Structure, Vol. 12, 831–838, May, 2004, ©2004 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.str.2004.02.031

Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography

Mark A. DePristo,* Paul I.W. de Bakker,[1]
and Tom L. Blundell
Department of Biochemistry
University of Cambridge
80 Tennis Court Road
Cambridge, CB2 1GA
United Kingdom

Summary

Proteins are dynamic molecules, exhibiting structural heterogeneity in the form of anisotropic motion and discrete conformational substates, often of functional importance. In protein structure determination by X-ray crystallography, the observed diffraction pattern results from the scattering of X-rays by an ensemble of heterogeneous molecules, ordered and oriented by packing in a crystal lattice. The majority of proteins diffract to resolutions where heterogeneity is difficult to identify and model, and are therefore approximated by a single, average conformation with isotropic variance. Here we show that disregarding structural heterogeneity introduces degeneracy into the structure determination process, as many single, isotropic models exist that explain the diffraction data equally well. The large differences among these models imply that the accuracy of crystallographic structures has been widely overestimated. Further, it suggests that analyses that depend on small differences in the relative positions of atoms may be flawed.

Introduction

Proteins are dynamic, heterogeneous molecules (Frauenfelder et al., 1991; McCammon and Harvey, 1987). They exhibit individual atomic anisotropic motion and collective, large-scale motion over a range of time scales (Frauenfelder et al., 1991). Their complex energy landscapes give rise to multiple, significantly populated conformations that are separated by large energy barriers (Burling et al., 1996; Rejto and Freer, 1996; Ringe and Petsko, 1986; Smith et al., 1986; Stec et al., 1995). Further, dynamics and heterogeneity have been increasingly recognized as essential for protein function (McCammon, 1999; Rader and Agard, 1997; Rejto and Freer, 1996; Wilson and Brunger, 2000).

Dynamics and heterogeneity remain even in the crystalline form, due to freedom afforded by the high solvent content in most protein crystals (Jensen, 1997; Ringe and Petsko, 1986). Crystalline proteins retain the ability to bind ligands reversibly, even when the ligand is large and the binding site is buried in the solvent-inaccessible core (Petsko, 1996). Atomic resolution crystal structures exhibit extensive, discrete conformational substates, where up to 30% of side chains can exist in multiple conformations (Rejto and Freer, 1996; Smith et al., 1986; Stec et al., 1995).

Modeling anisotropic motion and structural heterogeneity has been limited to proteins that diffract to atomic resolution, due to the necessity for a high parameter-to-observation ratio (Ringe and Petsko, 1986; Wilson and Brunger, 2000). The vast majority of proteins (>90%) diffract to worse than 1.6 Å resolution and are solved as a single, average conformation with Gaussian, isotropic thermal motion (Burling et al., 1996; Kuriyan et al., 1986; Ringe and Petsko, 1986; Wilson and Brunger, 2000).

The artifacts that are introduced by ignoring heterogeneity during structure determination remain largely uncharacterized. Kuriyan et al. (1986) refined a single, isotropic B factor structure against reflection data derived from a molecular dynamics simulation trajectory, concluding that many atoms had substantially incorrect mean positions and thermal motion distributions. Following a similar protocol, Vitkup et al. (2002) showed that inadequate modeling of anisotropy and structural heterogeneity is the principal limitation in accounting for all of the simulated reflection data. Consequently, ignoring structural heterogeneity leads to (i) an incomplete description of the crystallographic data (Vitkup et al., 2002) and (ii) a considerable degree of inaccuracy stemming from the inability to fit a single, average conformation to diffraction data generated by a dynamic, heterogeneous ensemble of molecules (Kuriyan et al., 1986; Rejto and Freer, 1996; Ringe and Petsko, 1986).

The introduction of uncharacterized inaccuracies in crystal structures is troubling, as estimates of the uncertainty in atomic positions are necessary to identify genuine features or differences among structures (Kleywegt, 1999). Without such estimates, overinterpretation of unreliable conformations is inevitable. Among the most famous examples is the steric hindrance hypothesis for carbon monoxide binding to myoglobin (Stec and Phillips, 2001). Based on a bent Fe-C-O angle in low-resolution crystal structures, this hypothesis flourished despite objections from chemists and spectroscopists, and was only recently abandoned as atomic resolution structures were shown to exhibit the expected linear arrangement (Stec and Phillips, 2001).

The accuracy of atomic positions in X-ray crystal structures remains an open and contentious question. Theoretical methods (Luzzati, 1952; Read, 1986; Tickle et al., 1998) estimate positional uncertainties on the order of 0.1–0.3 Å. X-ray restrained molecular dynamics simulations, on the other hand, report larger expected uncertainties of around 0.5 Å (Kuriyan et al., 1987, 1991), concluding that crystal structures are less accurate than anticipated by theoretical calculations. In contrast, comparison of structures solved independently (Ohlendorf, 1994; Zoete et al., 2002), structures with unrestrained noncrystallographic symmetry (Kleywegt, 1996), and from the same reflections by several crystallographers (Mowbray et al., 1999) suggest still larger differences of

*Correspondence: mdepristo@cryst.bioc.cam.ac.uk
[1] Present address: Department of Molecular Biology, Massachusetts General Hospital, 55 Fruit Street, Wellman 8, Boston, Massachusetts 02114.

0.6–1.0 Å. The rather large discrepancy between the theoretical, computational, and experimental estimates suggests that an important effect captured by the experimental techniques is being neglected in the theoretical calculations. We believe that this effect is in part the uncharacterized inaccuracy due to poor modeling of structural heterogeneity.

