated above the plots. (**C**) g^+, *trans*, and g^- rotamer populations vs. ρ_{point} (X_γ). g^+ in light gray bars, trans in medium gray bars, g^- in dark gray bars. The populations are given in percents. The sample sizes are selected to guarantee with 80% confidence that the populations lie in a 5% relative error interval. The dashed horizontal lines indicate the rotamer populations calculated for the whole ED range. The vertical line shows the ρ_{point} (X_γ) mean. The total number of residues is indicated in the top of the plots. (**a**) serine, (**b**) glutamic acid, (**c**) valine and (**d**) cysteine.

A total of 60% of the PDB χ_1 intermediate side chains have rotameric conformations in their ED distributions. To estimate consistency of our analysis we checked how many PDB χ_1 rotameric side chains can be refined to non-rotameric torsion angles. The data demonstrate that only 1% of these have this property, compared with 15% of the PDB non-rotameric side chains. The data contradict the notion common 10–15 years ago that protein side chains need not be rotameric because of strong environmental forces.

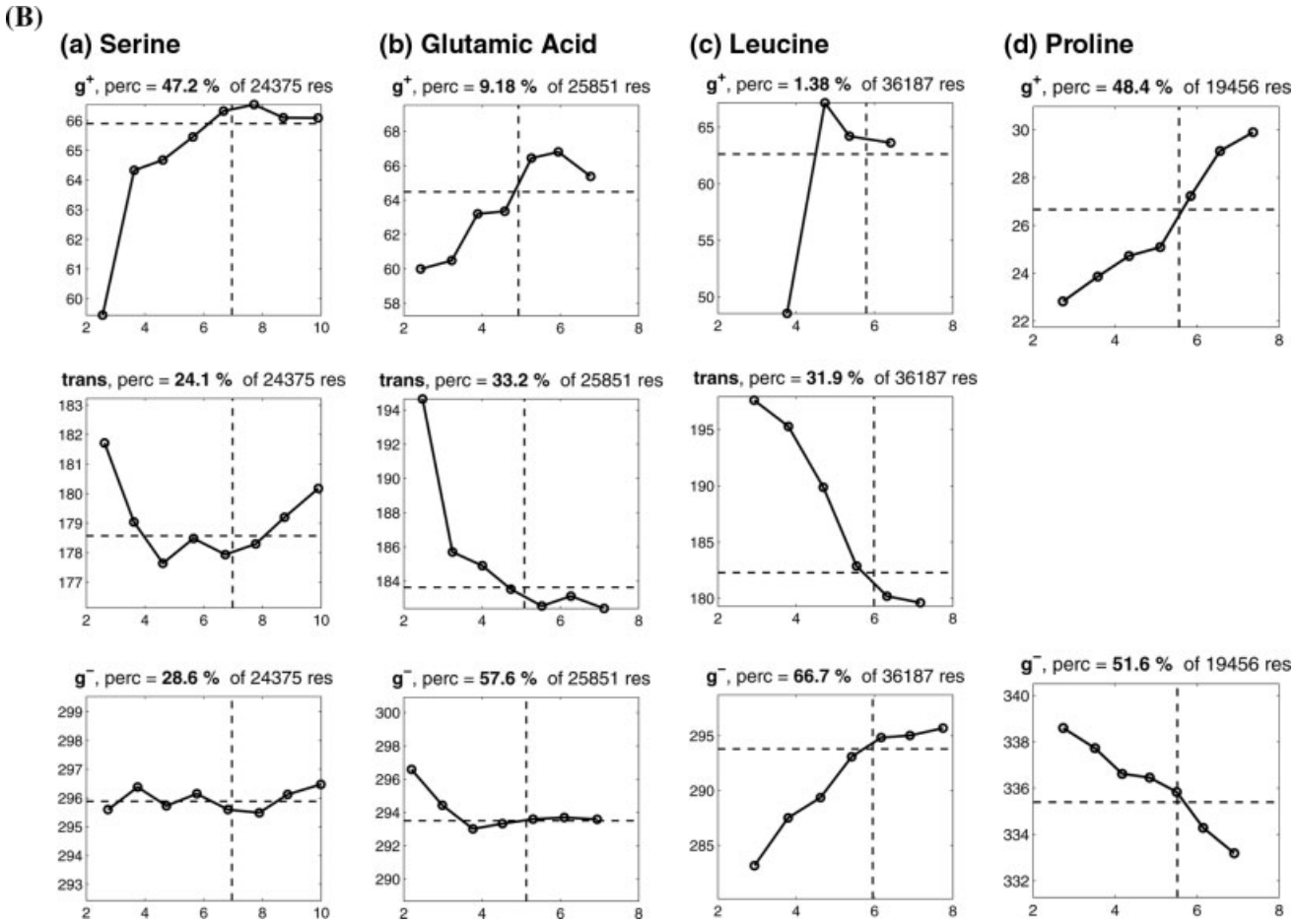

Figure 7. (Continued.)

Most PDB Non-Rotameric χ_1 Side Chains Have High Entropy and Are Not Fixed in Those Positions

Non-rotameric side chains not only have low electron density (as shown in Fig. 6) but consistent with this, they also have high entropy (Figs. 8 and 9; Table IV). We have demonstrated that 15% of PDB χ_1 non-rotameric residues are more consistent with rotameric conformations, 60% of χ_1 intermediate residues are X-ray rotameric. But 85% of non-rotameric and 40% of intermediate residues do not have rotameric electron density in our measurements, and these groups are worth further investigation.

We suspect that many non-rotameric side chains are in fact significantly disordered, moving between rotamers, whether or not they spend significant time between rotamers at room temperature or at the temperature before flash-freezing. We therefore investigated the χ_1 entropy dependence on the atomic model χ_1 for dataset 3. For every residue type (except proline) averaged entropy has strong minima at the canonical positions of the g^+, t, and g^- rotamers [Fig. 8(A2,B2,C2)] and maxima between them at about 120°, 240°, and 360°. Proline has minima at about 30° and −30° and a maximum at 0°. In terms of the side-chain dynamics these entropy results indicate that non-rotameric side chains are highly mobile and tend to exhibit more than one conformation. The results as shown in Figure 8 for all other residue types are given in the Supplementary Material.

To quantify the percentage of the disordered χ_1 non-rotameric and intermediate residues, we calculated the χ_1 entropy mean and standard deviation of the PDB single-conformational rotameric residues (as defined earlier) for every residue type analyzed. The entropy mean plus one standard deviation was used as a cutoff value to distinguish "disordered" side chains from more "ordered" ones (Table IV). Residues having entropy above the cutoff level are considered "disordered," and those below the cutoff level as "ordered." This value is relatively arbitrary, but serves as a reasonable reference point for comparing different sets of side chains (PDB-rotameric, PDB-non-rotameric, PDB-intermediate, etc.).

