Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Exploring structural variability in X-ray crystallographic models using protein local optimization by torsion-angle sampling

Jennifer L. Knight,[a] Zhiyong Zhou,[b] Emilio Gallicchio,[c] Daniel M. Himmel,[c] Richard A. Friesner,[d] Eddy Arnold[c] and Ronald M. Levy[c]*

[a]The Scripps Research Institute, La Jolla, CA 92037, USA, [b]Schrodinger Inc., New York, NY, USA, [c]Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, and [d]Columbia University, New York, NY, USA

Correspondence e-mail: ronlevy@lutece.rutgers.edu

Received 5 November 2007
Accepted 8 January 2008

Modeling structural variability is critical for understanding protein function and for modeling reliable targets for *in silico* docking experiments. Because of the time-intensive nature of manual X-ray crystallographic refinement, automated refinement methods that thoroughly explore conformational space are essential for the systematic construction of structurally variable models. Using five proteins spanning resolutions of 1.0–2.8 Å, it is demonstrated how torsion-angle sampling of backbone and side-chain libraries with filtering against both the chemical energy, using a modern effective potential, and the electron density, coupled with minimization of a reciprocal-space X-ray target function, can generate multiple structurally variable models which fit the X-ray data well. Torsion-angle sampling as implemented in the *Protein Local Optimization Program* (*PLOP*) has been used in this work. Models with the lowest $R_{\rm free}$ values are obtained when electrostatic and implicit solvation terms are included in the effective potential. HIV-1 protease, calmodulin and SUMO-conjugating enzyme illustrate how variability in the ensemble of structures captures structural variability that is observed across multiple crystal structures and is linked to functional flexibility at hinge regions and binding interfaces. An ensemble-refinement procedure is proposed to differentiate between variability that is a consequence of physical conformational heterogeneity and that which reflects uncertainty in the atomic coordinates.

1. Introduction

Structural flexibility and dynamics both play an important role in protein function. Local atomic fluctuations and large-scale conformational changes affect the ability of macromolecules to bind ligands, recognize protein surfaces and catalyze reactions (Koshland, 1963; Gutteridge & Thornton, 2005; Karplus *et al.*, 2005; Alberts *et al.*, 2002). Effectively modeling structural variability is a crucial step towards understanding the interplay between protein function, flexibility and dynamics and for developing reliable targets for *in silico* docking experiments. Recent studies have demonstrated the increased predictive power of structure-based drug-design strategies that account for structural variability (Bonvin, 2006; Ehrlich *et al.*, 2005; Halperin *et al.*, 2002; Sherman *et al.*, 2006).

In traditional protein crystallography, a single three-dimensional model is generally used to represent a dynamic ensemble of structures. Atomic fluctuations are encapsulated in the isotropic or anisotropic *B*-factor terms (Willis & Pryor, 1975; Kuriyan *et al.*, 1986; Westhof *et al.*, 1986). In related approaches, such as translation, libration and screw-rotation (TLS) refinement (Winn *et al.*, 2001) and normal-mode analysis (NMA; Kidera & Go, 1990; Chen *et al.*, 2007; Poon *et*

al., 2007), collective displacement variables are used to describe anisotropic and correlated atomic fluctuations. Side-chain flexibility tends to be modeled as an average conformation with elevated *B* factors or, less frequently, as multiple coordinates for a given residue with scaled occupancy factors (Stec *et al.*, 1995; Smith *et al.*, 1986; Rejto & Freer, 1996). In a recent letter to *Nature Structural and Molecular Biology*, Furnham and coworkers contend that these single-conformer models provide little information about the uncertainty in the model or the heterogeneity that is present in the crystal (Furnham *et al.*, 2006). These authors suggest that ensembles of models would be more appropriate representations of a macromolecule and that these ensembles would provide end-users with information about the range of structures that should be considered in subsequent analyses of the models.

More extensive descriptions of conformational variability have been achieved by ensemble refinement, in which multiple complete structures are refined simultaneously (Rader & Agard, 1997; Burling & Brünger, 1994; Kuriyan *et al.*, 1991). In this approach, each conformer is generally assigned a fractional occupancy equal to the reciprocal of the number of copies and, while each individual copy is not necessarily a good model of the macromolecule, the ensemble is in good agreement with the X-ray reflection data. A recent study using synthetic data has demonstrated that ensemble refinement of an ensemble of conformers can substantially reduce the $R_{\rm free}$ values and improve the estimation of the magnitude and anharmonicity of motions within macromolecular X-ray structures (Levin *et al.*, 2007). However, most current X-ray structure-refinement methods are labor-intensive as stepwise improvements are made to models by iterating between automated refinement and manual intervention. Therefore, more extensive modeling of structural heterogeneity and uncertainty will depend more heavily on automated procedures, especially for structures exhibiting concerted differences and where multiple models need to be refined in parallel.

A promising representation of structural variability, which we are exploring in this work, consists of the generation of families of single-conformer crystallographic models consistent with the reflection data. However, exploring the complex energy landscape of macromolecules to identify alternative structures is particularly challenging. Molecular-dynamics and simulated-annealing protocols have expanded the range of conformations that can be sampled using traditional crystallographic refinement (Brunger & Adams, 2002; Brunger *et al.*, 1999; Brunger, Adams & Rice, 1998; Brünger *et al.*, 1987). However, even with these tools it is difficult to overcome the large energy barriers associated with backbone rearrangements and/or side-chain re-packing (DePristo *et al.*, 2005).

It was recently demonstrated that the program *RAPPER*, which uses libraries of backbone dihedral angles and side-chain rotamers derived from high-resolution structures, can generate ensembles of single-conformer models in which each model satisfies given restraints; for example, agreement with experimental electron density (DePristo *et al.*, 2004, 2005). Using an automated protocol built around the program *RAPPER*, DePristo and coworkers constructed multiple models for three macromolecules that fitted the experimental X-ray crystallographic reflection data comparably well (DePristo *et al.*, 2004). These authors concluded that the uncertainty in crystallographic structures has been underestimated and information may be lost if only a single model is used to represent a macromolecule. Recently, Terwilliger *et al.* (2007) constructed sets of high-quality single-conformer models using a strategy that includes fragment-based loop building and splicing together segments from multiple loop candidates based on their fit to X-ray data. Terwilliger and coworkers suggest that the variation among the structures in the resulting ensemble provides an estimate of the precision of a macromolecular model and forms a lower bound on the uncertainty in the coordinates of the individual models.

In this paper, we describe how torsion-angle sampling with scoring *via* a modern effective potential can be used to efficiently explore the degree of structural variability that is consistent with a given set of X-ray reflections. We propose an iterative approach in which each cycle involves (i) an efficient torsion-angle search using backbone and side-chain rotamer libraries, hierarchical screening and clustering, and scoring with an all-atom effective potential function to generate an ensemble of low-energy conformations for a five-residue segment; (ii) identification of the conformation that has the best agreement with an experimental electron-density map in real space and (iii) a short optimization of the new structure using a reciprocal-space X-ray target function. This cycle is repeated using a target window to define which segment of five residues is modeled by torsion-angle sampling and then sliding the target window along the entire sequence of the macromolecule. We use the *Protein Local Optimization Program* (*PLOP*) to carry out the torsion-angle sampling (Jacobson *et al.*, 2004). *PLOP* has been used previously to model side-chain and loop conformations (Andrec *et al.*, 2002; Jacobson *et al.*, 2002, 2004; Zhu *et al.*, 2006), crystal-packing interactions (Jacobson *et al.*, 2002) and binding pockets in induced-fit docking (Sherman *et al.*, 2006).

While the current work is similar in spirit to the approaches described by DePristo and coworkers and by Terwilliger and coworkers, there are significant differences with respect to the methodology employed, the analysis of the results and the goals of the project. One of the fundamental differences between the torsion-angle sampling strategies is the adoption in this work of a physics-based effective potential to filter and score candidate structures for refinement. *RAPPER* avoids the use of a chemical energy function by fitting candidate structures to the X-ray data at an earlier stage in the filtering process and Terwilliger and coworkers filter and score candidates primarily using X-ray data criteria. While there are several other significant differences in the sampling strategies (Jacobson *et al.*, 2004; de Bakker *et al.*, 2003; Terwilliger *et al.*, 2007), perhaps the most important distinguishing feature of the current work is our focus on generating sets of models that are as diverse as possible which fit the X-ray data well. Finally, in our analysis, we use a multi-conformer model to explore the

possible contributions to conformational variability and propose criteria to identify where differences among models are likely to arise from the true conformational heterogeneity that exists in the crystal. Using synthetic data, Terwilliger and coworkers have shown that model variability can represent either the range of structures that are compatible with the experimental data (*i.e.* what we call positional uncertainty) or the set of structures that is actually present in the crystal (*i.e.* what we call conformational heterogeneity). However, they acknowledge that the latter effect is not addressed by their work and that it would be more appropriately studied by an ensemble-refinement procedure, such as that employed in this work, in which structures are refined as a group against the crystallographic data.