In this work we address the issue of X-ray crystal structure accuracy in the context of the single, isotropic B factor model. We have derived accuracy estimates by comparing the differences among an ensemble of structures produced by a novel automated refinement procedure. Significantly, the ensemble of conformations exhibits anisotropic atomic motion and discrete conformational substates consistent with theoretical and experimental expectations.

Mimicking manual model building and refinement using experimental reflection sets overcomes many of the limitations of previous approaches. As the same protocol was used to generate each model, the variation observed must result from intrinsic features of the reflection data and limitations of the single, isotropic B factor model, and not from differences in experimental conditions (Ohlendorf, 1994) or subjective human decisions (Mowbray et al., 1999). Our estimates are also free from questionable approximations (Luzzati, 1952), assumptions about the quality and convergence of refinement (Tickle et al., 1998), and potential artifacts due to the introduction of additional refinement parameters (Burling et al., 1996; Kuriyan et al., 1987; Rejto and Freer, 1996; Wilson and Brunger, 2000). Importantly, we use a range of measures of structure fit and quality, not only the R or R_{free} factors as in previous theoretical and computational approaches, which alleviates the concern that the alternate conformations are simply artifacts of insufficient quality control (Kuriyan et al., 1987, 1991). Finally, the discrete nature of our conformational sampling algorithm circumvents the inherent difficulties of crossing energetic barriers that limit and underestimate conformational diversity in conventional molecular mechanics methods (van Gunsteren and Berendsen, 1990).

Results

Resolving Structures

Between 10 and 20 independent conformers were generated with a discrete restraint-based modeling algorithm, called RAPPER, based on propensity-weighted ϕ/ψ and χ angle sampling (de Bakker et al., 2003; DePristo et al., 2003a, 2003b). The PDB structure was used to restrain conformational sampling to only conformations whose Cα coordinates were within 2 Å of the PDB Cαs (DePristo et al., 2003b). Further, all atoms were restrained to lie in positive electron density in a $2F_{obs}$-F_{calc} map phased with the PDB structure, though this restraint could be discarded for side chain atoms if, at a particular residue, the restraint was unsatisfiable.

The Cα-trace and PDB structures are mutually dissimilar, with pairwise 1.2–1.3 Å all-atom root-mean-square deviations (rmsd). The Cα-trace models are also poor models of the reflection data, with R and R_{free} factors of 0.4–0.5. Their fit to the reflection data was improved by iteratively (i) reassigning all side chains to the rotamer with best fit to an electron density omit map and (ii) refining atomic coordinates and B factors against a maximum likelihood residual until convergence. The five refined structures with lowest R_{free} factor were selected for further analysis.

The final five structures have R and R_{free} factors equivalent to or better than the PDB structures (Table 1). Measures of local correctness, such as rms deviation in bond lengths and angles, real-space R factor (Branden and Jones, 1990), fit to $2F_{obs}$-F_{calc} and F_{obs}-F_{calc} electron-density maps, ϕ/ψ compatibility with the Ramachandran plot (Lovell et al., 2003), and side chain rotamericity (Lovell et al., 2000) are also similar to those of the PDB structure (Table 1). B factor distributions are nearly identical among the alternate and PDB structures, with correlation coefficients of 0.90–0.99 over the average residue B factors (Figure 1).

Examination of $2F_{obs}$-F_{calc} simulated annealing (SA) omit maps indicate that the alternate structures are no worse fit than the original PDB structure (Figure 2). There are cases, however, where a local conformation is incorrect, but these constitute a small faction of the observed differences and, importantly, are no more frequent than incorrect fits in the original PDB structure. Finally, to further remove model bias, automated refinement was aborted after the final side chain reassigned step of the five final interleukin-1β models and subjected to 50 steps of restrained refinement with residues 51–55 (Figure 2) removed. The subsequent SA omit maps were virtually identical to those in Figure 2. We conclude that these new structures are equivalent solutions for the contents of the protein crystal. The application of stringent measures of global and local correctness is essential for this conclusion and is a critical improvement over previous approaches (Burling et al., 1996; Kuriyan et al., 1987; Rejto and Freer, 1996; Ringe and Petsko, 1986).

Analysis of Individual Proteins

Amicyanin, a blue-copper cupredoxin involved in electron transport from *Paracoccus denitrificans*, diffracted to 1.31 Å resolution and was solved with isotropic B factors and multiple side chain conformations for 9 of 85 residues using X-ray restrained molecular dynamics/ simulated annealing (MD/SA) refinement (Cunane et al., 1996). The alternate structures exhibit little variance from the PDB structure or among themselves (Figure 3). The main chain conformation is essentially invariant in all models, with only minor differences in the orientation of the carbonyl groups. Although each alternate structure includes only a single side chain conformation, five of the nine side chains modeled as multiple conformations by the authors of the original PDB structure are identifiable by their variability among the ensemble of structures, while the remaining four converged to one of the multiple conformations. An additional five side chains exist in multiple conformations in the ensemble, suggesting an even greater degree of plasticity. Almost all of the side chain variation in the models is attributable to differences at sites of multiple conformations. The sampling and refinement procedure employed here generates ensembles exhibiting heterogeneity similar to

Table 1. Crystallographic Quality Metrics

	Amicyanin		HIV Protease		Interleukin-1β	
	PDB	Models[a]	PDB	Models[a]	PDB	Models[a]
Resolution (Å)	8–1.3	20–1.3	50–1.8		15–2.3	55–2.3
R_{work} (%)						
Published[b]	15.5		19.5		15.7	
Refined[c,d]	15.0	14.3–14.8	19.4	17.5–18.3	16.0	15.7–16.2
R_{free} (%)						
Published[b]			23.0		21.0	
Refined[c,d]		16.8–17.1	22.6	20.9–21.9		20.3–21.7
Real-space R	0.981	0.980	0.964	0.967	0.850	0.846
Rms bond[e] (Å)	0.014	0.012	0.014	0.011	0.018	0.017
Rms angles[e] (°)	2.36	1.48	1.84	1.51	2.36	1.70
Allowed ϕ/ψ (%)	100.0	99.6	100.0	99.8	98.0	98.6
Bad rotamers[f]	0/85	0.2/85	2/166	3.3/166	12/129	12.8/140
Esu from R factor[g] (Å)	0.05		0.15		0.23	

[a] Averaged values over five alternate models.