As discussed above, a total of 15% of the PDB χ_1 non-rotameric residues have ED's more consistent with rotameric χ_1. Of the remaining 85%, those that have ED in

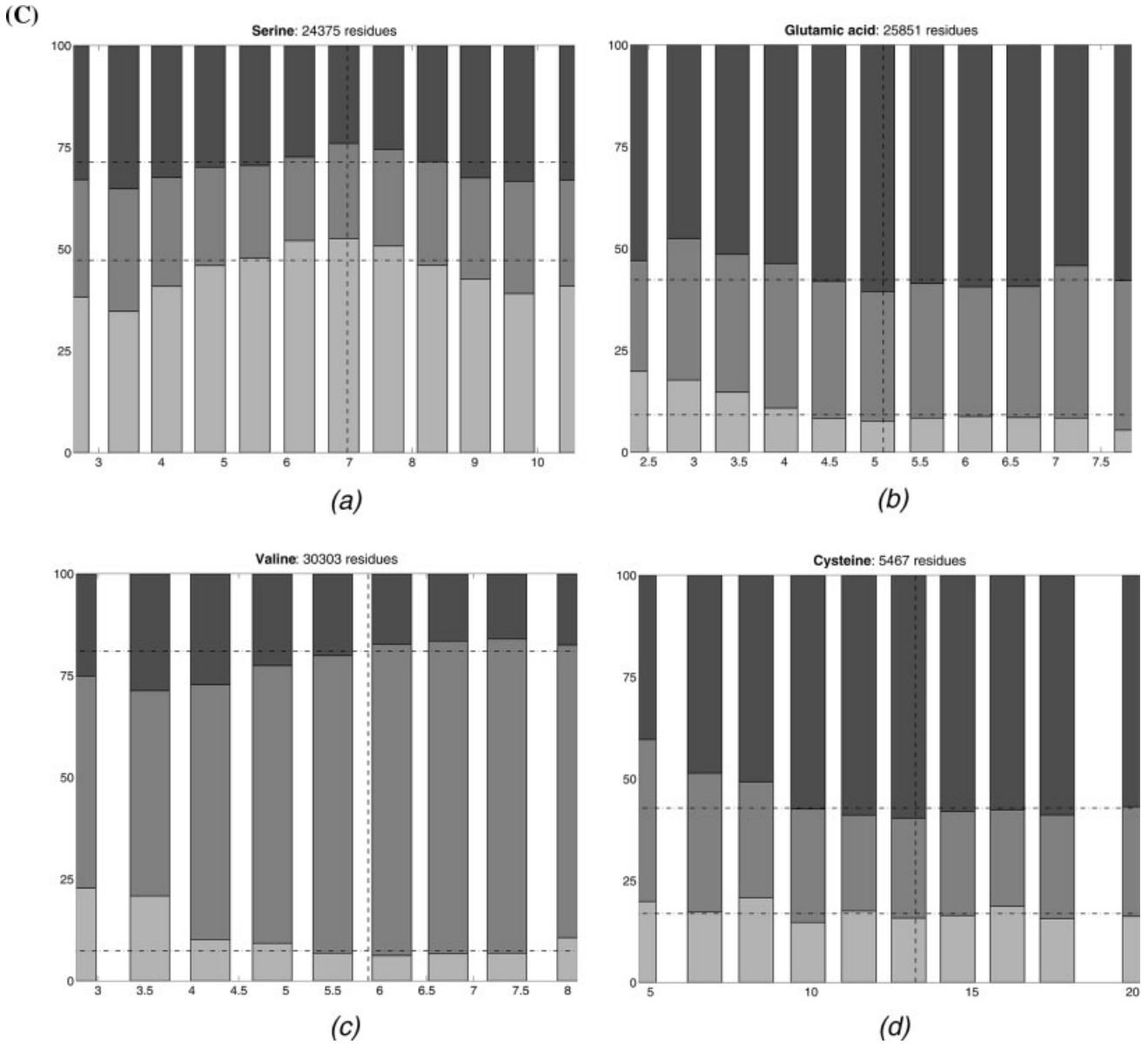

Figure 7. (Continued.)

the non-rotameric region of χ_1, 47% are "disordered" ($S \geq \overline{S} + \sigma$), and the rest, 38%, are "ordered" ($S < \overline{S} + \sigma$). Among the χ_1 intermediate residues those percentages are 60% rotameric, 15% non-rotameric and disordered, and 25% non-rotameric and ordered.

As a control point we estimated how many PDB-rotameric residues that are also rotameric in their electron density are disordered or ordered according to the entropy calculation. As shown above, only 1% of the PDB rotameric residues are more consistent with non-rotameric χ_1 according to their ED distributions. Of the PDB and ED rotameric residues (99% of PDB-rotameric residues), 11% are disordered and 88% are ordered according to our entropy criterion. To verify if the entropy cutoff was reasonable, we did the same calculations for the PDB multi-conformational residues with χ_1 at least 60° apart. Of these residues, 74% of them have entropy above the cutoff and are designated in our terminology as disordered.

We can compare these results with those of Petrella and Karplus[18] who used energy minimization of side chains to determine whether non-rotameric side chains were in local energy minima, or upon minimization would move into rotameric positions. Their definitions of rotameric and non-rotameric are slightly different from ours. Nevertheless, they found that 36% of non-rotameric side chains minimized into rotameric positions, while only 2.4% of rotameric side chains minimized into

TABLE II. Decrease in Dihedral Angle Variance of g^+, trans, g^- χ_1-Rotamers Comparing Side Chains With the Lowest and Highest X_γ Point Electron Density ρ_{point}

		χ_1 sigma decrease			
No.	Res	g^+	Trans	g^-	Average
1	TRP	2.1	6.1	2.7	3.6
2	PRO	3.6	—	5.3	4.5
3	TYR	12.4	4.3	5.7	7.5
4	PHE	7.6	9.4	5.7	7.6
5	HIS	7.7	9.4	10.7	9.3
6	ILE	9.6	10.5	9.5	9.9
7	ASP	9.3	9.0	12.5	10.3
8	ASN	13.7	7.5	12.4	11.2
9	VAL	13.3	10.5	12.5	12.1
10	LEU	13.4	12.2	11.8	12.5
11	THR	11.5	12.3	14.7	12.8
12	GLN	17.6	12.8	11.2	13.8
13	SER	16.9	14.9	11.4	14.4
14	LYS	19.7	13.6	11.8	15.0
15	CYS	10.7	17.1	18.4	15.4
16	MET	18.1	17.4	11.5	15.7
17	GLU	18.3	15.7	15.4	16.5
18	ARG	19.9	17.4	12.3	16.6
WPYFH		12.6	9.0	10.6	10.7
IDNVL		11.8	13.7	11.4	12.3
TESKCMER		17.5	15.3	19.6	17.5
ALL	ANY	17.7	15.6	18.8	17.3

non-rotameric positions. Our results based only on experimental data are in reasonable agreement with their results based purely on energy calculations.

Side Chains with Coordinates Clearly Consistent with Their ED Distributions

Before we present the ED features of side chains that may be incorrectly modeled, we would like to demonstrate different types of ED distributions of side chains with electron density highly consistent with the model coordinates. These are of course the large majority of side chains in high-resolution structures. We measured electron density as a function of χ_1 by rotating a pseudo-atom in a circular arc at the bond length and bond angle calculated from the C_β and X_γ coordinates in the PDB file (as described in Methods). We also calculated electron density as a function of two variables, χ_1 and $r_{\beta\gamma}$, the distance from the C_β atom along the C_β—X_γ bond direction.

The resulting plots are shown in Figure 10 for: (A) a single-conformational rotameric residue, Arg 10 in PDB entry 1DY5[41]; (B) a single-conformational non-rotameric residue, Trp 154B in 1GK9[42]; and (C) a multi-conformational rotameric residue, Ser 331 in 1GA5.[43] Each of those three residues has three types of plots in Figure 10: in (I) for each side chain, electron density is shown versus χ_1 alone, and in (II) and (III) we show two versions of ED versus χ_1 and $r_{\beta\gamma}$. The subplot (II) shows the density in a rectangular coordinate system, while (III) shows the density in polar coordinates. In the view in (II), the C_β atom density is spread out along the full bottom of the plot and the X_γ density is spread out vertically. In the view in (III), the C_β density appears in the center of the polar plot and the X_γ density spreads out radially along $r_{\beta\gamma}$.