In this paper, we chose five proteins spanning 1.0–2.8 Å resolution as test cases to illustrate how, starting from a PDB structure, ensembles of structurally variable high-quality structures can be obtained by iterative use of torsion-angle sampling on an effective potential surface with filtering by and optimizing against experimental X-ray data. We describe the structural variability among the resulting models, compare the fits of the individual models and of the ensemble to the experimental X-ray data and explore the role played by the effective potential in generating high-quality ensembles. Using HIV-1 protease, calmodulin and SUMO-conjugating enzyme as examples, we show how variability in the ensemble of *PLOP* structures captures the structural variability that is observed across multiple crystal structures and in NMR ensembles and is linked to functional flexibility at hinge regions and binding interfaces. Explicitly modeling structural variability is particularly important for gaining insight into binding-site plasticity as well as the conformational flexibility of binding interfaces and for constructing multiple structures that can be used as three-dimensional targets in structure-based drug design.

2. Methods

2.1. Protein structures and reflection data

Atomic coordinates and structure factors (including test-set and training-set assignments) for 1g35 (Schaal *et al.*, 2001), 1a3s (Tong *et al.*, 1997), 1exr (Wilson & Brunger, 2000), 9ilb (Yu *et al.*, 1999) and 1ew4 (Cho *et al.*, 2000) were obtained from the Protein Data Bank (Berman *et al.*, 2000, 2003). Two of the test cases (1g35 and 9ilb) were used by DePristo and coworkers to test the *RAPPER* protocol. *RAPPER*-generated models for HIV-1 protease were obtained from http://www-cryst.bioc.cam.ac.uk/rapper/ (DePristo *et al.*, 2004). To reduce the impact of model bias and allow a more thorough exploration of conformational space and increased structural variability among the final models, for each macromolecule ten different initial structures were generated by simulated annealing starting from the PDB structure (see supplementary material[1] for details).

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: EN5274). Services for accessing this material are described at the back of the journal.

2.2. Iterative X-ray structure refinement using protein local optimization with torsion-angle sampling

Each cycle in our iterative protocol consists of (i) an extensive torsion-angle search in *PLOP* to generate an ensemble of low-energy conformations for a segment of five residues, (ii) identification of the *PLOP* candidate with the best agreement to the X-ray data based on the real-space correlation coefficient (RSCC) of the modeled segment and (iii) a short optimization of the new structure in *CNS* using the maximum-likelihood function. Six different start sites for the target window were used on each of the 11 initial structures [ten simulated annealing with molecular dynamics (SA/MD) structures and the PDB model] to generate a total of 66 final *PLOP* structures for each protein. Between each cycle, the target window was translocated along the sequence by three residues: in general, from the start site to the C-terminus and then from the start site to the N-terminus. The resulting ensemble of *PLOP* models was filtered to remove variability among the models that did not represent comparable or improved alternatives relative to the PDB structure. Each of these steps in the cycle is described below.

Step 1: hierarchical torsion-angle sampling. Loop prediction in *PLOP* is accomplished *via* an *ab initio* construction procedure which, at the limit of highest resolution, exhaustively searches the phase space of possible loop geometries connecting the two loop stems. The method achieves both efficiency and high accuracy *via* deployment of a hierarchy of scoring functions; rapid screening functions are used to eliminate large numbers of high-energy loops at early stages, ultimately yielding a relatively small number of candidates that are evaluated *via* minimization with the accurate OPLS-AA/SGBNP effective energy function. See the supplementary material and Jacobson *et al.* (2004) for more details.

Step 2: filtering PLOP candidates. For each *PLOP* candidate that was within 84 kJ mol^{-1} of the lowest energy model (in practice 5–30 candidates), $2F_o - F_c$ ($3F_o - 2F_c$ for the low-resolution structures 1a3s and 9ilb) and F_c maps were generated in *CNS* v. 1.1 (Brünger, Adams, Clore *et al.*, 1998). The mean RSCC for the targeted five-residue segment in each *PLOP* candidate was calculated in *MAPMAN* (Jones *et al.*, 1991; Kleywegt & Jones, 1996) and the *PLOP* candidate with the highest mean RSCC for the remodeled segment was selected as the optimal *PLOP* candidate.

Step 3: refinement of the optimal PLOP candidate. The optimal *PLOP* candidate was subjected to a restrained coordinate optimization (two cycles of ten steps of conjugate-gradient energy minimization) and, for the high-resolution structures, 30 steps of *B*-factor optimization. The *CNS*-optimized structure became the seed structure for the subsequent cycle of *PLOP* modeling in which a new target window was defined. Steps 1–3 were repeated until each residue in the protein had been sampled by *PLOP* at least once.

Step 4: filtering the ensemble of PLOP models. The ensemble of 66 *PLOP* models was filtered to remove variability in the *PLOP* ensemble that was not achieved with a similar or improved RSCC relative to the PDB structure. The

residue-specific RSCCs (resRSCCs) for all *PLOP* candidates that were variable at a given residue were evaluated in *MAPMAN* from the corresponding $2F_o - F_c$ ($3F_o - 2F_c$ for the low-resolution structures 9ilb and 1a3s) map generated by *CNS*. If at a given variable residue all alternative conformations demonstrated degraded resRSCCs relative to the PDB structure, *PLOP* models which exhibited variability at this site were removed. Degradations in resRSCCs were described by

$$\text{resRSCC(PDB}, i) - \text{resRSCC}(PLOP_j, i) > 0.5[1 - \text{avg5resRSCC(PDB}, i)]$$

or

$$\text{resRSCC(PDB}, i) - \text{resRSCC}(PLOP_j, i) > 0.03$$

and

$$\text{resRSCC(PDB}, i) - \text{resRSCC}(PLOP_j, i) > 0.25[1 - \text{avg5resRSCC(PDB}, i)],$$

where avg5resRSCC is the resRSCC averaged over residues $i - 2$ through $i + 2$. In cases where over half of the *PLOP* models showed variability and all variability was degraded by the above criteria, rather than eliminate the structures, *PLOP* variability at this residue was described as a false positive. Out of the 214 variable residues across the five proteins, only nine were false positives.

2.3. Optimizing ensemble occupancy values

For each protein, all possible combinations of five *PLOP* models from the filtered ensemble were identified. The subset of five *PLOP* structures that had the largest number of distinct variable residues compared with the corresponding PDB structure was selected along with the PDB structure to undergo ensemble optimization. Where multiple sets of structures fitted these criteria, the subset with the lowest average *R* value was selected. Occupancy values for the PDB structure and each of the structures in the selected set were optimized by sampling 3500 different initial occupancy values *via* Monte Carlo sampling and minimization in *CNS*, *i.e.* the occupancy values were the only adjustable parameters while the X-ray target function was minimized.

3. Results

3.1. Trends in modeling structural variability in X-ray structure refinement

3.1.1. Summary of model quality. The automated iterative protocol we developed typically generates 20–60 structural models for each protein upon completion of the refinement. The model quality and variability in the ensembles of single-conformer structures are summarized in Table 1. For all five proteins studied, this procedure generates models of equal or higher quality than the original PDB structure, with similar mean real-space correlation coefficients (RSCC) and improvements in R_{free} of up to one percentage point. Because reciprocal-space criteria were not used to filter the ensembles, some models with relatively large R_{free} values containing improved local fits to the electron density in regions of variability are retained in the ensembles. In all the models the bond lengths and angles have close to ideal geometry; between 96 and 100% of the residues are found in Ramachandran core and allowed regions. Table 2 shows comparable model quality for the sets of *PLOP* and *RAPPER* single-conformer models of HIV-1 protease.

Table 1
Summary of model quality and variability.

Results for ensembles of single-conformer *PLOP* models for calmodulin (1exr), CyaY protein (1ew4), HIV-1 protease (1g35), human interleukin-1β (9ilb) and SUMO-conjugating enzyme (1a3s). Minimum and maximum values for structures in each ensemble are reported.

PDB code	1exr†	1ew4	1g35	9ilb	1a3s
Resolution (Å)	1.0	1.4	1.8	2.3	2.8
PDB‡ *R*	0.232	0.208	0.179	0.148	0.205
PDB‡ R_{free}	0.254	0.230	0.225	0.205	0.266
PDB‡ RSCC	0.93	0.93	0.95	0.95	0.92
No. of *PLOP* structures	32	20	40	53	38
Backbone r.m.s.d. (Å)	0.15–0.18	0.08–0.18	0.08–0.09	0.27–0.49	0.28–0.59
Non-H atoms r.m.s.d. (Å)	0.65–0.88	0.61–0.87	0.47–0.69	0.81–0.99	0.91–1.18
Total nonglycine residues	135	97	172	145	149
No. of variable side chains					
Ensemble total§	50	26	46	62	76
False positives¶	3	5	0	0	1
PLOP model quality††					
R	0.234–0.245	0.208–0.230	0.178–0.185	0.157–0.165	0.209–0.220
R_{free}	0.254–0.271	0.228–0.248	0.216–0.236	0.195–0.211	0.262–0.284
Average RSCC	0.92–0.93	0.92–0.93	0.95	0.95–0.96	0.91–0.92
Average RSR	0.211–0.221	0.193–0.202	0.149–0.155	0.104–0.115	0.164–0.173
Average *B* factor (Å^2)	17.1–17.7	19.5–19.8	20.5–20.8	40.3–41.2	42.1–43.6
Minimum *B* factor (Å^2)	5.6–6.0	6.5–8.8	5.4–6.3	11.4–12.3	13.4–14.9
Maximum *B* factor (Å^2)	53.5–57.6	65.6–99.0	58.3–96.5	177.0–178.9	96.4–97.9
R.m.s.d. bonds (Å)	0.022–0.025	0.020–0.025	0.025–0.028	0.029–0.033	0.009–0.010
R.m.s.d. angles (°)	1.7–1.9	2.0–2.1	2.2–2.8	2.5–2.8	1.5–1.6
R.m.s.d. dihedrals (°)	20.7–21.8	24.6–25.4	26.1–26.7	26.9–27.8	23.3–24.4
R.m.s.d. impropers (°)	1.3–1.5	1.3–1.4	1.6–1.9	1.5–1.8	1.1–1.3
Ramachandran plot: core (%)	93–95	94–95	96–98	84–90	85–90
Ramachandran plot: allowed (%)	5–7	4–6	3–4	12–16	8–14
Unfavorable χ_1–χ_2‡‡ (%)	0–2	0–2	0–2	1–5	1–7