[b] R_{work} and R_{free} factors of original PDB structure as reported by the authors. May differ from the refined value due to differences in R factor calculation and bulk solvent correction.

[c] Following 20 rounds of restrained refinement on the PDB structure.

[d] Range of R_{work} and R_{free} factors of the five alternate models (models columns).

[e] As calculated by REFMAC; may differ from published values.

[f] The denominators differ for h-IL1β because eleven side chain conformations were not modeled in the PDB structure.

[g] The estimated standard uncertainty based on an approximation to matrix inversion (Murshudov and Dodson, 1997).

that identified by manual model building with the additional value that our method is fully automated.

HIV protease, a major drug target against the human immunodeficiency virus, was crystallized with a cyclic sulfamide inhibitor and solved at 1.8 Å as a single conformation with isotropic B factors using MD/SA refinement (Schaal et al., 2001). Compared to the original free reflection set, the R_{free} values of our models are 1%–3% lower than those of the original or refined PDB structures (Table 1). In our models for HIV protease, the main chain exhibits even less variability than that of amicyanin, due to a higher percentage of buried residues (37% versus

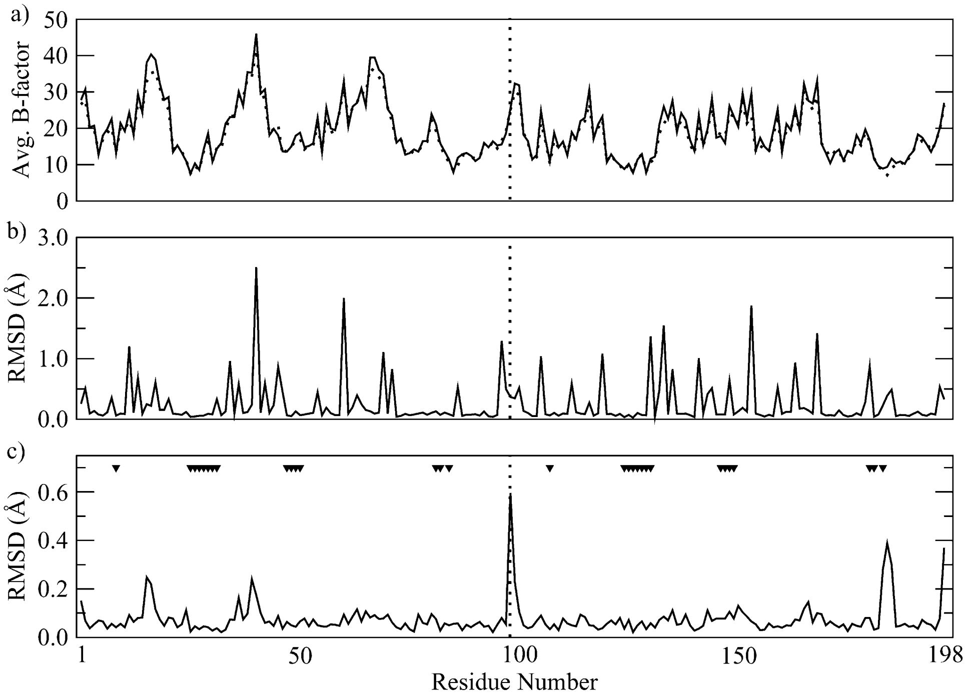

Figure 1. B Factors and Rmsd per Residue for HIV Protease

Averaged B factor (A) of the PDB structure (dots) and the five alternate models (line). Note the similarity of the average B factors between the PDB and RAPPER models. All-atom (B) and main chain (C) rmsd for each residue of the alternate models compared to the PDB structure. Triangles indicate residues in contact with the inhibitor molecule. The vertical dotted line denotes the break between the two chains of the protease dimer.

