Supplementary Material

Exploring structural diversity in X-ray crystallographic refinement 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*]

1. Detailed experimental procedures

1.1. Protein Structures and Reflection Data

Atomic coordinates and structure factors (including test 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., 2003; Berman et al., 2000). RAPPER-generated models for HIV-1 protease were obtained from http://www-cryst.bioc.cam.ac.uk/rapper/ (DePristo et al., 2004). For the 37 residues in 1exr which had multiple conformations, only coordinates from the "B" conformation were included in our initial structures; the occupancy values for these coordinates were reset to 1.0. The occupancy and B-factors for loop atoms in residues 32-36 in 1a3s were reset to 1.0 and 100, respectively.

1.2. Initial structures generated by simulated annealing with molecular dynamics

To reduce the impact of model bias and allow a more thorough exploration of conformational space and increased structural diversity among the final models, for each macromolecule, ten different initial structures were generated by simulated annealing starting from the PDB structure. Simulated annealing slow-cooling molecular dynamics (SA/MD) was used in CNS (Brunger *et al.*, 1998; Brunger & Adams, 2002). Cartesian MD was used for 1g35, 1exr, and 1ew4 and torsion angle MD was used for the low-resolution structures 1a3s and 9ilb. The initial temperature was 1000 K, and the temperature dropped by 25 K after each dynamics cycle. The parameters for solvent treatment and w_a were optimized in CNS and were applied to all calculations. To provide additional variation in the SA/MD structures, the optimal crystallographic weighting term, w_a, in the SA/MD simulations was reduced ten-fold.

1.3. Iterative X-ray 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 eleven initial structures (ten SA/MD structures and the PDB model) to generate a total of 66 final PLOP structures for each protein. Regions with low RSCCs were identified as start sites since larger conformational changes may be tolerated in these regions and, due to the model bias that is inherent in X-ray refinement, including differences earlier on in the modeling will yield larger differences in the phasing and generate more divergent final structures. 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. For the 1g35 homodimer, the target window alternated between residues on chain A and B. Sixty, 47, 46, 47, and 32 cycles were required for 1g35, 1a3s, 1exr, 9ilb and 1ew4 structures respectively. 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. All PLOP and CNS calculations were performed on a cluster of 2.1 GHz AMD Athlon processors.

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 OPLS-AA/SGBNP effective energy function. Each execution of PLOP is composed of four stages: buildup, closure, clustering, and scoring. Crystal unit cells are explicitly reconstructed by using the dimensions and space group reported in the PDB files. The simulation system consists of one asymmetric unit (which may contain more than one protein chain) and all atoms from other, surrounding symmetric units that are within 30 Å of any atom in the primary asymmetric unit. Every copy of the asymmetric unit is identical at every stage of the calculation; that is, if the conformation of a side chain is modified, all copies of the side chain

in the simulation system are updated simultaneously. In the following subsections we briefly review the elements of a single loop prediction; see (Jacobson et al., 2004) for more details.

Build up: "Buildup" refers to the generation of an initial set of loop conformations that will be passed on to the subsequent three stages. The cornerstone of the loop-sampling methodology is dihedral angle search, which is conducted via "rotamer libraries" for backbone dihedral angles (i.e., discretized versions of the well-known Ramachandran plot). Dihedral angle libraries were obtained from a large (>500), nonredundant database of high-resolution (<2 Å) protein crystal structures. Every backbone dihedral angle was recorded and then binned every 10°; every (ϕ,ψ) combination that appeared more than five times in the database was included in the backbone library. The resultant library, at 5° resolution, contains 747 (ϕ,ψ) combinations for Gly, 215 for Pro, and 866 for all other residue types. The high resolution of the libraries ensures that discretization error does not fundamentally limit the achievable accuracy of a prediction. This stage generates between 2^n and 10^6 loops, where n is the number of residues in the loop. The overlap factor parameter ofac is defined as the minimum permitted ratio of the interatomic distance over the sum of the Lennard-Jones radii of the atoms of interest. For this work, ofac is set to 0.5 to allow more overlap between atoms which, in effect, allows for more loop conformations to be investigated which would otherwise be eliminated due to steric clashes.

Loop closure: The buildup procedure continues independently from both sides of the loop up to the Cα atom on the closure residue. Then, all pairs of loop fragments built from the two sides which have the closure Cα atoms within 0.5 Å of each other are identified. Averaging the positions of the closure Cα atom from the two fragments and adding the Cα, Hα, and side-chain atoms to the closure residue using standard geometries generates a closed loop. Additional screens are employed at this stage to avoid retaining loops with unacceptable geometries at the site of the closure. The screens include: acceptable N-Cα-C angles and dihedral angles on the closure residue, absence of steric clashes, and sufficient volume for the side chain.

Clustering: Even with screening techniques, the buildup and closure stages can generate tens of thousands of loops. Optimizing and scoring every candidate with a high-resolution energy model is prohibitively expensive. Furthermore, many of the loops would ultimately optimize into similar structures. Therefore, a K-means clustering algorithm is used to select a representative set of structures for side chain optimization and energetic scoring.