† Coordinates for the 'A' conformations for the 37 discretely disordered residues in 1exr were omitted and the occupancy values were set to 1.0; thus, the *R* and R_{free} are substantially poorer than those for the full published model. ‡ PDB results were computed by performing cycles of *CNS* optimization (minimization for all proteins and *B*-factor optimization for 1exr, 1ew4 and 1g35) without *PLOP* torsion-angle sampling. § Total number of distinct side chains in the ensemble of *PLOP* models that are in different conformations relative to the PDB structure. ¶ False positives are defined as variable residues at which over half the *PLOP* structures exhibit variability and the quality of the alternative conformations is degraded relative to the PDB structure. †† *R*, R_{free} and coordinate errors were computed in *CNS*. Real-space correlation coefficient (RSCC) and real-space *R* value (RSR) values averaged over each residue were computed in *MAPMAN*. ‡‡ Number of residues lying in unfavorable regions of the χ_1–χ_2 torsion-angle plots.

Table 2
HIV-1 protease structures: summary of model quality and variability.

Results for 1g35 are reported as means and standard deviations where applicable. Results for the *RAPPER* and *PLOP* structures are reported as a range representing the minimum and maximum values for structures in each ensemble of single-conformer models. The same abbreviations are used as in Table 1.

Set of models	1g35	*RAPPER*	*PLOP*
No. of models	1	5	40
Backbone r.m.s.d. (Å)	—	0.08–0.21	0.08–0.09
Heavy-atom r.m.s.d. (Å)	—	0.53–0.57	0.47–0.69
No. of variable side chains			
Ensemble total	—	32	46
Relative to 1g35	—	19–25	14–23
Pairwise†	—	6–13	5–25
Measures of model quality			
R	0.178	0.179–0.186	0.178–0.185
$R_{\rm free}$	0.225	0.216–0.224	0.216–0.236
Average RSCC	0.949 ± 0.027	0.948–0.951	0.948–0.952
Average RSR	0.152 ± 0.037	0.150–0.154	0.149–0.155
Average B factor (Å^2)	20.6 ± 9.2	20.5–20.7	20.5–20.8
Minimum B factor (Å^2)	6.3	5.1–5.9	5.4–6.3
Maximum B factor (Å^2)	58.6	61.1–75.1	58.3–96.5
R.m.s.d. bonds (Å)	0.028	0.030–0.031	0.025–0.028
R.m.s.d. angles (°)	2.9	3.2–3.4	2.2–2.8
R.m.s.d. dihedrals (°)	25.6	26.2–27.0	26.1–26.7
R.m.s.d. impropers (°)	2.1	2.2–2.5	1.6–1.9

† Number of side chains in different conformations between two *PLOP* models. Differences between all pairs of *PLOP* models were evaluated.

3.1.2. Summary of model variability. The *PLOP* structures have mean backbone r.m.s.d.s of 0.08–0.6 Å and mean heavy-atom r.m.s.d.s of 0.5–1.2 Å relative to the corresponding PDB structures. The range of backbone and heavy-atom r.m.s.d.s depicted in Fig. 1(*a*) shows that both the variability in atomic coordinates relative to the PDB as well as within a *PLOP* ensemble tend to increase with decreasing resolution. HIV-1 protease at 1.8 Å resolution is an outlier, but its more limited variability is explained by its significantly higher proportion of buried residues relative to the other proteins.

Side chains are defined as being in an alternative conformation if any atom in the *PLOP*-generated side chain is more than 1 Å away from the closest atom of the same residue in the PDB structure. This working definition provides an estimate of when atoms in a given conformation will be positioned outside the envelope of electron density associated with the reference structure. In this way, ring flips or concerted changes in dihedral angles that result in side-chain atoms occupying the same volume as the reference structure will not be identified as 'variable'.

With the above definition, 15–40% of the side chains in any individual *PLOP* structure are modeled in a different conformation relative to the corresponding PDB structure. The filtered ensemble of *PLOP* structures for each protein contains alternative side-chain conformations for 25–50% of the residues in the sequence. The side-chain variability we measure is comparable to the 30% reported previously with a less stringent criterion of variability (Stec *et al.*, 1995). Systematically omitting structures generated from a given initial condition (*i.e.* any one of the ten initial simulated-annealing structures or any one of the six different start sites for the target windows) yielded reduced variability at no more than two side chains or less then 4% of the total variability

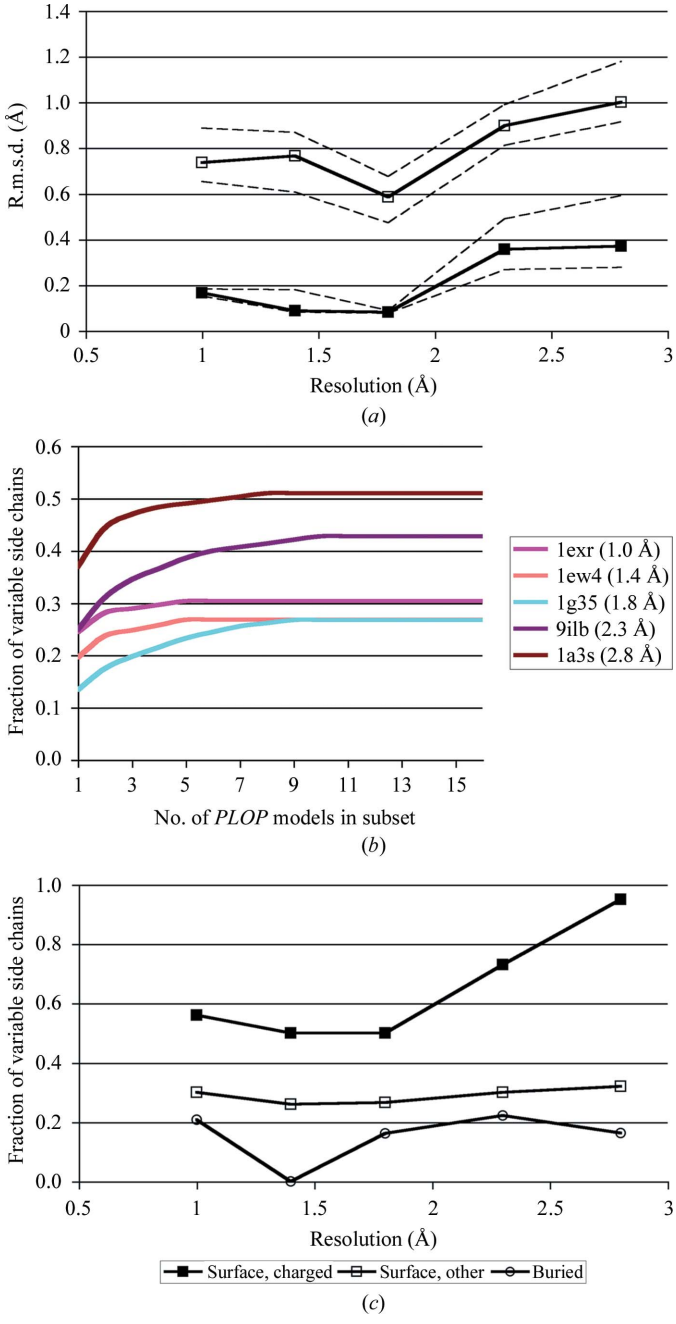

Figure 1
Structural variability among *PLOP* ensembles. (*a*) Median backbone (filled squares) and heavy-atom (empty squares) r.m.s.d. between *PLOP* models and the PDB structure as a function of resolution. Dashed lines indicate the corresponding minimum and maximum r.m.s.d. values in the respective ensembles of single-conformer *PLOP* models. (*b*) The number of distinct side chains that are in different conformations relative to the PDB structure was evaluated for every combination of n *PLOP* models. The reported percentage side-chain variability is the maximum number of variable side-chain conformations for a given number of *PLOP* models (n) divided by the number of nonglycine residues in the corresponding protein. (*c*) Side-chain variability in the respective *PLOP* ensembles relative to the PDB structure categorized by charged surface residues (filled squares), neutral surface residues (empty squares) and buried residues (circles).

Table 3
HIV-1 protease: distribution of variable side chains based on 1g35.

RAPPER and *PLOP* structures were generated from the X-ray reflection data associated with 1g35. Each residue was classified according to its side-chain electron density in the 1g35 $2F_o - F_c$ map, as well as its real-space correlation coefficient (RSCC) and mean *B* factor in 1g35. Each variable side chain using the 1 Å distance cutoff is categorized as unique to the *RAPPER* structures, unique to the *PLOP* structures or common to both sets of structures.