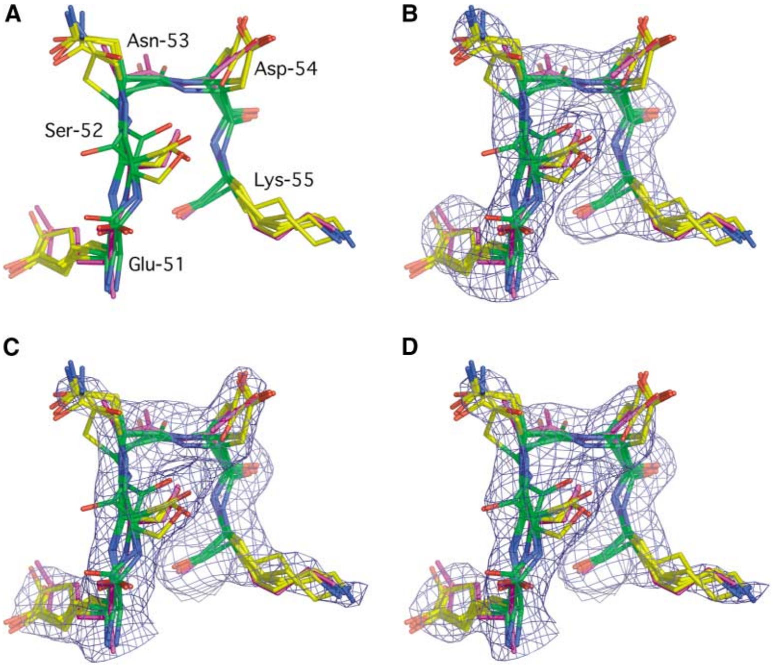

Figure 2. Main Chain and Side Chain Heterogeneity in Human Interleukin-1β

Shown are residues 51–55 from h-IL1β. The PDB structure is in magenta and the five alternate models are colored according to: nitrogen, blue; oxygen, red; main chain carbon, green; side chain carbon, yellow. Note the pronounced backbone variability and side chains with anisotropic motion (Ser52, Asn53, Lys55) and multiple discrete conformations (Glu51, Asp54, Lys55). Asn53 is omitted in the PDB structure. (B)–(D) show simulated-annealing omit maps contoured at 1 σ, for the original PDB structure (B) and alternate models 2 (C) and 3 (D). Note the density differences at Glu51, Asp54, and Lys55. Images were created with PYMOL (DeLano, 2002).

26%) where almost no variability is observed. The side chains, in contrast, are significantly more variable (Figure 3), with 41 of the 166 side chains (25%) clustering into distinct conformations (35) or becoming completely disordered (6). The differences among the models are less than those observed between the protomers of the protease dimer: 0.5 Å main chain and 1.0 Å all-atom rmsd averaged over the PDB and models. Further, the pattern of variation is similar in both protomers, suggesting that it results from an intrinsic property of the protein, consistent with previous observations (Zoete et al., 2002). The region near the inhibitor molecule is highly conserved, as expected from the stabilization of the protease dimer upon ligand binding. Suggestively, the differences observed here are consistent with a recent analysis of 73 crystal structures of HIV protease carried out by Zoete et al. (2002) who found an average all-atom rmsd of 0.5 Å among all structures and 0.1–0.4 Å backbone rmsd between three pairs of independently solved structures.

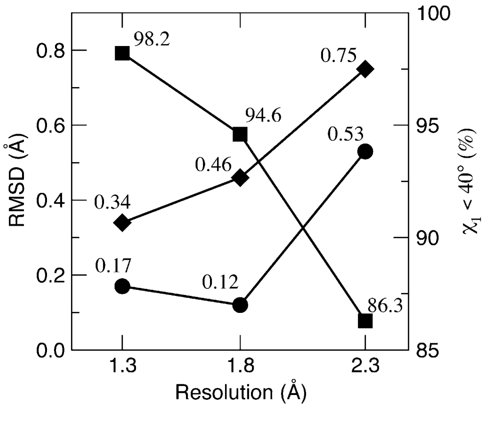

Figure 3. Pairwise Differences among the PDB and Alternate Models by Resolution

Main chain rmsd (circles), all-atom rmsd (diamonds), and rotamer state conservation (squares) versus resolution for amicyanin (1.3 Å), HIV protease (1.8 Å), and h-IL1β (2.3 Å). The χ_1 percentage within 40° measures the fraction of residues with side chains in a similar rotameric state.

Human interleukin-1β (h-IL1β) was solved at 2.32 Å as a single conformation with isotropic B factors using least-squares refinement (Yu et al., 1999). h-IL1β exhibits a surprising degree of both main chain and side chain variability in our models (Figures 2–4). The greatest differences in the main chain are localized to 7 of the 11 surface loops, where models can differ by as much as 1 Å. A total of 43 of the 140 side chains (31%) in h-IL1β occur in multiple conformations (33) or are completely disordered (10), with especially pronounced variability in the regions of large main chain movement. Of the 11 side chains omitted in the PDB model due to poor electron density, 5 are completely disordered, 4 exist in multiple conformations, and 2 appear as single conformations after substantial backbone rearrangements. Our results for h-IL1β are consistent with a comparative analysis of four independent determinations of h-IL1β, with regions of greatest disparity recurring among our models (Ohlendorf, 1994). Thus, it appears that differences in the four h-IL1β crystal structures reflect an inability to fit uniquely a single conformation with isotropic B factors to heterogeneous low-resolution data, and do not result from refinement protocol or human intervention.

General Trends

Several general trends can be drawn from our analyses. Variability increases with distance from the protein core, consistent with the surface-molten solid character of proteins (Zhou et al., 1999). The main chain is more conserved than the side chains, though main chain de-

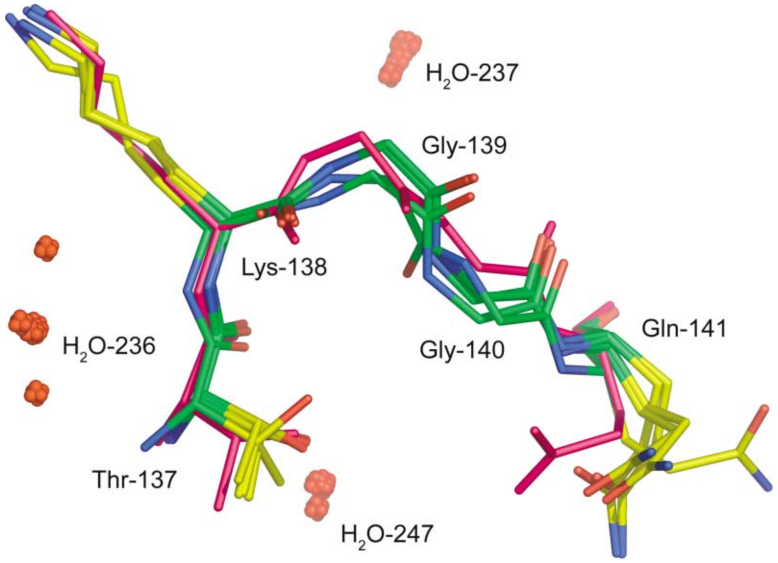

Figure 4. Main Chain and Water Heterogeneity in Human Interleukin-1β

Residues 137–141 from h-IL1β are shown, highlighting backbone variability and disordered side chains and waters. Coloring is the same as in Figure 2. Note the significant variability in the main chain (Gly139 and Gly140) and side chain (Thr137 and Lys138) conformations, while Gln141 appears to be total disordered. Waters H_2O-237 and H_2O-247 are well ordered, whereas H_2O-236 has a mean square displacement of 3.5 Å. Images were created with PYMOL (DeLano, 2002).