In Figure 10(A) we see the most common situation found in PDB models—a single-conformational, rotameric side chain (93%, Table III). This arginine's χ_1 is rotameric in a trans conformation with a 173° torsion angle. Its C_γ has a very narrow and strong ED peak. The model dihedral angle very precisely fits the ED distribution leaving no doubts about the χ_1 conformation. In (B) there are plots for a single-conformational, non-rotameric tryptophan with a 233° χ_1—almost in the middle of the non-rotameric χ_1 region between t and g^-. Again, the C_γ ED peak is obvious and narrow, and is nicely fitted with the PDB angle. We discuss this very uncommon side-chain conformation in detail later in the paper.

Depositors of structures to the PDB can indicate multiple positions for atoms (labeled A, B, etc.) and occupancies less than 1.0. The electron density plots for these residues typically look like the multi-conformational serine shown in Figure 10(C), which shows annotated occupancies of 68 and 32% for the A and B rotameric conformations at χ_1 values of 173°(*trans*) and 299°(g^-) respectively. Both model torsion angles agree with the ED peak positions.

The ED Distributions of Non-Rotameric, Low-Density, and/or High-Entropy Side Chains

Of greater interest are the electron density distributions of non-rotameric side chains and those with high B-factors, for which determination of the correct coordinates in the model is more difficult and may be in some cases not ideal. For rotameric side chains with low electron density and high entropy, we found a number of side-chain ED distribution patterns very similar to the declared multi-conformational ones [Fig. 10(C)], but these side chains were not annotated as multi-conformational in their atomic models in the PDB. We show two examples in Figure 11: Ser 4010 in 1G61[44] [Fig. 11(A)] and Lys 145 in 1A6M[45] [Fig. 11(B)]. We believe the serine [Fig. 11(A)] should be reported as a g^+ and g^- rotameric multi-conformational side chain and the lysine [Fig. 11(B)] as a t and g^- rotameric multi-conformer.

Then we analyzed non-rotameric side chains that either have rotameric electron density or have high entropy. These residues constitute 62% of all PDB non-rotameric side chains (Table IV, 62% = 15%+47%). We found the following four common cases for these side chains. As an example of a PDB-non-rotameric side chain that has ED that is more consistent with a rotameric conformation, in Figure 11(C) we show a clearly single-conformational rotameric leucine while its χ_1^{PDB} has a non-rotameric value of 225°. Another representative case is shown Figure 11(D) in which the non-rotameric PDB model occurs between two clearly defined

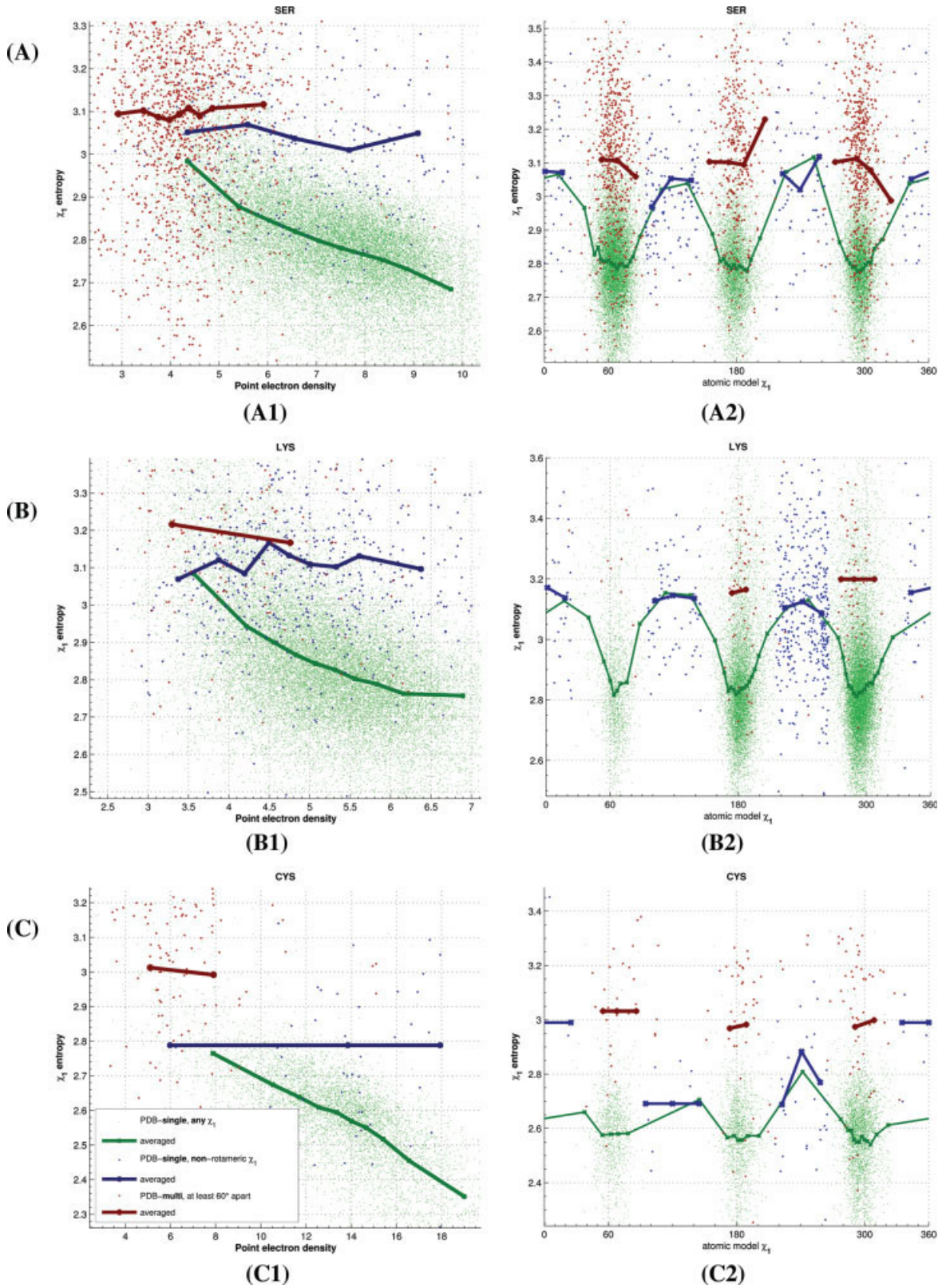

Fig. 8. χ_1 entropy versus point electron density and χ_1 dihedral angle for : (**A**) serine, (**B**) lysine, (**C**) cysteine. (A1,B1,C1) χ_1 rotamer entropy vs. X_γ point electron density ρ_{point} scatter plots. (A2,B2,C2) χ_1 rotamer entropy vs. atomic model χ_1 scatter plots. Each residue is represented by one dot on the entropy-PED and entropy-χ_1 plots. The solid lines represent the averaged entropy vs. averaged PED atom confidence level (A1,B1,C1) or averaged model torsion angle (A2,B2,C2). The χ_1 axis (A2,B2,C2) has ticks and dashed lines to designate the canonical g^+, *trans*, g^- χ_1 torsion angles at 60°, 180°, and 300° respectively. Three residue types are shown. Three sets of data are presented on each plot: single-conformational residues with any model χ_1 (green dots and solid line), χ_1 PDB non-rotameric residues (blue dots and solid lines), and PDB multi-conformational residues having X_γ atoms at least 60° apart (red dots and solid lines).