	Total residues	Total variable *RAPPER* residues	Total variable *PLOP* residues	Total variable residues
Total nonglycine residues	172	32	46	52

	Total residues	Variable residues unique to *RAPPER*	Variable residues unique to *PLOP*	Variable residues common to both	Total variable residues
Environment					
Surface, charged	34	4	4	13	21
Surface, other	64	0	7	10	17
Surface, total	98	4	11	23	38
Buried, total	74	2	9	3	14
Total	172	6	20	26	52
Electron density					
None, weak	10	1	0	8	9
Ambiguous	39	2	11	17	30
Well defined	123	3	9	1	13
RSCC					
0.843–0.936	43	0	5	16	21
0.936–0.956	49	0	10	9	19
0.956–0.969	46	3	4	0	7
0.969–1.000	34	3	1	1	5
Mean *B* factor (Å^2)					
23.0–45.2	43	0	8	17	25
17.0–23.0	43	2	8	7	17
13.5–17.0	41	0	1	2	3
7.0–13.5	45	4	3	0	7

that was present in the filtered ensemble. In most cases, there was no loss of variability. These tests indicate that this automated procedure is capturing most of the allowed variability and that extending the cycles of refinement to generate more single-conformer models will not greatly increase the observed variability that is consistent with the crystallographic data. Owing to the redundancy in the ensemble of *PLOP* structures, we wanted to identify the minimum number of structures that are required to represent the side-chain variability that is observed in the full ensemble of *PLOP* structures. By comparing all possible combinations of subsets of *PLOP* structures for each protein, we computed the maximum number of distinct variable side-chain conformations relative to the PDB structure that are present for a given number of *PLOP* structures. Fig. 1(*b*) shows for each protein the percentage of side chains that are in alternative conformations relative to the PDB structure as a function of the number of *PLOP* structures in the subset. For each of the test cases, five *PLOP* structures are sufficient to capture at least 85% of the structural variability present in the full *PLOP* ensembles and with ten *PLOP* structures the full variability can be represented.

Fig. 1(*c*) illustrates the side-chain variability that is observed for the different classes of residues in the ensemble of *PLOP* structures and Table 3 summarizes, as an example, the characteristics of variable residues for *PLOP* and *RAPPER* ensembles of HIV-1 protease. Several trends have been described previously (Smith *et al.*, 1986; Stec *et al.*, 1995; Rejto & Freer, 1996). The majority of the residues exhibiting structural variability are on the surface of the protein and over half of these are long charged residues: lysine, arginine and glutamate residues. Most of the nonvariable charged surface side chains are restrained by ion pairing and intermolecular and intramolecular hydrogen bonding. Buried side chains account for ∼10% of the conformational variability and tend to involve valine or isoleucine residues in which the side chains have been rotated ∼120° about the χ_1 torsion angles so that one of the C^γ atoms in each conformation occupies the same density.

Unexpectedly, variability is observed among residues that in more manual crystallographic refinement strategies would be assumed to be in well defined single conformations and traditionally would neither be targeted for further refinement nor be regarded as candidate sites at which multiple conformations should be modeled. In fact, for the five proteins under investigation, 15–45% of the *PLOP* side chains that are modeled in different conformations relative to the PDB structure have low *B* factors and/or high RSCC values. However, rather than being indicative of *PLOP* errors, nearly all of these alternative *PLOP* conformations exhibit high RSCCs in the context of their own electron-density maps. This unanticipated variability suggests that the extensive automated sampling protocol based on backbone and side-chain rotamer libraries used in *PLOP* can be particularly effective at thoroughly exploring and refining regions that may be overlooked by manual model building. These results also serve as a cautionary note about the uncritical use of RSCC values as validation parameters owing to inherent difficulties with model bias (Kleywegt, 2000). In principle, some of these incongruities could be explored further using simulated-annealing OMIT maps during refinement.

3.1.3. Importance of the effective potential for refinement. The standard effective potential used in the *PLOP* modeling consists of the all-atom OPLS force field (Kaminski *et al.*, 2001; Jorgensen *et al.*, 1996; including bond, angle and dihedral energies as well as van der Waals and electrostatics energies) together with the SGB/NP continuum solvation model [the surface implementation of the generalized Born model (Ghosh *et al.*, 1998) and also including a nonpolar hydration free-energy estimator which represents the nonpolar energies as the sum of an unfavorable cavity work term plus a favorable

van der Waals dispersion term (Gallicchio *et al.*, 2002); 'sgbnp' in Table 4]. To assess the role of the effective potential in the strategy for crystallographic refinement using *PLOP*, we generated ensembles of single-conformer structures in which the electrostatics and solvation terms were eliminated from the *PLOP* energy ('noelec' in Table 4). This potential mimics the chemical energy generally used in *CNS* and *X-PLOR* (Moulinier *et al.*, 2003). The results summarized in Table 4 demonstrate that inclusion of the electrostatics term with solvation improves the refinement and generates structures with lower $R_{\rm free}$ values. Fig. 2 shows the distributions of $R_{\rm free}$ values for ensembles of single-conformer structures for 1g35 generated with and without the electrostatic and solvation terms included in the effective potential. 40% of the models generated with the full potential have $R_{\rm free}$ values that are improved relative to 1g35, compared with only 11% of the corresponding ensemble modeled without the electrostatics and solvation terms. In addition, twice as many 'sgbnp' models pass through the real-space filtering criteria compared with the 'noelec' models. This suggests that when the same discretized backbone and side-chain rotamer libraries are used, *PLOP* optimization and scoring using the OPLS-AA/SGBNP effective potential is significantly more effective than the truncated potential for generating conformations with the lowest $R_{\rm free}$ values and most favorable local features.

3.1.4. Ensemble models for distinguishing conformational heterogeneity from uncertainty in atomic positions. We further analyzed each of the five proteins to identify residues for which structural variability could be attributed to physical conformational heterogeneity of the sample rather than to ambiguities in the data (related to, for example, imperfections in the crystal, limitations in data acquisition or inadequacies in the refinement model). In order to address this question, it is necessary to resort to ensemble-refinement approaches (Levin *et al.*, 2007). We performed limited ensemble-refinement calculations based on the single-conformer sets described above. For each protein, the multi-conformer model consists of the original PDB structure and the subset of five single-conformer *PLOP* structures that together exhibit side-chain variability at the most residues in the sequence. The ensemble R and $R_{\rm free}$ were determined by minimizing the X-ray target function as a function of the occupancy values for each structure *via* a Monte Carlo procedure. The $R_{\rm free}$ values resulting from this procedure are summarized in Table 4 and indicate that the ensemble models provide equivalent or better representations of the reflection data than the best single-conformer models. In all cases, the optimized overall occupancy of the PDB structure is less than 0.5, confirming that the models generated from the automated *PLOP* refinement procedure indeed capture variability that is consistent with the X-ray reflection data.

We examined each of the ensembles to identify those residues for which conformational variability could be attributed to either conformational heterogeneity or positional uncertainty. Firstly, we visually inspected each of the corresponding $\sigma_{\rm A}$-weighted $2F_{\rm o} - F_{\rm c}$ electron-density maps contoured at the 1σ level. Variability at residues that did not have good side-chain density was attributed to positional uncertainty. Weak electron density in these areas generally indicates that no single conformation or small subset of conformations was adopted significantly more frequently than others.

Table 4
Summary of model quality: ensemble-refinement measurements and role of effective potential.

PDB code	1exr	1ew4	1g35	9ilb	1a3s
Resolution (Å)	1.0	1.4	1.8	2.3	2.8
PDB†					
R	0.232	0.208	0.179	0.148	0.205
$R_{\rm free}$	0.254	0.230	0.225	0.205	0.266
Best individual in ensemble‡					
R	0.235	0.208	0.178	0.159	0.209
$R_{\rm free}$	0.254	0.228	0.221	0.199	0.264
Ensemble§					
R	0.238	0.207	0.181	0.162	0.214
$R_{\rm free}$	0.253	0.224	0.214	0.181	0.265
Fractional ensemble occupancy¶					
PDB	0.39	0.40	0.31	0.44	0.23
PLOP	0.07–0.17	0.01–0.43	0.06–0.23	0.04–0.19	0.03–0.27
Best individual (sgbnp)††					
R	0.235	0.208	0.179	0.158	0.209
$R_{\rm free}$	0.254	0.228	0.216	0.195	0.262
Best individual (noelec)††					
R	0.236	0.211	0.180	0.158	0.210
$R_{\rm free}$	0.262	0.236	0.223	0.200	0.268

† PDB results are the same as described in Table 1. ‡ The lowest $R_{\rm free}$ structure in the highly diverse ensemble. The highly diverse ensemble was identified as the PDB structure and the subset of five *PLOP* structures (from the filtered ensemble) that yielded the largest number of distinct side-chain conformations relative to the PDB structure. § The lowest ensemble R resulting from optimizing the fractional occupancies of the highly diverse ensembles. ¶ The corresponding fractional occupancies after optimization. †† The lowest $R_{\rm free}$ structure from the filtered ensemble of single-conformer models generated using the full ('sgbnp') or truncated ('noelec') potentials in *PLOP*.