generacy can become pronounced in surface loops at low resolutions, especially those containing flexible residues such as glycine (Figure 4). The atomic positions of side chains are quite variable, as reflected by the large all-atom rmsd (Figure 3). Further, gross side chain movement becomes increasingly frequent with decreasing resolution, as shown by the plummeting conservation in rotamer state (Figure 3). High B factors, albeit suggestive of heterogeneity (Stec et al., 1995), are poor predictors of the degree or type of variability. Conversely, low B factors, a commonly used indicator of a reliable conformation, do not in fact ensure freedom from heterogeneity, as multiple side chain conformations frequently occur at such positions. Many crystallographic waters are poorly conserved: in amicyanin and h-IL1β nearly 20% exhibit mean-square displacements >1 Å, although fewer than 5% of waters in HIV protease are so disordered. Such mobile waters are probably artifacts of model bias arising from the different protein models.

Inconsistent Outliers

Amino acids with poor ϕ/ψ angles and nonrotameric side chains vary among the PDB and alternate models for all three proteins. For example, although eleven residues are nonrotameric in at least one model of HIV protease, nine recur only once or twice, and only one, Val-32B, recurs in more than half of the six structures. Most residues with Ramachandran outliers and nonrotameric side chains occur at sites of multiple conformations in the models, suggesting that they result from the inability to fit a single conformation with isotropic B-factors to heterogeneous data, and are not genuinely strained conformations demanded by the experimental data.

Ensemble Is Superior to Individual Structures

The structure factors F_{calc} of the alternate structures are partially independent, with pairwise R factors between 0.2 and 0.3. This suggests that collectively the ensemble of models may account for more of the reflection data than any individual model, a hypothesis that can be tested by calculating an ensemble R factor (see Experimental Procedures). For HIV protease and h-IL1β, the ensemble R_{free} factor is slightly better than any individual model and $\sim$1% lower than the average R_{free} of the models. The ensemble R_{free} factor for amicyanin, however, is worse by $\sim$1%, presumably due to the specificity of the high-resolution data and the simplicity of our ensemble structure factor calculations. Further, the multiple RAPPER models often fit simulated-annealing omit maps better than the single model used to phase the maps (Figure 2). The ensemble of models is a more complete description of the reflection data, for the two lower resolution structures.

Discussion

The greater heterogeneity of protein structures defined at lower resolution is often due to decreasing intermolecular interactions and increased solvent content (Jensen, 1997; Zhang et al., 1995). A single conformation with isotropic B factors is incapable of capturing the anisotropic motion typical of protein crystals and leads directly to large R and R_{free} factors (Kuriyan et al., 1986; Vitkup et al., 2002). The freedom afforded by large R and R_{free} factors allow individual models to converge to different conformational substates. Consequently, variability and inaccuracy increase at lower resolution because (i) the single, isotropic B factor model becomes a progressively worse approximation of the underlying heterogeneity, and (ii) the limited diffraction data reduces the specificity of the atomic coordinates.

An important point to note is the differences in the alternate models arise from correlated changes throughout the molecule. This phenomenon is well illustrated by comparing simulated-annealing omit maps (Figure 2), as differences in these maps result solely from differences in the surrounding protein structure. This behavior is expected given the global relationship between the structure factors and phases and the electron-density map through the Fourier transformation. Nevertheless,

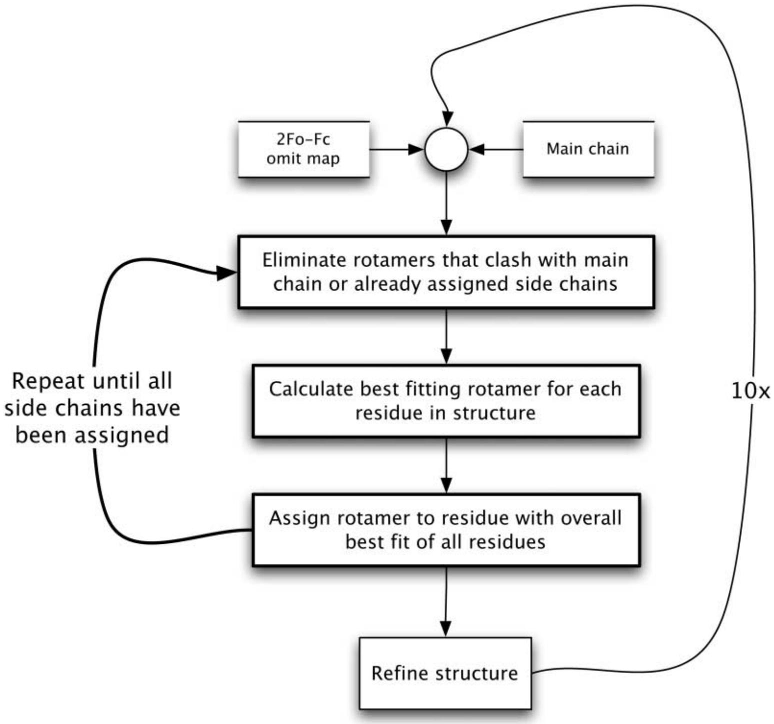
2Fo-Fc omit map
Main chain
Eliminate rotamers that clash with main chain or already assigned side chains
Calculate best fitting rotamer for each residue in structure
Assign rotamer to residue with overall best fit of all residues
Repeat until all side chains have been assigned
Refine structure
10x