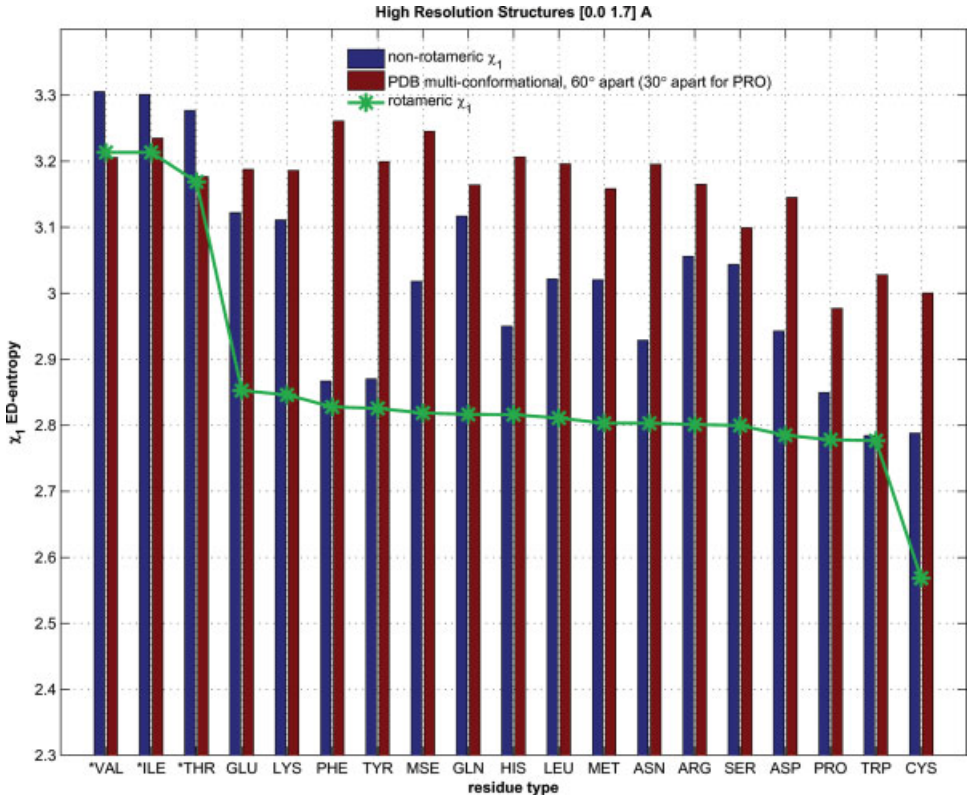

Fig. 9. Mean χ_1 rotamer entropy. The entropy data are given for three categories: (1) non-rotameric χ_1 side chains (blue bars); (2) PDB multi-conformational side chains with alternative X_γ atoms at least 60° apart (red bars); and (3) rotameric χ_1 side chains (green line and stars). The non-rotameric and multi-conformational side chains express significantly higher χ_1 entropy, and therefore, are more disordered. * Valine, isoleucine and threonine with two X_γ atoms produce two X_γ ED peaks on χ_1-rotation plots even in the single-conformation case; as a consequence, they have higher values of χ_1 rotamer entropy when calculated with the formula used for the single X_γ atoms. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

TABLE III. χ_1 Statistics of PDB Entries in the (0, 1.7] Å Range (Dataset 3)[a]

	PDB statistics					
	Single			Multi		
Residue type	Rotam χ_1	Nonrot χ_1	Interm χ_1	≥60°	others	Total no.
LEU	94	1.9	2.6	0.4	1.2	38,308
SER	90	1.0	1.7	5.0	2.0	25,913
ASP	93	1.6	3.5	0.8	1.1	26,112
ASN	92	1.8	4.3	1.0	1.3	19,240
LYS	92	2.2	2.5	0.5	3.1	24,094
GLU	91	2.3	2.7	1.4	2.4	27,172
GLN	93	1.7	2.0	1.0	2.5	15,855
ARG	92	1.7	2.4	0.5	3.3	20,736
HIS	94	0.8	3.1	0.4	1.3	10,323
PHE	95	1.0	3.3	0.1	0.6	17,363
CYS	93	0.8	1.7	2.1	2.5	5,903
TRP	95	0.9	3.0	0.1	0.7	6,624
TYR	95	0.8	3.4	0.1	0.8	15,623
MET	88	1.9	2.1	1.2	6.5	8,360
ALL	93	1.6	2.8	1.1	2.0	261,622

[a]Amino acid residues are split into two groups: (1) single-conformational and (2) multi-conformational. Single-conformational residues (1) are the residues having all atoms in single positions in their PDB entries. Multi-conformational residues (2) are the residues having at least one atom with alternative coordinates declared in the coordinate section of PDB entries. The single-conformational group (1) is subdivided in accordance with χ_1 torsion angle values: 1(a) rotameric χ_1, 1(b) nonrotameric χ_1, and 1(c) intermediate χ_1. The multi-conformational group (2) is subdivided only in two subgroups: 2(a) residues having alternative positions of the X_γ atom with at least 60° χ_1 difference (≥60°), 2(b) the remaining multi-conformational residues (others) consisting of those with χ_1 difference less than 60°, or side chains having multi-conformations at X_δ or beyond, and those with multiple C_α.positions. The last column represents the total number of residues of each residue type. Thr, Ile, Val, Pro, Ala, and Gly are omitted.

TABLE IV. Electron Density and Entropy Analysis of PDB χ_1 Nonrotameric Single-Conformational Side Chains[a]

PDB	Single-conformational									60° multi-conformational	
	Rotameric χ_1			Nonrotameric χ_1			Intermediate χ_1				
ED		Rotam χ_1			Nonrot χ_1			Nonrot χ_1			
	Nonrot χ_1	$>\sigma$	$\leq\sigma$	Rotam χ_1	$>\sigma$	$\leq\sigma$	Rotam χ_1	$>\sigma$	$\leq\sigma$	$>\sigma$	$\leq\sigma$
LEU	1	11	88	21	53	26	56	22	22	79	21
SER	0	10	89	11	61	28	64	16	20	73	27
ASP	1	14	86	8	44	48	63	10	27	76	24
ASN	1	13	86	6	40	54	55	9	36	79	21
LYS	1	11	87	19	51	30	56	25	19	72	28
GLU	2	12	86	19	50	31	56	25	19	70	30
GLN	1	12	87	18	55	27	58	20	22	74	26
ARG	1	12	87	20	51	28	61	19	20	71	29
HIS	1	10	89	5	35	60	61	4	34	81	19
PHE	0	10	90	3	17	80	67	4	29	87	13
CYS	0	12	88	11	41	48	61	14	25	79	21
TRP	0	11	89	7	12	81	68	2	30	67	33
TYR	1	10	90	2	20	77	64	4	32	85	15
MET	1	11	88	10	51	39	49	27	24	80	20
ALL	1	11	88	15	47	38	60	15	25	74	26

[a]Side chains are classified according to their PDB rotamer status and that predicted by their electron density. Single-conformational side chains are divided into three groups: rotameric, nonrotameric, and intermediate based on their χ_1 torsion angle calculated from the PDB coordinates (top line). These are further broken down according to the rotamer/nonrotamer status and entropy as calculated from the point electron density (second line). From PED-analysis, it is found that 15% of all nonrotameric side-chains are more consistent with rotameric conformations. The other 85% are subject to χ_1 entropy analysis. It shows that 47% of all jointly PDB nonrotameric side chains are also ED-nonrotameric and significantly disordered ($S \geq \overline{S} + \sigma$). The remaining PDB-nonrotameric side chains (38%) are also ED nonrotameric but more ordered $S < \overline{S} + \sigma$. To demonstrate pertinence of the ED and entropy analysis, similar results are presented for the PDB-rotameric side chains. It shows that only 1% is more consistent with nonrotameric χ_1 according to the ED, while 88% are ordered rotameric single-conformational side chains. The remaining 11% have high entropy ($S \geq \overline{S} + \sigma$). The PDB multi-conformational data confirms the usefulness of the entropy for the disorder analysis: 74% of multi-conformational side chains in the PDB having alternative X_γ atoms at least 60° apart demonstrate high entropy. Thr, Ile, Val, Pro, Ala, and Gly are omitted.