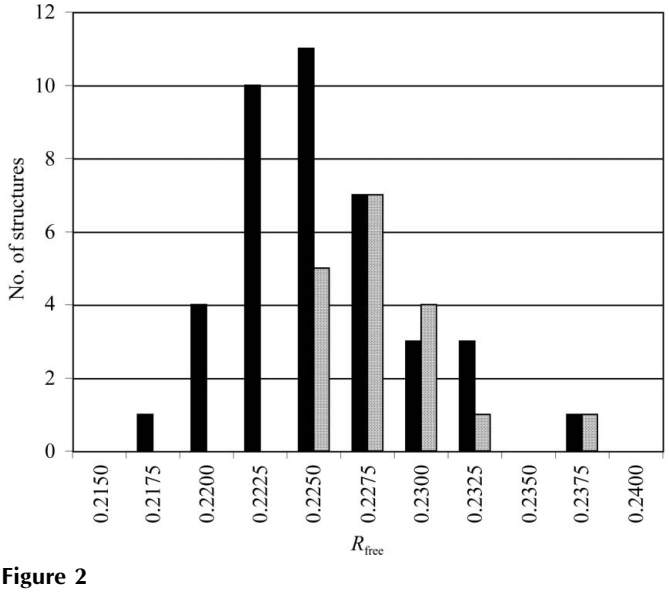

Figure 2
Higher quality models generated using the full potential in *PLOP*. The distribution of $R_{\rm free}$ values is shown for the filtered ensembles of *PLOP* models that were generated using the full potential ('sgbnp', solid black) and without the electrostatics and solvation terms ('noelec', shaded) in *PLOP*. 40 structures are in the filtered 'sgbnp' ensemble and 18 structures are in the filtered 'noelec' ensemble. The $R_{\rm free}$ of the PDB structure is 0.225.

Table 5
Distinguishing conformational heterogeneity from positional uncertainty.

PDB code	1exr	1ew4	1g35	9ilb	1a3s
Resolution (Å)	1.0	1.4	1.8	2.3	2.8
Total No. of nonglycine residues	135	97	172	145	149
No. of variable side chains†	40	16	36	55	73
Average $\sigma(B)$‡	1.00	1.19	1.09	1.64	1.41
Assignment					
Positional uncertainty	18	11	18	53	56
Physical heterogeneity	22	5	18	2	17
Physical heterogeneity with multiple density envelopes	12	3	2	0	0

† False positives in the highly diverse *PLOP* ensembles were omitted from further analysis and were not included in the number of variable side chains. ‡ For residues which contain good backbone and side-chain density.

Secondly, for each variable residue k for which there was continuous backbone and side-chain electron density, an occupancy-weighted spread in atomic coordinates from the variance of the distribution of positions in the ensemble was defined by

$$\sigma_k(S) = \max_j \left[\left(\sum_{i=1}^{N} p_i \left| \mathbf{r}_{ijk} - \sum_{i=1}^{N} p_i \mathbf{r}_{ijk} \right|^2 \right)^{1/2} \right], \qquad (1)$$

where N is the number of conformations in the ensemble, p_i is the occupancy of structure i in the optimized ensemble and $\mathbf{r}_{ijk}$ is the position of atom j in residue k in the ith member of the ensemble. The positional uncertainty for each residue was similarly estimated from the atomic B factors,

$$\sigma_k(B) = \max_j \left(\frac{3B_{jk}}{8\pi^2} \right)^{1/2}. \qquad (2)$$

We propose that to a first approximation the observed conformational variability of residue k is more likely to reflect the true structural heterogeneity present within the crystal when

$$\sigma_k(S) > \sigma_k(B). \qquad (3)$$

Fig. 3 shows a plot of $\sigma(S)$ *versus* $\sigma(B)$ for the variable residues in each of the five proteins studied. Minor changes in the threshold criterion will affect the assignment of variable residues with $\sigma(S) \simeq \sigma(B)$. Nevertheless, the results of this analysis, summarized in Table 5, shows that for four of the five proteins studied between 25% and 50% of the variable residues satisfy the criteria that $\sigma(S) > \sigma(B)$. The exception, 9ilb, has proportionally many more variable residues that are assigned to positional uncertainty than the other four proteins, primarily owing to the lack of side-chain density for many of its residues and the systematically higher B factors for the remaining variable residues.

To determine where side-chain variability would be anticipated by a PDB-phased electron-density map, we inspected the σ_A-weighted $2F_o - F_c$ electron-density maps phased by the corresponding PDB structure and identified residues at which the density at the 1σ contour level was visibly consistent with multiple conformations. The results are shown in Table 5. For each of the proteins except 1exr, alternative electron density has not been clearly identified for many residues as a consequence of one or more of the following: model bias in the phasing, weakly populated states or systematic refinement errors.

To validate the approach outlined above in discriminating physical heterogeneity from positional uncertainty, we compared the side chains modeled in multiple conformations in the 1exr PDB structure of calmodulin with those in the corresponding ensemble obtained in this work. 30 of the 37 residues modeled in alternative conformations in the PDB structure are described as 'variable' given our 1 Å criteria and half of these residues exhibit 'structural heterogeneity' according to the criteria above. In both sets of calmodulin models, the PDB structure as well as our ensemble, between 10% and 15% of the nonglycine residues exhibit physical heterogeneity (*i.e.* 15 residues for 1exr and between 12 and 22 for the ensemble). For 11 of the 15 1exr variable residues that are attributed to physical heterogeneity, very similar alternative conformations are also observed among the ensemble. For the remaining four variable residues, the spread in atomic positions within the ensemble is smaller than the corresponding spread observed in 1exr; for these residues, the variability in the ensemble is assigned to uncertainty. Thus, in general, there is a good correspondence between the variable residues that are assigned to physical heterogeneity using the automated procedure adopted here and from traditional model-building efforts.

3.2. Capturing structural variability: three case studies

3.2.1. Multiple crystal structures and *RAPPER* models: HIV-1 protease (1.8 Å). A thorough analysis of the structural flexibility of HIV-1 protease was performed by Zoete *et al.* (2002) using a database of X-ray structures. Mobility of the

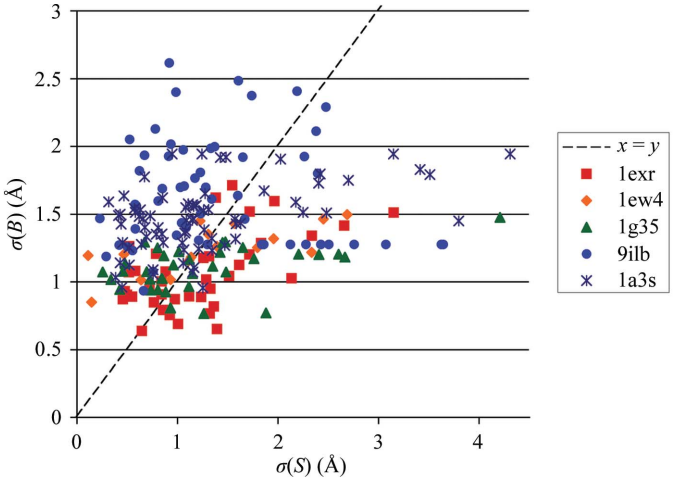

Figure 3
Contribution of variable side-chain conformations attributed to structural heterogeneity. Scatter plot of $\sigma(S)$ and $\sigma(B)$ values for each residue identified as 'variable' using the volumetric definition and a 1 Å cutoff. Residues for which there is poor side-chain density have been omitted for clarity. The dashed line indicates the threshold for attributing the modeled variability to structural heterogeneity (right of the line) or positional uncertainty (left of the line).

flap tips has been implicated in the function of HIV-1 protease by allowing substrate access to the catalytic aspartate residues (Erickson & Kempf, 1994; Miller *et al.*, 1989). Fig. 4(*a*) depicts these flexible regions in HIV-1 protease. Fig. 4(*b*) shows that while each *PLOP* structure exhibits different structural details, the largest deviations among the *PLOP* models for HIV-1 protease are localized in three surface loops (residues 14–20, 35–46 and 62–70). These variations are consistent with the results from normal-mode analyses or Gaussian network model analyses of HIV-1 protease crystal structures and snapshots along molecular-dynamics trajectories (Kurt *et al.*, 2003; Zoete *et al.*, 2002). The corresponding surface loops on chain *B* are stabilized by crystal contacts, thus breaking any symmetry of the dimer flexibility. In fact, the qualitative features of backbone variability observed across many crystal structures or predicted by molecular dynamics and normal-mode analysis are captured by both *RAPPER* and *PLOP* ensembles for HIV-1 protease. The magnitude of the structural variations in the *RAPPER* and *PLOP* ensembles, however, is significantly smaller than that observed by comparing multiple crystal structures (Zoete *et al.*, 2002), primarily owing to the specific physical packing forces which restrain the conformational flexibility in a single X-ray crystallographic experiment corresponding to a single crystal form.