Figure 5. Flow Diagram of the Automated Refinement Procedure

Side chain reassignment (think boxes and lines) operates on a protein main chain conformation and an omit map. Each cycle assigns the best-fitting rotamer according to the electron-density map, constructing an all-atom model one side chain at a time. This final model is subjected to a round of refmac refinement (thin boxes and lines), and the whole cycle repeats.

it does explain why the alternate but self-consistent models observed in this work are not routinely encountered by crystallographers: traditional model building and refinement is labor intensive and crystallographers are therefore disinclined to alter the entire molecule at every model-building step and to pursue several solution paths in parallel. Indeed, large differences among alternate models produced by experimentalists have been limited to the rare cases where several groups have crystallized and solved the same protein independently (Ohlendorf, 1994; Zoete et al., 2002).

Our results highlight the importance of structural validation in assessing the quality of a crystal structure. A wide-range of metrics was used to assess the quality of the PDB and alternate models. Indeed, the inability to differentiate systematically among the original and alternate models underlies our assertion that the latter are equivalent solutions for the contents of the crystal cell. We do not suggest that our ensemble of models is complete or correct, but rather, that they represent a range of solutions consistent with current measures of crystal structure quality. A future direction of this research would be to develop more sophisticated structure determination and validation tools that reduce the range of "acceptable" solutions and thereby reduce the degree of degeneracy observed here.

It should be emphasized, however, that in many cases the problem is not selecting the best single conformation, but that several conformations are equally plausible interpretations of the electron-density map, especially at lower resolution (see, for example, Figure 2). This ambiguity is likely to be an unavoidable consequence of the underlying molecular heterogeneity of proteins in the crystalline state, which implies that no validation tool or quality metric could reasonably differentiate among the alternate conformations. This argument suggests that degeneracy can only be overcome, not by restricting the scope of acceptable solutions, but instead by actively incorporating heterogeneity into the model.

The magnitude of the observed variation within the ensemble implies that many structural analyses and comparisons may be flawed. Descriptors such as secondary structure and solvent accessibility, as well as estimates of electrostatic and potential energy, are highly sensitive to the relative positions of atoms, and will therefore be affected by the degeneracy observed here. For example, both the calculated solvent accessibility and secondary structure of h-IL1β differ among the models, with residues changing from buried to exposed and even a short strand dissolving into its surrounding loop. In comparisons of site-directed mutants and ligand complexes with wild-type uncomplexed proteins, it is unsafe to presume that small, observed differences are significant and are caused by the mutation or complexation. Likewise, it is dangerous to rank comparative models (or even modelers) by slight differences from low-resolution experimental structures, which has become common practice in the Critical Assessment of Structure Prediction (CASP) exercises (Tramontano et al., 2001).

In summary, our results highlight the need to develop a better representation of protein heterogeneity in X-ray crystallography. Individual anisotropic vibration and discrete conformational substates cannot be ignored without introducing significant degeneracy and inaccuracy into the structure determination process. It is likely that models comprising multiple backbone and side chain structures selected from our ensembles of conformers will lead to more realistic descriptions of protein struc-

ture and account better for low-resolution reflection data. Our approach represents a significant step toward reconciling experimental data from X-ray crystallography with the large body of theoretical work that emphasizes the dynamic nature of proteins.

Experimental Procedures

Protein Structures and Reflection Data

Atomic structures and structure factors of amicyanin (Cunane et al., 1996) (PDB code 1AAC), HIV protease (Schaal et al., 2001) (1G35), and human interleukin-1β (Yu et al., 1999) (9ILB) were obtained from the PDB. Amicyanin was selected because it of its high resolution; HIV protease because of its well-characterized dynamics and abundant crystal structures (Zoete et al., 2002); and h-IL1β because of the comparative analysis of its four independent, simultaneous solutions (Ohlendorf, 1994). The original free set was obtained for HIV protease. Amicyanin was solved without a free set and the h-IL1β free set was unavailable; new sets were generated including 10% of the reflections. For h-IL1β, this is equivalent to free set generation using a molecular replacement solution. The free set was excluded from all refinement and map calculations.

Initial Cα-Trace Models

Initial Cα-trace models were generated with idealized stereochemistry, favored ϕ/ψ angles, and rotameric side chains, free of heavy-atom, hard-spheres overlap (DePristo et al., 2003b) as part of the RAPPER restraint-based modeling program (de Bakker et al., 2003; DePristo et al., 2003a, 2003b; Shetty et al., 2004). RAPPER employs a discrete build-up algorithm to construct a complete conformation satisfying a set of restraints by iteratively extending a polypeptide chain of valid peptides in the N-terminal to C-terminal direction. The Cα atoms of the initial models were restrained to lie within 2 Å of the PDB Cα atoms. To ensure compatibility with reflection data, all atoms were further restrained to lie in positive density according to a $2F_{obs}$-F_{calc} map phased with the PDB structure. However, the electron density for some side chains is so poor that no residue can be constructed that fits all side chain atoms into positive density. Consequently, RAPPER was permitted to discard the positive density restraint for side chain atoms of residues where the chain extension fails systematically. Finally, B factors were reset to 20 Å^2 for main chain and 30 Å^2 for side chain atoms. Ligands and waters were taken directly from the PDB to ensure that an equivalent numbers of atoms were modeled. Ligand and water atoms were, however, allowed full freedom of movement during refinement. The initial Cα-trace models are available in the Supplemental Materials at http://www.structure.org/cgi/content/full/12/5/831/DC1.