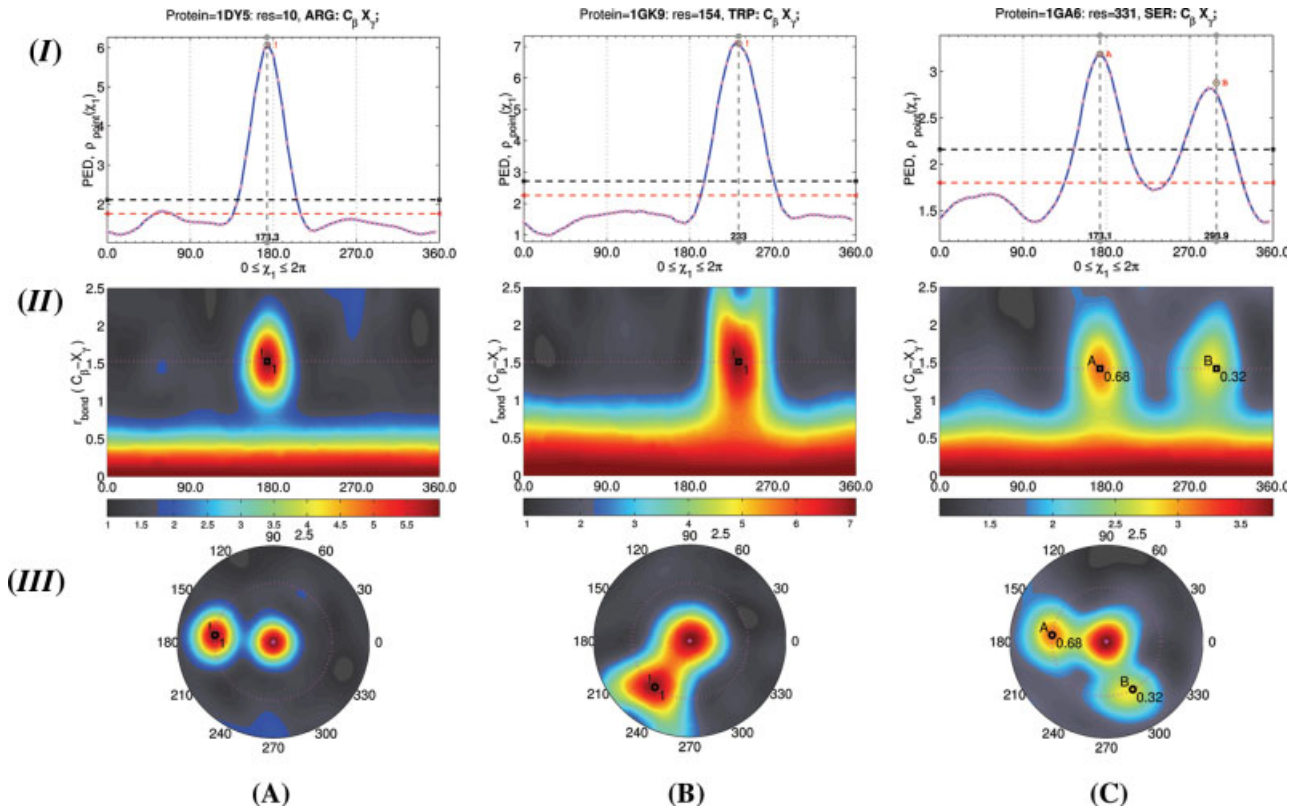

Fig. 10. Three different types of χ_1 ED distribution of side chains likely to be correctly modeled in the structure deposited in the PDB: (**A**) a single-conformational rotameric arginine; (**B**) a single-conformational non-rotameric tryptophan; (**C**) a multi-conformational rotameric serine. χ_1-rotations analysis without altering the position of the C_β atom or changing the $C_\beta-X_\gamma$ bond length and $C_\alpha-C_\beta-X_\gamma$ angle. The top plot (*I*) is $\rho_{point} = \rho_{point}(\chi_1)$, while the middle and bottom plots (*II*) and (*III*) are $\rho_{point} = \rho_{point}(\chi_1, r_{\beta\gamma})$ in the rectangular and polar coordinate systems respectively. On the plots in (*II*) and (*III*), the color scheme represents the level of ED ρ_{point} in the (χ_1, $r_{\beta\gamma}$) space. The vertical black dash lines (*I*), black squares (*II*), and black circles (*III*) indicate the atomic model χ_1 conformations. (*I*) The symbols A, B (multi-conformational) and ! (single-conformational) characterize the conformations according to the PDB entry. (*II*, *III*) The black numbers close to the black squares and circles represent occupancies of the alternative conformations. The purple horizontal line (*II*) and dashed circle (*III*) show the PDB bond length. The red dashed line (*I*) demonstrates the average crystal unit cell ED, the black dashed line the 15% excess over it. The shadowed black-and-white colors (*II*, *III*) describe the area having ED below the average unit cell ED, The pale colors surround the area 15% above the average unit cell ED. The strong central peak corresponds to the C_β atom, the peaks shifted from the center and at the $r_{\beta\gamma} = C_\beta - X_\gamma$ bond distance the X_γ atom(s).

rotameric peaks. The modeled position is placed between two rotameric positions trans and g^-. Both cases (C) and (D) belong to the 15% group of PDB non-rotameric side chains more consistent with rotametic conformations (either single (C) or multiple (D)).

Other PDB non-rotameric side chains do not have strictly rotameric ED but have relatively high entropies. Two cases are shown in Figure 11(E,F). The side chain in Figure 11(E) has an annotated non-rotameric peak and not declared g^+ rotameric strong peak. In fact, the density at the PDB dihedral angle may belong to electron density from other nearby atoms of the same side chain or other residues, as shown in the two-dimensional plots. Figure 11(F) shows that a PDB non-rotameric lysine that has very broadly distributed electron density between 160° and 300°, with a maximum at about 210°. χ_1^{PDB} is 254°, approximately in the center of the well-dispersed electron density. This side chain is very likely to be moving back and forth between two rotameric positions t and g^-, at least before flash cooling.

Finally in Figure 11(G), we show a PDB-rotameric side chain for which the ED shows a non-rotameric distribution. This occurs for only 1% of rotameric side chains. However, in this case it is noticeable that the C_β atom has lower density than the C_γ and this may indicate a problem with the modeling of the backbone and C_β atom positions rather than the side chain.

In summary, we have observed many residues with multiple peaks in the electron density that should very likely be modeled with multiple positions for the X_γ atoms. Examples are shown in 11(A–E). And there are cases such as that in 11(F) where the density is spread out and overlaps two rotameric positions. These may also be better modeled as two separate approximately rotameric conformations. Most likely such misinterpretations arise either because the software used is not designed to refine multiple positions and occupancies or