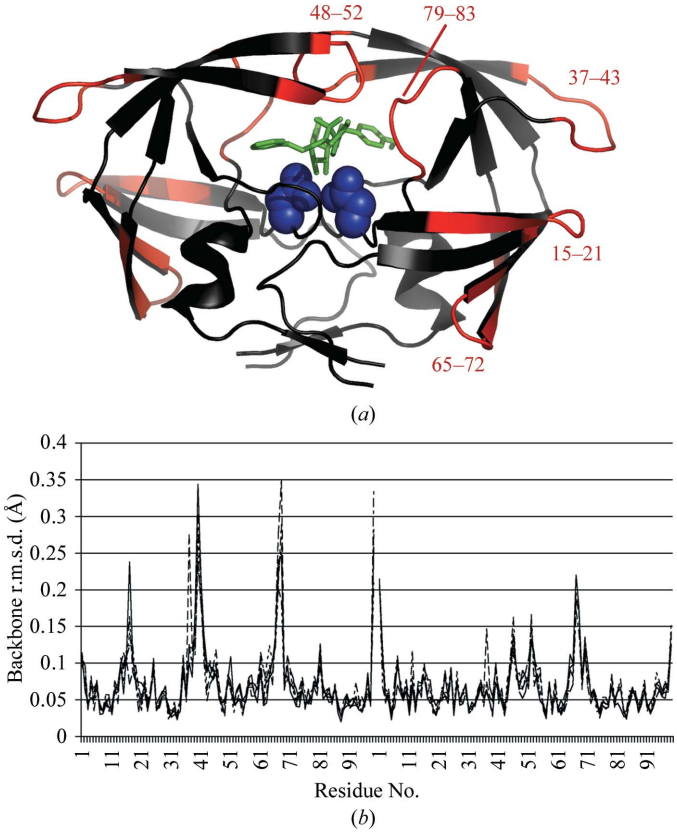

Figure 4
Structural variability in HIV-1 protease. (*a*) Cartoon representation of HIV-1 protease (1g35; Schaal *et al.*, 2001). The variable loops described by Zoete *et al.* (2002) are colored in red and the residue numbers are indicated; the ligand for 1g35 is colored in green and the catalytic aspartate residues are represented by blue spheres. (*b*) Backbone r.m.s.d. values as a function of residue number for five *PLOP* structures. The first 99 residues correspond to chain *A* and the last 99 residues correspond to chain *B*.

We observe greater side-chain variability among the ensemble of HIV-1 protease structures generated with *PLOP* than in the corresponding ensembles reported by DePristo *et al.* (2004). Of the 172 nonglycine residues in 1g35, 32 have a different conformation in at least one *RAPPER* model and 46 in at least one *PLOP* model; these results are shown in Table 2. Table 3 summarizes the environment of the HIV-1 protease side chains that exhibit variable conformations. The flap hinges and tips, the flexibility of which is necessary for protease activity, are densely populated with variable side-chain conformations: 43% and 50% of the nonglycine residues of the flap tip and hinge regions are variable in the *RAPPER* and *PLOP* models, respectively, compared with only 27% of the remaining nonglycine residues. The primary differences between the *PLOP* and *RAPPER* structures are the increase in variability among the neutral surface residues and buried residues. Five of the six side chains that are variable in the *RAPPER* structures and not in the *PLOP* structures have RSCCs for the alternative *RAPPER* side chains that are degraded by more than 0.04 relative to 1g35 and thus are indicative of *RAPPER* modeling errors or the modeling of weakly populated side-chain conformations. In contrast, all of the variable residues among the *PLOP* structures for 1g35 have alternative conformations that have comparable or better RSCCs for that residue relative to the PDB structure. 20 side chains are variable in the *PLOP* structures but not variable in the *RAPPER* structures; 19 of these residues have the 1g35/*RAPPER* conformation modeled by at least one of the *PLOP* structures, suggesting that multiple low-energy conformations exist in these regions that are consistent with the crystallographic data.

It is not clear whether the larger structural variability found for HIV-1 protease among the *PLOP* structures reported here relative to those reported by DePristo *et al.* (2004) using *RAPPER* reflects inherent differences in the search algorithms, since in the current work our explicit goal was to generate as diverse a set of high-quality structures as possible, whereas in DePristo *et al.* (2004) the goal was to construct a set of alternative structures with comparably low $R_{\rm free}$ values. In any case, we note that the underlying torsion-angle sampling algorithms are significantly different. *PLOP* is closer in spirit to exhaustive enumeration of backbone and side-chain rotamer libraries, the results of which are filtered using the chemical energy and then fitted to the electron density. In contrast, *RAPPER* torsion-angle sampling uses a genetic algorithm to build up fragments which are filtered by their fit to the electron density at an earlier stage in the build-up process.

3.2.2. Modeling multiple side-chain conformations at high resolution: calmodulin (1.0 Å). Wilson & Brunger (2000) published a high-resolution calmodulin structure, 1exr, in which 37 of 146 residues are modeled in two conformations. These residues are predominantly in the central helix and the

two hydrophobic binding pockets, which may permit target-specific recognition. Our initial model contained atomic coordinates from only one conformation (labeled 'B' in the PDB) for each residue in calmodulin. Among the final *PLOP* structures in the multi-conformer ensemble, seven residues are modeled exclusively in the 1exr 'A' conformation; in 1exr, these have high occupancy values for conformation A that range from 0.63 to 0.91. Three side chains are modeled by the *PLOP* refinement exclusively in the 1exr 'B' conformation; in 1exr, these residues have systematically lower occupancy values for the A conformation, *i.e.* between 0.31 and 0.48. For eight residues, the *PLOP* conformations appear to be staggered between the A and B side-chain conformations, reminiscent of interpolated structures based on A and B 'end points'. These residues have mid-range occupancy values (0.44–0.60) for the A conformation in the PDB structure. The remaining 12 side chains that have multiple conformations in the 1exr structure are modeled in multiple conformations in the *PLOP* ensemble.

Wilson and Brunger noted that there were indications of structural variability beyond that which they modeled explicitly using their contour-based criterion. In agreement with this observation, 24 additional side chains are modeled in alternative conformations in the *PLOP* ensemble. Two-thirds of these residues have long charged side chains. Of particular interest, however, is the observed variability at six out of eight methionine residues (four of these side chains are variable in 1exr). The calmodulin methionine residues are dynamic in solution and this malleability is proposed to facilitate the binding of a diverse set of target proteins (O'Neil & DeGrado, 1990). These *PLOP* models could represent snapshots of the potential substates occupied by calmodulin in solution which are required for binding specificity and versatility.

3.2.3. Multiple crystal structures, NMR experiments: SUMO-conjugating enzyme (2.8 Å). Two regions of SUMO-conjugating enzyme (also called Ubc9) were determined by NMR experiments to be more mobile than the remainder of the protein (Liu, Yuan *et al.*, 1999). Comparison of multiple crystal structures at high resolution suggests flexibility in these same regions (Tong *et al.*, 1997). The flexible N-terminal region has been identified as the SUMO-binding site (Tatham *et al.*, 2003; Liu, Jin *et al.*, 1999), whereas the region near the C-terminus corresponds to the binding site of the target proteins (Bernier-Villamor *et al.*, 2002). Ubc9 can recognize a variety of protein targets and it has been proposed that the variability of the binding interface aids in substrate-specific recognition.

We generated 38 structures for SUMO-conjugating enzyme starting from the 1a3s crystal structure, using X-ray reflection data resolved to 2.8 Å. The backbone deviations are largest in the loop consisting of residues 32–36, with r.m.s.d.s of 1–3 Å relative to 1a3s. There are no crystal contacts restricting the conformational flexibility of this loop; thus, *PLOP* modeling is able to sample large variations in loop structures to provide alternative low-energy conformations in the absence of well resolved electron density in this region. Half of the 149 nonglycine residues of Ubc9 are modeled in different conformations in the *PLOP* models relative to 1a3s. Variable side-chain conformations are concentrated in regions in which variability has been observed across multiple crystal structures (Tong *et al.*, 1997) and NMR experiments (Liu, Yuan *et al.*, 1999); *i.e.* 67% of residues 32–36 and 121–146 exhibit different conformations.

Although there are relatively small variations in the backbone coordinates along the active-site cleft and protein-binding interfaces (∼0.2–0.3 Å r.m.s.d.), there are significant variations in the side-chain conformations in these regions. The target protein-binding surface (residues 85–92 and 123–143) shows 57% variability, while the SUMO-binding surface (residues 10–27) exhibits variability in 53% of the side-chain conformations. This conformational variability suggests that using a single Ubc9 structural target in drug–protein or protein–protein interactions may limit the reliability of results from high-throughput docking. Flexible fitting approaches to docking could be implemented by using multiple high-quality *PLOP* models that depict the range of conformations that is observed at the interfaces.

4. Discussion

4.1. Role of torsion-angle sampling in structure refinement

X-ray structure-refinement programs traditionally use target functions to optimize the agreement of an atomic model both with observed X-ray diffraction data and *a priori* chemical information. Whereas efficient algorithms exist for local optimization of the target function, the problem of locating the global minimum remains challenging owing to the high dimensionality of the search space. In 1987, simulated annealing with molecular dynamics (SA/MD) was adapted for X-ray structure refinement (Brünger *et al.*, 1987); SA/MD explores conformational space more extensively than local minimization methods. In principle, simulated annealing identifies the global minimum of the target function. However, in practice difficulties exist in locating the global minimum for complex systems in a finite period of time using realistic annealing schedules. An alternative to SA/MD is torsion-angle sampling, which has recently been introduced into X-ray structure refinement (DePristo *et al.*, 2004, 2005; Terwilliger *et al.*, 2007), in which backbone dihedral angle and side-chain rotamer libraries are used to sample many conformations within a macromolecule and conformations are scored by their fit to an experimental electron-density map. A torsion-angle sampling and rebuilding algorithm with an all-atom force field, as implemented in the protein-folding program *Rosetta*, has recently been shown to be able to provide *ab initio* high-quality initial structures for protein crystallographic refinement (Qian *et al.*, 2007).