Automated Refinement Protocol

Each model was subjected to one round of unrestrained refinement, two rounds of restrained refinement with loose geometric restraints, and ten cycles of side chain reassignment and refinement with increasingly strict geometric restraints (Figure 5). Finally, the hydrogen bond network was optimized with REDUCE (Word et al., 1999) and subjected to a final step of refinement. See the Supplemental Materials at http://www.structure.org/cgi/content/full/12/5/831/DC1 for a detailed description of the refinement protocol.

Side Chain Reassignment

Side chain reassignment was performed with RAPPER to reposition poorly fit side chain conformations. Given a fixed main chain conformation, new side chain conformations were selected for all residues from the penultimate rotamer library according to their fit to a $2F_{obs}$-F_{calc} omit map calculated with OMIT (CCP4, 1994). The fit was scored by the average atomic electron density σ over all side chain heavy atoms, with a significant weight on negative electron density σs:

$$\frac{1}{N}\sum_{i}^{N} w_i \times \sigma(a_i) \quad w_i = \begin{cases} 1 & \text{if } \sigma(a_i) \geq 0 \\ 10 & \text{if } \sigma(a_i) < 0 \end{cases},$$

where N is the number of side chain heavy atoms, and $\sigma(a_i)$ is the number of deviations from the rms electron density of the i^{th} side chain heavy atom. Further, only rotamers free of heavy-atom, hard-spheres overlaps were considered valid, but with side chain atomic radii reduced by 50% to eliminate only grossly overlapping conformations.

The assignment algorithm proceeds as depicted in Figure 5. First, rotamers that clash with the fixed main chain or previously assigned side chains are eliminated. Next, the valid rotamer with the highest score among all rotamers of all residues is selected and assigned. This process repeats until all residues have assigned side chains. It is possible, mostly due to a grossly incorrect main chain conformation, that all rotamers for a residue are invalid; in such cases no side chain is assigned

Finally, B factors were reset to uniform values for main chain and side chain atoms (for values, see Supplemental Materials at http://www.structure.org/cgi/content/full/12/5/831/DC1). Ligand and water atoms are ignoring during side chain assignment—i.e., they are not included in the excluded volume calculations—and are simply copied from the input to the output structure.

Refinement

Refinement and R factors calculations were performed using REFMAC (version 5.1.24) with a maximum-likelihood target function and the Babinet bulk solvent correction (Murshudov et al., 1997). REFMAC default values were used except that the bonded-atom B factor restraints were loosened; full details of the refinement parameters are available in the Supplemental Materials at http://www.structure.org/cgi/content/full/12/5/831/DC1. Refinement was carried out for 20 cycles; increasing the number of cycles to 200 or 2000 does not systematically improve R factors or move the resulting models closer to each other.

Miscellaneous

The ensemble structure factors F_{ens} are calculated by including all five final models in a single asymmetric unit, each contributing 20% to the atomic scattering. Simulated annealing omit maps were calculated with CNS using the Babinet bulk solvent correction (Brunger et al., 1998; Hodel et al., 1992). Side chain rotamers for both initial Cα-trace and automated refinement were taken from the penultimate rotamer library (Lovell et al., 2000) with idealized bond lengths and angles. Ramachandran and rotamer outliers were calculated with RAMPAGE and MOLPROBITY, respectively (Lovell et al., 2003). Main chain and all-atom rmsds were calculated over all heavy atoms without superposition; superimposed values are virtually identical.

RAPPER binaries and all data and models are publicly available at http://www-cryst.bioc.cam.ac.uk/rapper/.

Acknowledgments

We thank Fred E. Cohen and colleagues at UCSF for their support during the initial phases of this project and Ian Tickle, Andrew McCammon, and David Moss for insightful comments on this manuscript. This work was funded by the Marshall Aid Commemoration Commission, U.S. National Science Foundation, Cambridge Overseas Trust (M.A.D.), Cambridge European Trust, Isaac Newton Trust, and BBSRC (P.I.W.D.B.).

Received: October 17, 2003
Revised: January 21, 2004
Accepted: February 10, 2004
Published: May 11, 2004

References

Branden, C., and Jones, A.T. (1990). Between objectivity and subjectivity. Nature *343*, 687–689.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Burling, F.T., Weis, W.I., Flaherty, K.M., and Brunger, A.T. (1996). Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. Science *271*, 72–77.

CCP4 (Collaborative Computational Project 4) (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D *50*, 760–763.

Cunane, L.M., Chen, Z.W., Durley, R.C.E., and Mathews, F.S. (1996). X-ray structure of the cupredoxin amicyanin, from *Paracoccus denitrificans*, refined at 1.31 Å resolution. Acta Crystallogr. D Biol. Crystallogr. *52*, 676–686.

de Bakker, P.I.W., DePristo, M.A., Burke, D.F., and Blundell, T.L. (2003). Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins *51*, 21–40.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System on World Wide Web http://www.pymol.org/.

DePristo, M.A., de Bakker, P.I.W., Lovell, S.C., and Blundell, T.L. (2003a). Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins *51*, 41–55.

DePristo, M.A., de Bakker, P.I.W., Shetty, R.P., and Blundell, T.L. (2003b). Discrete restraint-based protein modeling and the Cα-trace problem. Protein Sci. *12*, 2032–2046.

Frauenfelder, H., Sligar, S.G., and Wolynes, P.G. (1991). The energy landscapes and motions of proteins. Science *254*, 1598–1603.

Hodel, A., Kim, S.H., and Brunger, A.T. (1992). Model bias in macromolecular crystal structures. Acta Crystallogr. A *48*, 851–858.

Jensen, L.H. (1997). Refinement and reliability of macromolecular models based on X-ray diffraction data. In Methods in Enzymology, C.W. Carter Jr., and R.M. Sweet, eds. (New York: Academic Press, Inc.), pp. 353–366.