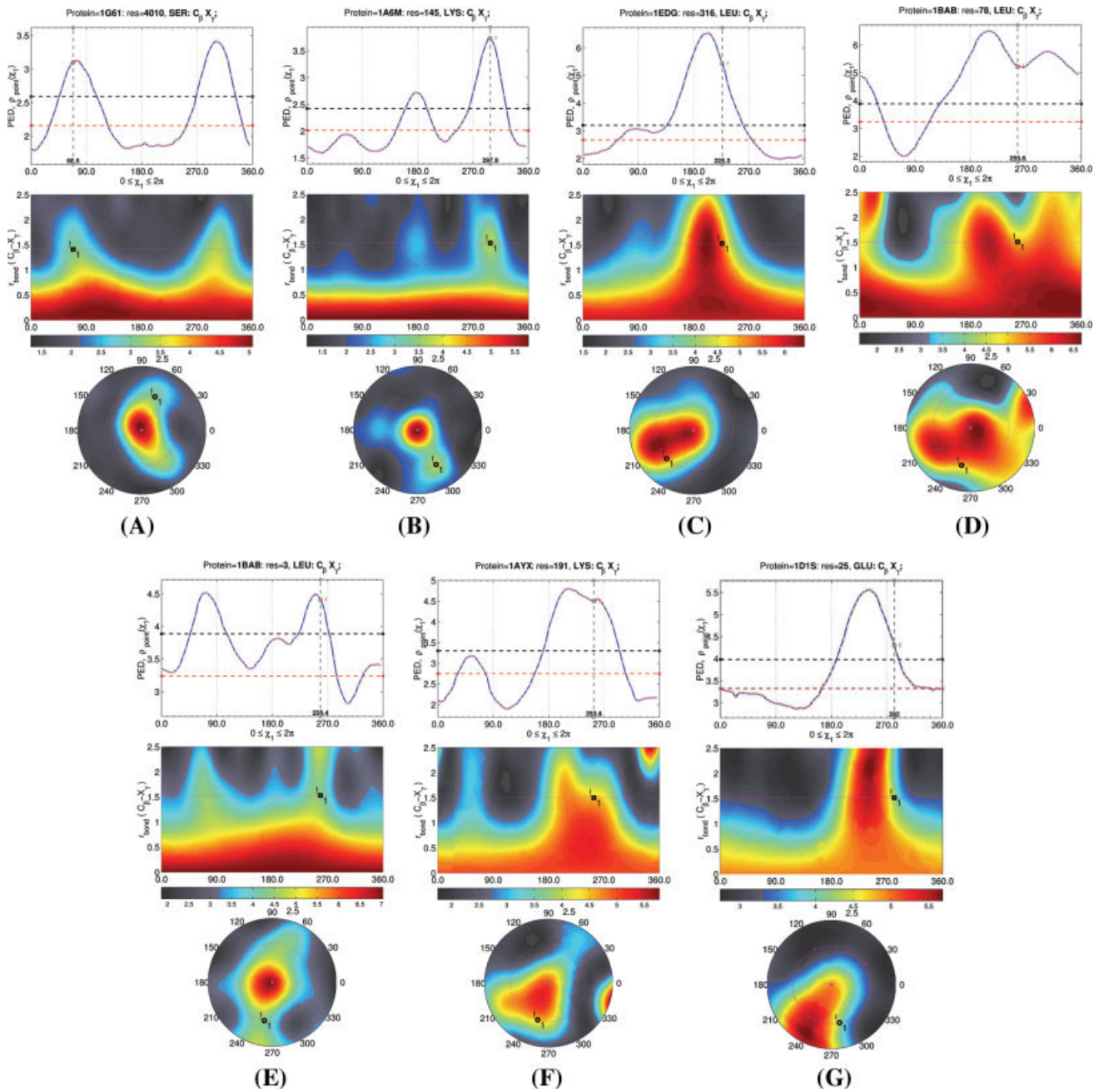

Fig. 11. The ED distribution of incorrectly or questionably modeled single-conformational, non-rotameric and rotameric side chains. (**A**,**B**) Serine and lysine: PDB single-conformational side chains have ED distribution patterns very similar to the PDB declared multi-conformational residues [compare with Fig. 10(C)]. They should probably be modeled as multi-conformational residues. (**C**) Leucine: PDB non-rotameric side chain is more consistent with a single-conformational rotamer. (**D**) Leucine: PDB non-rotameric side chain is placed between two rotameric ED peaks and should be modeled as a multi-conformational rotamer. (**E**) Leucine: PDB not-rotameric single-conformational side chain has another ED peak in the rotameric area and should be annotated as multi-conformational. (**F**) Lysine: PDB non-rotameric side chain with a very broad ED peak covering two rotameric positions and would be also better modeled as a multi-conformational side chain. (**G**) Glutamic acid. Inaccurate PDB single-conformational rotameric residue refinable to a non-rotameric conformation, although the C_β density is weak. For details and notations see the legend of Figure 10.

the density is too weak and unresolved to place two sets of coordinates easily. In addition, if the rest of the side chain is not visible (X_δ atoms etc.), crystallographers may be reluctant to place X_γ atoms. This is probably why there are fewer residues with discordant ED/PDB positions among the shorter side chains (Ser and Cys) compared with the longer ones (Lys, Arg, Glu, Gln etc.).

We are in the process of developing methods of automatic detection of multi-conformational side chains, that

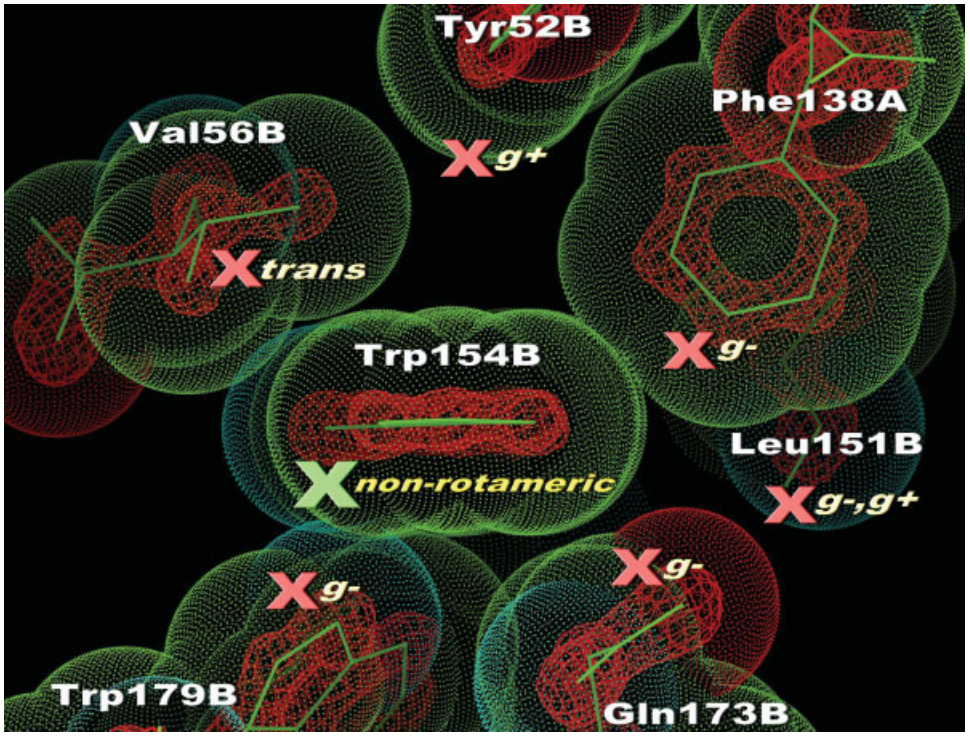

Fig. 12. Trp 154B taken from PDB entry 1GK9 (resolution 1.3 Å) demonstrates a non-rotameric χ_1. It is held there by a large number of neighboring interactions. At more rotameric positions, it strongly clashes with the neighboring side chains: at *trans*: with Val56B, at g^-: with Phe138A, Gln173B, Trp179B and at g^+: with Tyr52B, Leu151B. Thus the side chain cannot occupy any of the standard rotamers: g^+, *trans*, and g^-. The PDB model atoms of the non-rotameric Trp154B and its neighboring residues are precisely fitted into the high-dense electron clouds represented by the red 4σ contour lines. The dashed colored surfaces depict van der Waals radii of the atoms, and demonstrate the very tight environment. The green cross designates the only allowed χ_1 non-rotameric conformation; the three restricted rotamers and the residues they would clash with are indicated with the red crosses. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

are robust and produce as few false positives or false negatives as possible. The details of those rules and algorithms will be the subject of a future paper.