A primary advantage of torsion-angle enumeration is that sampling is not directly affected by the roughness of the target function, so alternative low-energy conformations that might be separated from the initial structure by large energy barriers can be explored systematically. However, concerted changes in side-chain or backbone conformations are difficult to model

unless all the atomic coordinates involved in the changes are optimized simultaneously. Discrete torsion-angle sampling does not depend on the initial conformation for the segment of residues that are under investigation. For a local region, the torsion-angle conformational space can be sampled almost exhaustively, although the conformation of the remainder of the structure directly determines the quality of the selected local candidates. Torsion-angle sampling scales exponentially with the number of degrees of freedom. In *PLOP*, this exponential scaling is tempered by an adaptive build-up procedure and by using clustering as well as screening techniques. Both efficiency and accuracy are achieved *via* the deployment of a hierarchy of scoring functions; rapid screening functions are used to eliminate a large number of high-energy conformations in the early stages, ultimately yielding a relatively small number of candidates whose energies are evaluated *via* minimization of an all-atom molecular-mechanics energy function with continuum solvent model. Even so, in practice, ∼50 degrees of freedom or 13 residues is currently the upper limit of the size of segment that can be modeled with *PLOP* (Jacobson *et al.*, 2004; Zhu *et al.*, 2006). Moreover, owing to the discretized nature of the torsion-angle libraries (resolution of 5–10°), the sampled conformations require an additional step of local optimization to the nearest energy minimum.

There is no one-size-fits-all approach to X-ray structure refinement. SA/MD is suitable for refining entire structures which are far from the global minima. SA/TLS and NMA are well suited to exploring collective anisotropic thermal motions of the macromolecules and refining modest-resolution crystallographic structures. Torsion-angle sampling has advantages in refining parts of structures (such as loops) assuming that the coordinates in the rest of the structure are almost correct. The challenge of effectively and explicitly modeling structural variability in X-ray structure refinement, in which the starting structure is nearly correct, is best addressed by torsion-angle sampling, which gives the most aggressive exploration in local regions. We developed our automated iterative procedure to take advantage of (i) the systematic yet rapid enumeration afforded by hierarchical torsion-angle sampling guided by an accurate modern effective potential with (ii) filtering using fitting to the real-space electron-density map and (iii) minimization of a target function containing chemical and reciprocal-space X-ray energy terms that would optimize all atomic coordinates yet retain well refined regions of the model. Owing to the inherently local nature of torsion-angle sampling, we cycle through the sequence of the macromolecule in order to capture structural variability that may exist throughout the model and that is consistent with experimental X-ray reflection intensity data.

4.2. Distinguishing physical conformational heterogeneity from uncertainty in atomic positions

Multiple sources can contribute to conformational variability among structural models refined from the same crystallographic data set. Frequently, structural differences reflect uncertainty in the atomic positions which are associated with the limited resolution of the experimental data and/or with inadequate assumptions in refinement protocols. Conformational heterogeneity arising from protein motions is often absorbed into the *B* factors and interpreted formally as positional uncertainty when the refinement is performed using a single-conformer isotropic *B*-factor model. Isotropic *B*-factor models assume uncorrelated atomic harmonic fluctuations that are described by single-particle isotropic Gaussian functions. This assumption leads to ambiguities in the location of atomic positions in cases when the actual distributions of atomic positions are anisotropic and/or multimodal.

In our single-conformer isotropic *B*-factor models, the distance cutoff criterion of 1 Å identifies alternative conformations that intuitively would be considered to be 'different' when observed in a molecular viewer. Our proposed criterion to distinguish between physical heterogeneity from positional uncertainty is based on ensemble models and considers the electron density that would surround the atomic coordinates given the magnitude of the corresponding *B* factors. If we imagine constructing an electron-density map from the atomic coordinates and *B* factors, the larger the *B* factors, the larger the associated envelope of density and the larger are the atomic displacements that are required to achieve an assignment of 'heterogeneity' over 'uncertainty'.

We suggest that the requirement that the electron-density map corresponding to the PDB structure clearly shows alternative conformations in order to associate model variability with true heterogeneity may be unduly conservative. For each protein studied, the ensemble calculations demonstrate a slight improvement in the $R_{\rm free}$ value relative to the PDB structure, which suggests that the ensemble models have characteristics that are at least comparable in quality to the highest quality single-conformer models (see Table 4). Moreover, the automated *PLOP*/*CNS* protocol for generating ensemble models from high-quality single-conformer models demonstrates reasonably good agreement with the location and extent of side-chain variability in the manually curated high-resolution calmodulin PDB model 1exr. It should also be noted that while local errors can still be present even when the global $R_{\rm free}$ measure is improved, the lack of multiple density envelopes for variable residues in some cases reflects limitations in the maps themselves. It is well known that the model can bias the features of the electron density and it is not unreasonable to expect that different combinations of maps and phases could reveal alternative conformations. The construction of simulated-annealing OMIT maps would be one way to explore these effects, but we have not pursued this.

The identification of signatures leading to correct interpretations of the physical underpinnings of modeled structural variability would benefit the crystallographic community as well as all those relying on structures to develop scientific hypotheses. It is therefore important to undertake further work to investigate general methods to reliably distinguish between positional uncertainty and conformational heterogeneity.

4.3. Utility of explicit representation of structural variability

In protein crystallography it is standard procedure to represent the conformation of a protein as a single structure, unless there is strong evidence in the electron-density maps for the inclusion of alternative conformations in the model. However, as pointed out by Furnham *et al.* (2006), this convention unfortunately gives little indication of either the accuracy or the conformational heterogeneity in the crystal structure. These authors suggest that an ensemble of models would be a more suitable representation of a protein and that the range of structures in the ensemble represents the range of structures that should be considered by any user of the structural information. Until the recent application of torsion-angle sampling to the protein structure-refinement problem, it has been difficult to generate in an automated way a diverse ensemble of high-quality models which fit the X-ray data well. With programs such as *RAPPER* and *PLOP* now available, it is much easier to generate an ensemble of models to represent a protein structure and to explore the advantages of an ensemble representation of the structure.

One motivation for modeling structural variability is to explore the relationship between protein flexibility and biological function. Conformational variability at binding sites, hinge regions and at interfaces between domains often has functional relevance and this knowledge can be important for subsequent modeling research and experimental design (Gutteridge & Thornton, 2005; Gerstein & Echols, 2004; Rajamani *et al.*, 2004; Karplus *et al.*, 2005). A single-conformer representation of a macromolecular X-ray crystal structure reflects the dominant state of the system. However, this is only a partial picture that overlooks the true structural heterogeneity of the system, as well as the uncertainty in the atomic coordinates. Crystallographic models based on conformational ensembles (Gros *et al.*, 1990; Wilson & Brunger, 2000; Schiffer & Hermans, 2003), such as those explored in this study, provide a representation of the underlying variability within the macromolecular structure that can be used as an aid to help understand the relationship between protein conformational flexibility and mechanism. Ensembles can be used to highlight the range of conformations that should be taken into account in any subsequent analysis. For side chains displaying physical heterogeneity, one can ensure that emergent hypotheses are consistent with the presence of all high-quality conformations. Conversely, ensembles can also be used to derive estimates of model uncertainty, which should be considered when mechanisms are proposed that are based on the details of short-range atomic interactions. In addition, ensemble representations of a macromolecule can provide a structural rationale for interpreting unanticipated experimental results that are difficult to rationalize using single-conformer models.

The refinement of ensembles of single-conformer models which individually and/or collectively fit the crystallographic data well particularly benefit virtual screening and molecular-docking applications. With minimal flexibility at the binding site, rigid-body docking can succeed when softened potentials describe the interactions between the subunits or when alternative rotamer states within the binding pocket are sampled. Protein–ligand and protein–protein docking simulations demonstrate increased predictive power when they allow flexibility of the subunits (Bonvin, 2006; Ehrlich *et al.*, 2005; Halperin *et al.*, 2002; Sherman *et al.*, 2006). Receptor conformational flexibility can also be accounted for by using multiple protein structures obtained in a variety of ways: from NMR ensembles, from multiple X-ray crystallographic structures and from modeling. Distributions of conformations can be used to select multiple discrete targets or to construct a composite receptor target which is then used in docking studies (Damm & Carlson, 2007; Huang & Zou, 2007). By incorporating structural variability explicitly and docking against an ensemble of structures, more robust virtual screening protocols become possible and this may also lead to improved protocols for modeling-induced fit effects.

5. Conclusions

Structural biology provides a molecular perspective for understanding biological phenomena based on the analysis of the three-dimensional structures of proteins, nucleic acids and other macromolecules. The primary source of structural information at the atomic level is crystallography. Even though most practitioners in the field understand that a single PDB structure represents some sort of average structure, the atomic positions are almost sacred when observed in a molecular viewer. In a recent letter, Furnham *et al.* (2006) proposed that the representation of a macromolecular crystal structure as an ensemble of models is a more suitable representation, for which there is a precedent in the way NMR structures are represented, and that such a representation can improve our understanding of the relationship between structure and function. Generating ensembles of models that fit the X-ray data well provides a direct measure of the structural variability resulting from a combination of factors involving conformational heterogeneity and model uncertainty. However, generating ensembles of high-quality models which fit the X-ray data is a nontrivial task. Automated refinement methods are needed to generate ensembles of models and the most obvious approach, simulated-annealing molecular dynamics, is not particularly well suited to the task. With the application of torsion-angle sampling methods to X-ray refinement such as those contained in the programs *RAPPER* and *PLOP*, it becomes much easier to generate ensembles of models which individually and together fit the X-ray data well.