Side chain optimization and loop minimization: Sampling is accomplished by using a highly detailed (10° resolution) rotamer library constructed by Xiang and Honig from a database of 297 proteins (Xiang & Honig, 2001). All side chains are initially built onto the fixed backbone in a random rotamer state, and then each side chain in the protein is optimized

one at a time, holding the others fixed. The procedure is iterated until no side chains change rotamer states. After convergence is achieved, the complete loop (side chains plus backbone) is energy minimized in Cartesian coordinates to remove any remaining clashes, and to obtain a reliable estimate of the energy that can be used to compare fairly the energies of the diverse representative cluster members generated in the previous steps, using a rapid, novel multiscale minimization algorithm (Jacobson and Friesner, unpublished results).

Energy functions: The PLOP algorithm relies on an all-atom energy function to discriminate between proposed conformations. Specifically, we use the all-atom force field, OPLS 2000 (Jorgensen et al., 1996; Kaminski et al., 2001), and an implicit solvent model. OPLS-AA models the intramolecular bonded and nonbonded interactions of the protein. The bonded interactions include bond stretching, angle bending, and torsion dihedral interactions. The nonbonded interactions are the van der Waals dispersion interactions (modeled with the 6-12 Lennard-Jones potential function) and the direct Coulomb interactions between the formal and partial charges on each atom. The implicit solvent model was a surface implementation of the generalized Born model (Ghosh et al., 1998) with a non-polar hydration free energy estimator (Gallicchio et al., 2002). To test the effectiveness of the OPLS-AA/SGBNP potential in generating high-quality structures, we also ran PLOP without an implicit solvent model and without non-bonded electrostatics interactions, to mimic the energy terms that are used routinely in CNS.

Step 2: Filtering PLOP candidates

For each PLOP candidate that was within 20 kcal/mol of the lowest energy model (in practice between 5 and 30 candidates), 2Fo-Fc (3Fo-2Fc for the low-resolution structures 1a3s and 9ilb) and Fc maps were generated in CNS version 1.1 (Brunger et al., 1998). The mean real-space correlation coefficient for the targeted 5-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 re-modeled 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 (2 cycles of 10 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. The cycles were repeated until each residue in the protein had been sampled by PLOP at least once. Ligands and water molecules were free to move for SA/MD and minimization in CNS, but were frozen during PLOP torsion angle sampling.

1.4. Filtering the ensemble of PLOP models

The ensemble of 66 PLOP models were 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 RSCC (resRSCC) for all PLOP candidates that were variable at a given residue were evaluated in Mapman from the corresponding 2Fo-Fc (3Fo-2Fc for the low-resolution structures 9ilb and 1a3s) map and Fc map generated by CNS. Absolute and relative degradations in resRSCCs relative to the PDB structure were considered. Specifically, all PLOP candidates, *j*, that exhibit variability at residue *i* were removed from the ensemble, if:

$$resRSCC(PDB,i) - resRSCC(PLOP_j,i) > 0.5(1 - avg5resRSCC(PDB,i))$$

or

$$\begin{cases} resRSCC(PDB,i) - resRSCC(PLOP_j,i) > 0.03 \\ and \\ resRSCC(PDB,i) - resRSCC(PLOP_j,i) > 0.25(1 - avg5resRSCC(PDB,i)) \end{cases}$$

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. PLOP candidates were eliminated and variable residues were scanned through until each variable side chain in the filtered ensemble, other than the false positives, had at least one PLOP candidate with a comparable or better RSCC for that residue relative to the PDB structure.

1.5. 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 for further optimization. Where multiple sets of structures fit this criteria, the subset with the lowest average R value was selected. Occupancy values for each of the structures in the ensemble were optimized by sampling 3,500 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. The fractional occupancies and ensemble R and R^{free} values were reported for the combination of fractional occupancies that gave the lowest R^{free} value.

References

Berman, H., Henrick, K. & Nakamura, H. (2003). *Nat 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.

Brunger, A. T. & Adams, P. D. (2002). *Acc Chem Res* **35**, 404-412.

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., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Crystallogr D Biol Crystallogr* **54 (Pt 5)**, 905-921.

Cho, S. J., Lee, M. G., Yang, J. K., Lee, J. Y., Song, H. K. & Suh, S. W. (2000). *Proc Natl Acad Sci U S A* **97**, 8932-8937.

DePristo, M. A., de Bakker, P. I. & Blundell, T. L. (2004). *Structure* **12**, 831-838.

Gallicchio, E., Zhang, L. Y. & Levy, R. M. (2002). *J Comput Chem* **23**, 517-529.

Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998). *J Phys Chem B* **102**, 10983-10990.

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 (1991). *Acta Crystallogr A* **47 (Pt 2)**, 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.

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Crystallogr D Biol Crystallogr* **52**, 826-828.

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.

Tong, H., Hateboer, G., Perrakis, A., Bernards, R. & Sixma, T. K. (1997). *J Biol Chem* **272**, 21381-21387.

Wilson, M. A. & Brunger, A. T. (2000). *J Mol Biol* **301**, 1237-1256.

Xiang, Z. & Honig, B. (2001). *J Mol Biol* **311**, 421-430.

Yu, B., Blaber, M., Gronenborn, A. M., Clore, G. M. & Caspar, D. L. (1999). *Proc Natl Acad Sci U S A* **96**, 103-108.