Correct Ordered Non-Rotameric Residues

We have shown that non-rotameric side chains are rare instances of the PDB model (1.6%, Table III) at high resolution. Of the PDB-non-rotameric residues, only 38% (Table IV) are relatively ordered in their χ_1 ED. Among those 38% there are relatively few residues having χ_1 in the center of the non-rotameric regions (near the fully eclipsed positions at 0°, 120°, 240°). The majority tend to stay closer to the intermediate area. We were interested in how such ordered non-rotameric conformations arise and what their trends are.

We show one representative example in Figure 12, Trp 154B, taken from PDB entry 1GK9 (resolution 1.3Å),[42] which demonstrates a non-rotameric χ_1. It is held there by a large number of neighboring interactions. If placed at rotameric positions, it would strongly clash with neighboring side chains: at trans with Val56B, at g^- with Phe138A, Gln173B, and Trp179B and at g^+ with Tyr52B and Leu151B. Thus the side chain cannot occupy any of the staggered rotamers: g^+, trans and g^-. Its χ_1 ED distribution pattern [shown in Fig. 10(B)] demonstrates a very narrow non-rotameric peak. Such a narrow peak is very uncommon for non-rotameric side chains because in those positions they clash with backbone atoms: either Hα at 240° between trans and g^-, or backbone N at 0° between g^- and g^+, or backbone C at 120° between g^+ and trans. In the case of Trp154B it happens because any slight change of χ_1 torsion angle towards trans or g^- leads to strong clashing with Val56B or Trp179B respectively. In other words Trp154B is held in a very tight environment. It is notable that in other structures of the same protein, this residue has very similar χ_1 values to the one in PDB entry 1GK9 (data not shown). Such strained conformations may have functional roles.[46,47]

Based on Figures 6(A,B) and 8(A2,B2,C2) we may conclude that for most side-chain types, 0° non-rotameric conformations occur more rarely than 120° non-rotameric conformations, and the 120° non-rotameric conformations happen more rarely than conformations near 240°. This is expected, since the eclipsed dihedral of the side-chain X_γ heavy atom occurs with a hydrogen atom at 240° (H_α), while for the other non-rotameric positions,

0° and 120°, the heavy atom X_γ is eclipsed with heavy atoms N and C of the backbone respectively.

DISCUSSION

It is tempting when using structures from the Protein Data Bank to treat all atom positions as "true," rather than as a model that explains the observed structure factors. Since the advent of crystallographic software that contains energy functions, such as CNS,[20] it is not always possible to tell when atoms have been placed in real electron density or placed in part because of a strong energy function component (least-squares and maximum likelihood methods also use stereochemical restraints, but with different weights and units). These programs of course have been of tremendous benefit since the density can often be fit in more than one way, and the correct way is likely to be the one of lower internal energy. Especially in moderate to low-resolution structures it is always possible to assign the wrong atom labels to observed density, or in fact to assign atom coordinates with high B-factors and almost no density at all.

The atomic B-factors are useful in identifying atoms with low electron density, potential disorder, and uncertainty in the coordinate position. They have proved very useful in determining which side chains used to build a rotamer library might best be discarded in order to provide the highest quality data for statistical analysis.[17] However, they do not provide an understanding of why a side chain may be placed in an unfavorable rotameric position, or what the origin of the low electron density (and hence high B-factor) might be.

To achieve a better understanding of side chains either with high B-factors or non-rotameric dihedral angles (or both), we have undertaken a statistical analysis of the electron density properties of protein side chains in three large data sets. For non-rotameric side chains, as defined by the model provided in the PDB, we have found that about 15% of these side chains actually have density more consistent with one or more rotameric positions. A further 47% have high entropy as a function of χ_1 due to density that is spread out over more than one rotameric region. These side chains are likely to be moving back and forth between different positions, even though the density may not be resolved clearly enough to identify two (or more) rotameric positions. The remaining 38% of non-rotameric side chains have fairly well-resolved density at their model positions, and may in fact be true "non-rotameric" side chains. Examination of some of these demonstrates that they are held there by a large number of neighboring interactions with other side chains. Some may in fact also be due to errors in backbone modeling.

Even among side chains that are rotameric in the PDB models, we find significant numbers with high entropy. Upon looking at the electron density as a function of χ_1, we observe for many of these side chains clear evidence of two or more rotameric positions for the γ heavy atom that is not annotated in the crystallographic model in the PDB. We are currently developing methods for clearly identifying these types of side chains, and determining whether modeling a larger proportion of side chains as multi-conformational will have a beneficial effect on X-ray crystallographic refinement, as demonstrated by improved residue-based real-space *R* values, and by tracking *R* and free-*R* factor values in order not to overfit the diffraction data.

Most side-chain conformation prediction programs do not predict multiple positions for side chains, and their existence certainly has an effect on the accuracy of such programs when applied to test-sets of known structures. Indeed, we found that for the side chains in the top quintile of electron density (thus side chains with single conformations nearly all with rotameric χ_1), our program SCWRL[48] has a prediction accuracy of 88% correct χ_1 within 40°, while in the bottom quintile of density (multi-conformational and/or non-rotameric side chains) the accuracy is 67%. A further analysis of these results will be presented elsewhere. In any case, the existence of a large number of side-chains in multiple rotameric positions is a challenge for further research in structure prediction as well as structure determination.

REFERENCES

1. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
2. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 2004; 13:400–411.
3. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34 (database issue):D291–D295.
4. Dunbrack RL, Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440.
5. Sasisekharan V, Ponnuswamy PK. Backbone and sidechain conformations of amino acids and amino acid residues in peptides. Biopolymers 1970;9:1249–1256.
6. Sasisekharan V, Ponnuswamy PK. Studies on the conformation of amino acids. X. Conformations of norvalyl, leucyl, aromatic side groups in a dipeptide unit. Biopolymers 1971;10: 583–592.
7. Janin J, Wodak S, Levitt M, Maigret B. Conformations of amino acid side-chains in proteins. J Mol Biol 1978;125:357–386.
8. Bhat TN, Sasisekharan V, Vijayan M. An analysis of sidechain conformation in proteins. Int J Peptide Protein Res 1979;13: 170–184.
9. Benedetti E, Morelli G, Nemethy G, Scheraga HA. Statistical and energetic analysis of sidechain conformations in oligopeptides. Int J Peptide Protein Res 1983;22:1–15.
10. Ponder JW, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–792.
11. McGregor MJ, Islam SA, Sternberg MJE. Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. J Mol Biol 1987;198:295–310.
12. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
13. Dunbrack RL, Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230:543–574.