The automated protocol described here combines aggressive local exploration of torsion-angle sampling with scoring using a modern physics-based effective potential to generate low-energy candidates that are then filtered by real-space X-ray criteria (electron density) and minimized using a reciprocal-space X-ray target function. This protocol generates models that fit the X-ray data as well as or better than the original deposited PDB structures for the five macromolecules reported here which serve as test cases. The ensemble of *PLOP* structures show significant variability relative to the

PDB structure, with backbone r.m.s.d.s from 0.08 to 0.6 Å and 25–50% of side chains in alternate conformations. Using HIV-1 protease and SUMO-conjugating enzyme as examples, we demonstrated how the modeled structural variability captured the variability that is observed in multiple crystal structures and in NMR ensembles and, for calmodulin, variability observed at high resolution which was modeled in the structure submitted to the PDB by reporting multiple occupancies.

An approximate approach based on ensemble refinement is proposed to differentiate between the variability arising from physical heterogeneity and that which reflects positional uncertainty. From 25% to 50% of the variability that is modeled by our protocol can be attributed to structural heterogeneity associated with discrete conformers for which the spread in conformer positions exceeds the positional uncertainty as estimated by the atomic *B* factors. However, the fraction of residues which exhibit variability and for which the PDB-phased electron-density map clearly shows multiple occupancies is generally smaller than 25%. We have also shown that a more physically realistic description of the effective potential in *PLOP* is able to generate higher quality conformations compared with a standard potential commonly used for X-ray structure refinement which excludes electrostatic interactions. Arguably, the role of the potential function in generating high-quality conformations becomes even more important when there are fewer experimental data and for modeling multiple conformations with unequal population distributions where the dominant conformer may mask minor ones in the electron-density map.

In summary, the automated iterative strategy described in this work based on torsion-angle sampling in combination with filtering and optimization of the fit to the X-ray data is a powerful tool by which multiple high-quality models may be generated and refined in parallel. These models can provide explicit representation of the structural variability that exists within macromolecular complexes and may be more reliable than a single-conformer model when used as aids to understand enzyme mechanisms or molecular recognition or as three-dimensional scaffolds in structure-based drug design.

This work was supported in part by grants (GM-30580 to RML and GM-52018 to RAF) from the National Institutes of Health and by NIH NRSA fellowships to Daniel Himmel (F32 AI060310) and Zhiyong Zhou (R90 DK071502). We thank Jeff Bell for stimulating discussions and for kindly reviewing the manuscript prior to publication and Matt Jacobson for providing the *PLOP* code and documentation.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (2002). *Molecular Biology of the Cell.* New York: Garland Science.
Andrec, M., Harano, Y., Jacobson, M. P., Friesner, R. A. & Levy, R. M. (2002). *J. Struct. Funct. Genomics*, **2**, 103–111.
Bakker, P. I. de, DePristo, M. A., Burke, D. F. & Blundell, T. L. (2003). *Proteins*, **51**, 21–40.
Berman, H., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. (2002). *Cell*, **108**, 345–356.
Bonvin, A. M. (2006). *Curr. Opin. Struct. Biol.* **16**, 194–200.
Brunger, A. T. & Adams, P. D. (2002). *Acc. Chem. Res.* **35**, 404–412.
Brünger, 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., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Brünger, A. T., Adams, P. D. & Rice, L. M. (1998). *Curr. Opin. Struct. Biol.* **8**, 606–611.
Brunger, A. T., Adams, P. D. & Rice, L. M. (1999). *Prog. Biophys. Mol. Biol.* **72**, 135–155.
Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Burling, F. T. & Brünger, A. T. (1994). *Isr. J. Chem.* **34**, 165–175.
Chen, X., Poon, B. K., Dousis, A., Wang, Q. & Ma, J. (2007). *Structure*, **15**, 955–962.
Cho, S. J., Lee, M. G., Yang, J. K., Lee, J. Y., Song, H. K. & Suh, S. W. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 8932–8937.
Damm, K. L. & Carlson, H. A. (2007). *J. Am. Chem. Soc.* **129**, 8225–8235.
DePristo, M. A., de Bakker, P. I. & Blundell, T. L. (2004). *Structure*, **12**, 831–838.
DePristo, M. A., de Bakker, P. I., Johnson, R. J. & Blundell, T. L. (2005). *Structure*, **13**, 1311–1319.
Ehrlich, L. P., Nilges, M. & Wade, R. C. (2005). *Proteins*, **58**, 126–133.
Erickson, J. & Kempf, D. (1994). *Arch. Virol. Suppl.* **9**, 19–29.
Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. (2006). *Nature Struct. Mol. Biol.* **13**, 184–185.
Gallicchio, E., Zhang, L. Y. & Levy, R. M. (2002). *J. Comput. Chem.* **23**, 517–529.
Gerstein, M. & Echols, N. (2004). *Curr. Opin. Chem. Biol.* **8**, 14–19.
Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998). *J. Phys. Chem. B*, **102**, 10983–10990.
Gros, P., van Gunsteren, W. F. & Hol, W. G. (1990). *Science*, **249**, 1149–1152.
Gutteridge, A. & Thornton, J. (2005). *J. Mol. Biol.* **346**, 21–28.
Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). *Proteins*, **47**, 409–443.
Huang, S. Y. & Zou, X. (2007). *Proteins*, **66**, 399–421.
Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. (2002). *J. Mol. Biol.* **320**, 597–608.
Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. (2004). *Proteins*, **55**, 351–367.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). *J. Am. Chem. Soc.* **188**, 11225–11236.
Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). *J. Phys. Chem. B*, **105**, 6474–6487.
Karplus, M., Gao, Y. Q., Ma, J., van der Vaart, A. & Yang, W. (2005). *Philos. Trans. R. Soc. Ser. A*, **363**, 331–355.
Kidera, A. & Go, N. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 3718–3722.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Koshland, D. E. Jr (1963). *Science*, **142**, 1533–1541.
Kuriyan, J., Osapay, K., Burley, S. K., Brünger, A. T., Henrickson, W. A. & Karplus, M. (1991). *Proteins*, **10**, 340–358.
Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. (1986). *J. Mol. Biol.* **190**, 227–254.
Kurt, N., Scott, W. R., Schiffer, C. A. & Haliloglu, T. (2003). *Proteins*, **51**, 409–422.

Levin, E. J., Kondrashov, D. A., Wesenberg, G. E. & Phillips, G. N. Jr (2007). *Structure*, **15**, 1040–1052.
Liu, Q., Jin, C., Liao, X., Shen, Z., Chen, D. J. & Chen, Y. (1999). *J. Biol. Chem.* **274**, 16979–16987.
Liu, Q., Yuan, Y. C., Shen, B., Chen, D. J. & Chen, Y. (1999). *Biochemistry*, **38**, 1415–1425.
Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., Clawson, L., Selk, L., Kent, S. B. & Wlodawer, A. (1989). *Science*, **246**, 1149–1152.
Moulinier, L., Case, D. A. & Simonson, T. (2003). *Acta Cryst.* D**59**, 2094–2103.
O'Neil, K. T. & DeGrado, W. F. (1990). *Trends Biochem. Sci.* **15**, 59–64.
Poon, B. K., Chen, X., Lu, M., Vyas, N. K., Quiocho, F. A., Wang, Q. & Ma, J. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 7869–7874.
Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.
Rader, S. D. & Agard, D. A. (1997). *Protein Sci.* **6**, 1375–1386.
Rajamani, D., Thiel, S., Vajda, S. & Camacho, C. J. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 11287–11292.
Rejto, P. A. & Freer, S. T. (1996). *Prog. Biophys. Mol. Biol.* **66**, 167–196.
Schaal, W., Karlsson, A., Ahlsen, G., Lindberg, J., Andersson, H. O., Danielson, U. H., Classon, B., Unge, T., Samuelsson, B., Hulten, J., Hallberg, A. & Karlen, A. (2001). *J. Med. Chem.* **44**, 155–169.
Schiffer, C. & Hermans, J. (2003). *Methods Enzymol.* **374**, 412–461.
Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. (2006). *J. Med. Chem.* **49**, 534–553.
Smith, J. L., Hendrickson, W. A., Honzatko, R. B. & Sheriff, S. (1986). *Biochemistry*, **25**, 5018–5027.
Stec, B., Zhou, R. & Teeter, M. M. (1995). *Acta Cryst.* D**51**, 663–681.
Tatham, M. H., Kim, S., Yu, B., Jaffray, E., Song, J., Zheng, J., Rodriguez, M. S., Hay, R. T. & Chen, Y. (2003). *Biochemistry*, **42**, 9959–9969.
Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Adams, P. D., Moriarty, N. W., Zwart, P., Read, R. J., Turk, D. & Hung, L.-W. (2007). *Acta Cryst.* D**63**, 597–610.
Tong, H., Hateboer, G., Perrakis, A., Bernards, R. & Sixma, T. K. (1997). *J. Biol. Chem.* **272**, 21381–21387.
Westhof, E., Chevrier, B., Gallion, S. L., Weiner, P. K. & Levy, R. M. (1986). *J. Mol. Biol.* **191**, 699–712.
Willis, B. T. & Pryor, A. W. (1975). *Thermal Vibrations in Crystallography*. Cambridge University Press.
Wilson, M. A. & Brunger, A. T. (2000). *J. Mol. Biol.* **301**, 1237–1256.
Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.
Yu, B., Blaber, M., Gronenborn, A. M., Clore, G. M. & Caspar, D. L. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 103–108.
Zhu, K., Pincus, D. L., Zhao, S. & Friesner, R. A. (2006). *Proteins*, **65**, 438–452.
Zoete, V., Michielin, O. & Karplus, M. (2002). *J. Mol. Biol.* **315**, 21–52.