Kleywegt, G.J. (1996). Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallogr. D Biol. Crystallogr. *52*, 842–857.

Kleywegt, G.J. (1999). Experimental assessment of differences between related protein crystal structures. Acta Crystallogr. D Biol. Crystallogr. *55*, 1878–1884.

Kuriyan, J., Petsko, G.A., Levy, R.M., and Karplus, M. (1986). Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics. J. Mol. Biol. *190*, 227–254.

Kuriyan, J., Karplus, M., and Petsko, G.A. (1987). Estimation of uncertainties in X-ray refinement results by use of perturbed structures. Proteins *2*, 1–12.

Kuriyan, J., Osapay, K., Burley, S.K., Brunger, A.T., Hendrickson, W.A., and Karplus, M. (1991). Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins *10*, 340–358.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. (2000). The penultimate rotamer library. Proteins *40*, 389–408.

Lovell, S.C., Davis, I.W., Arendall, B., III, de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003). Structure validation by Cα geometry: ϕ, ψ, and Cβ deviation. Proteins *50*, 437–450.

Luzzati, P.V. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. Acta Crystallogr. *5*, 802–810.

McCammon, J.A. (1999). Are molecular motions important for function? In Simplicity and Complexity in Proteins and Nucleic Acids, H. Frauenfelder, J. Deisenhofer, and P.G. Wolynes, eds. (Berlin: Dahlem University Press), pp. 193–198.

McCammon, J.A., and Harvey, S.C. (1987). Dynamics of Proteins and Nucleic Acids (Cambridge, MA: Cambridge University Press).

Mowbray, S.L., Helgstrand, C., Sigrell, J.A., Cameron, A.D., and Jones, T.A. (1999). Errors and reproducibility in electron-density map interpretation. Acta Crystallogr. D Biol. Crystallogr. *55*, 1309–1319.

Murshudov, G.N., and Dodson, E.J. (1997). Simplified error estimation *a la* Cruickshank in macromolecular crystallography. In CCP4 Newletter, S Bailey, ed. (Cheshire, UK: Daresbury Laboratory).

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240–255.

Ohlendorf, D.H. (1994). Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin-1β. Acta Crystallogr. D Biol. Crystallogr. *50*, 808–812.

Petsko, G.A. (1996). Not just your average structures. Nat. Struct. Biol. *3*, 565–566.

Rader, S.D., and Agard, D.A. (1997). Conformational substates in enzyme mechanism: the 120 K structure of alpha-lytic protease at 1.5 Å resolution. Protein Sci. *6*, 1375–1386.

Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallogr. A *42*, 140–149.

Rejto, P.A., and Freer, S.T. (1996). Protein conformational substates from X-ray crystallography. Prog. Biophys. Mol. Biol. *66*, 167–196.

Ringe, D., and Petsko, G.A. (1986). Study of protein dynamics by X-ray diffraction. In Methods in Enzymology, C.H.W. Hirs and S. N. Timasheff, eds. (Orlando, FL: Academic Press, Inc.), pp. 389–433.

Schaal, W., Karlsson, A., Ahlsen, G., Lindberg, J., Andersson, H.O., Danielson, U.H., Classon, B., Unge, T., Samuelsson, B., Hulten, J., et al. (2001). Synthesis and comparative molecular field analysis (CoMFA) of symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors. J. Med. Chem. *44*, 155–169.

Shetty, R.P., de Bakker, P.I.W., DePristo, M.A., and Blundell, T.L. (2004). On the advantages of fine-grained side chain conformer libraries for protein modelling. Protein Eng. *16*, 963–969.

Smith, J.L., Hendrickson, W.A., Honzatko, R.B., and Sheriff, S. (1986). Structural heterogeneity in protein crystals. Biochemistry *25*, 5018–5027.

Stec, B., and Phillips, G.N., Jr. (2001). How the CO in myoglobin acquired its bend: lessons in interpretation of crystallographic data. Acta Crystallogr. D Biol. Crystallogr. *57*, 751–754.

Stec, B., Zhou, R., and Teeter, M.M. (1995). Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. Acta Crystallogr. D Biol. Crystallogr. *51*, 663–681.

Tickle, I.J., Laskowski, R.A., and Moss, D.S. (1998). Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of gammaB- and betaB2-crystallin. Acta Crystallogr. D Biol. Crystallogr. *54*, 243–252.

Tramontano, A., Leplae, R., and Morea, V. (2001). Analysis and assessment of comparative modeling predictions in CASP4. Proteins Suppl. 5, 22–38.

van Gunsteren, W.F., and Berendsen, H.J.C. (1990). Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. Angew. Chem. Int. Ed. Engl. *29*, 992–1023.

Vitkup, D., Ringe, D., Karplus, M., and Petsko, G.A. (2002). Why protein R-factors are so large: a self-consistent analysis. Proteins *46*, 345–354.

Wilson, M.A., and Brunger, A.T. (2000). The 1.0 Å crystal structure of Ca(2+)-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. J. Mol. Biol. *301*, 1237–1256.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. *285*, 1735–1747.

Yu, B., Blaber, M., Gronenborn, A.M., Clore, G.M., and Caspar, D.L. (1999). Disordered water within a hydrophobic protein cavity visualized by X-ray crystallography. Proc. Natl. Acad. Sci. USA *96*, 103–108.

Zhang, X.J., Wozniak, J.A., and Matthews, B.W. (1995). Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. *250*, 527–552.

Zhou, Y., Vitkup, D., and Karplus, M. (1999). Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. J. Mol. Biol. *285*, 1371–1375.

Zoete, V., Michielin, O., and Karplus, M. (2002). Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J. Mol. Biol. *315*, 21–52.