14. Schrauber H, Eisenhaber F, Argos P. Rotamers: to be or not to be? An analysis of amino acid sidechain conformations in globular proteins. J Mol Biol 1993;230:592–612.
15. Dunbrack RL, Jr, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1:334–340.
16. Dunbrack RL, Jr, Cohen FE. Bayesian statistical analysis of protein sidechain rotamer preferences. Protein Sci 1997;6:1661–1681.
17. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
18. Petrella RJ, Karplus M. The energetics of off-rotamer protein side-chain conformations. J Mol Biol 2001;312:1161–1175.
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
20. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54 (Part 5):905–921.
21. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D Biol Crystallogr 1999;55 (Part 1):191–205.
22. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47 (Part 2):110–119.
23. Chapman MS. Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron density function. Acta Crystallogr A 1995;51:69–80.
24. Jones TA, Kjeldgaard M. Electron density map interpretation. Methods Enzymol 1997;277:173–208.
25. Kleywegt GJ. Experimental assessment of differences between related protein crystal structures. Acta Crystallogr D Biol Crystallogr 1999;55 (Part 11):1878–1884.
26. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A 1991; 47:392–400.
27. Davis IW, Arendall WB, III, Richardson DC, Richardson JS. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure 2006;14:265–274.
28. Rhodes G. Crystallography made crystal clear: a guide for users of macromolecular models. San Diego, CA: Academic Press; 2000.
29. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33 (web server issue):W94–W98.
30. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. The Uppsala electron-density server. Acta Crystallogr D Biol Crystallogr 2004;60 (Part 12, No. 1):2240–2249.
31. Wlodawer A, Li M, Dauter Z, Gustchina A, Uchida K, Oyama H, Dunn BM, Oda K. Carboxyl proteinase from *Pseudomonas* defines a novel family of subtilisin-like enzymes. Nat Struct Biol 2001;8:442–446.
32. Park YJ, Luger K. The structure of nucleosome assembly protein 1. Proc Natl Acad Sci USA 2006;103:1248–1253.
33. Tonrud DE. Knowledge-based B-factor restraints for the refinement of proteins. J Appl Crystallogr 1996;29:100–104.
34. Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Eng 1997;10:777–787.
35. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003; 12:1060–1072.
36. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72:212, 213.
37. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
38. Carugo O, Argos P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. Proteins 1998;31:201–213.
39. Karplus M, Parr RG. An approach to the internal rotation problem. J Chem Phys 1963;38:1547–1552.
40. MacArthur MW, Thornton JM. Protein side-chain conformation: a systematic variation of chi 1 mean values with resolution—a consequence of multiple rotameric states? Acta Crystallogr D Biol Crystallogr 1999;55:994–1004.
41. Esposito L, Vitagliano L, Sica F, Sorrentino G, Zagari A, Mazzarella L. The ultrahigh resolution crystal structure of ribonuclease A containing an isoaspartyl residue: hydration and stereochemical analysis. J Mol Biol 2000;297:713–732.
42. McVey CE, Walsh MA, Dodson GG, Wilson KS, Brannigan JA. Crystal structures of penicillin acylase enzyme-substrate complexes: structural insights into the catalytic mechanism. J Mol Biol 2001;313:139–150.
43. Sierk ML, Zhao Q, Rastinejad F. DNA deformability as a recognition feature in the reverb response element. Biochemistry 2001;40:12833–12843.
44. Groft CM, Beckmann R, Sali A, Burley SK. Crystal structures of ribosome anti-association factor IF6. Nat Struct Biol 2000; 7:1156–1164.
45. Vojtechovsky J, Chu K, Berendzen J, Sweet RM, Schlichting I. Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophys J 1999;77:2153–2174.
46. Herzberg O, Moult J. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins 1991;11:223–229.
47. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 1996;5: 1406–1420.
48. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.

APPENDIX A: ELECTRON DENSITY (PED/IED) VERSUS DEBYE–WALLER TEMPERATURE B-FACTOR RELATIONSHIP DEVIATION

The B-factor can be related to the mean displacement of a vibrating atom $\langle u \rangle$ by the Debye–Waller equation: $B = 8 \cdot \pi^2 \cdot \langle u \rangle^2$. The ED interpolated to the center of an atom $\rho_{\text{point}}(\vec{r}_{\text{atom}})$, e/Å^3 is approximately proportional to the number of the atom electrons divided by the effective volume that the atom ED cloud occupies: $\rho_{\text{point}}(\vec{r}_{\text{atom}}) \propto \frac{Q_{\text{atom}}}{V_{\text{eff}}} = \frac{Q_{\text{atom}}}{4/3 \cdot \pi \cdot R_{\text{eff}}^3}$.

When the atom is at the absolute zero temperature, it does not vibrate ($\langle u \rangle = 0$ Å, $B = 0$ Å^2) and has its effective radius equal to the covalent atom radius: $R_{\text{eff}} = r_{\text{coval}}$. When the atom does vibrate ($u > 0$ Å, $B > 0$ Å^2), its effective radius is increased by the mean displacement: $R_{\text{eff}} = r_{\text{coval}} + \langle u \rangle$. After the substitution $\langle u \rangle = \sqrt{B/(8 \cdot \pi^2)}$ into $\rho_{\text{point}}(\vec{r}_{\text{atom}})$ we have $\rho_{\text{point}}(\vec{r}_{\text{atom}}) \propto \frac{Q_{\text{atom}}}{4/3 \cdot \pi \cdot (r_{\text{coval}} + \sqrt{B/(8 \cdot \pi^2)})^3}$, or $\rho_{\text{point}} \propto \left[r_{\text{coval}} + \left(\frac{B}{8 \cdot \pi^2} \right)^{\frac{1}{2}} \right]^{-3}$.

Since $\rho_{\text{integ}}(\vec{r}_{\text{atom}})$ is the average atom electron density calculated based on the theoretical probability density function of electron positions, the same expression approximately describes the relationship between ρ_{integ} and B: $\rho_{\text{integ}} \propto \left[r_{\text{coval}} + \left(\frac{B}{8 \cdot \pi^2} \right)^{\frac{1}{2}} \right]^{-3}$.

APPENDIX B: ESTIMATION OF SAMPLE SIZE PROPERLY EVALUATING g^+, t, g^- ROTAMER POPULATIONS

For 17 out of 20 amino acid residues (all except glycine, alanine and proline) three χ_1 rotamers occur: g^+, t, g^-. Their frequencies are specific for each residue type and also depend on how a sample is chosen. As long as g^+, t, g^- rotamers do not have overlapping χ_1 intervals, the χ_1

rotamer events are independent ($g^- \cap trans = 0$, $trans \cap g^+ = 0$, $g^- \cap g^+ = 0$) and the sum of their probabilities gives 1: $P(g^-) + P(trans) + P(g^+) = 1$. For a multi-conformational rotamer the events are still independent but their frequencies have to be quantified appropriately (according to the occupancies of the alternative conformations).

We make the designations: $P(g^-) = \theta_1$, $P(trans) = \theta_2$ and $P(g^+) = \theta_3 = 1 - (\theta_1 + \theta_2)$. If N is the sample size, k is the g^- rotamer number, l is the *trans* rotamer number, and $m = N - (k + l)$ is the g^+ rotamer number then the probability of such distribution into the rotamer wells is

$$P(k,l,N;\theta_1,\theta_2) = \frac{N!}{k!l!(N-k-l)!} \cdot \theta_1^k \cdot \theta_2^l \cdot (1-\theta_1-\theta_2)^{(N-k-l)}$$

The sample frequencies k/N, l/N, and m/N give the estimates of θ_1, θ_2, and θ_3 respectively. It is clear that those frequencies differ from the real probabilities and carry a statistical error varying from a sample to a sample.

We would like to determine by how much the sample frequencies may differ from their real probabilities, so that they do not differ by more than some relative error $\delta\theta$; that is, so that $k/N \in \theta_1 \cdot [1 - \delta\theta, 1 + \delta\theta]$, $l/N \in \theta_2 \cdot [1 - \delta\theta, 1 + \delta\theta]$ and $m/N \in \theta_3 \cdot [1 - \delta\theta, 1 + \delta\theta]$. The sum of the probabilities of the (k_i, l_i, m_i) sets having frequencies in those relative error intervals is:

$$F(k,l,N;\theta_1,\theta_2,\delta\theta) = \sum_i P\Bigg(k_i,l_i,N;\theta_1,\theta_2 | k_i + l_i \leq N : \left|\frac{k_i}{N} - \theta_1\right| \leq \delta\theta, \left|\frac{l_i}{N} - \theta_2\right| \leq \delta\theta\Bigg)$$

Since we do not know the real θ_1 and θ_2, they can be approximated as k/N and l/N respectively.

For every ED range bin we required a minimum sample size N that guaranteed the 80% confidence that the frequencies represent the real probabilities with no more than the 5% relative error.