Structure
Article

Ensemble Refinement of Protein Crystal Structures: Validation and Application

Elena J. Levin,[1,2] Dmitry A. Kondrashov,[1] Gary E. Wesenberg,[2] and George N. Phillips, Jr.[2,3,*]
[1]Department of Biochemistry
[2]Center for Eukaryotic Structural Genomics
[3]Departments of Biochemistry and Computer Science
University of Wisconsin, Madison, Madison, WI 53706, USA
*Correspondence: phillips@biochem.wisc.edu
DOI 10.1016/j.str.2007.06.019

SUMMARY

X-ray crystallography typically uses a single set of coordinates and B factors to describe macromolecular conformations. Refinement of multiple copies of the entire structure has been previously used in specific cases as an alternative means of representing structural flexibility. Here, we systematically validate this method by using simulated diffraction data, and we find that ensemble refinement produces better representations of the distributions of atomic positions in the simulated structures than single-conformer refinements. Comparison of principal components calculated from the refined ensembles and simulations shows that concerted motions are captured locally, but that correlations dissipate over long distances. Ensemble refinement is also used on 50 experimental structures of varying resolution and leads to decreases in R_{free} values, implying that improvements in the representation of flexibility observed for the simulated structures may apply to real structures. These gains are essentially independent of resolution or data-to-parameter ratio, suggesting that even structures at moderate resolution can benefit from ensemble refinement.

INTRODUCTION

X-ray crystallography has yielded a wealth of macromolecular structures, and atomic positions are being determined to ever-increasing precision. Static structures, however, tell only a part of the story of biochemical function. Diverse tasks require conformational flexibility, including many enzymatic reactions, the regulation of access of the substrate to buried active sites, and signal transduction via ligand or protein binding. Accurate measurement of the dynamic properties of proteins is central to understanding the relationship between structure and function. Experimental techniques have made enormous strides in this area, but detailed characterization of molecular conformational changes remains both laborious and limited in applicability. NMR spectroscopy can be used to determine both the structure and the dynamics of proteins (Lindorff-Larsen et al., 2005); mass spectrometry coupled with hydrogen/deuterium exchange and proteolysis has been used to determine changes in the relative solvent accessibility of amide hydrogens (Lanman and Prevelige, 2004), and single-molecule experiments with optical trapping have resulted in spectacular observations of the motion of motor proteins (Abbondanzieri et al., 2005). X-ray diffraction can be used to probe the time evolution of electron density in crystals (Moffat, 2001), but its application is limited to reactions that can be triggered by light or trapped by clever manipulations.

Classical crystallography is also a source of information about conformational flexibility, despite the confines of crystal packing. In addition to conformational changes observed between structures determined with different ligands or under varying conditions, local flexibility data can be observed in a single crystal data set. These small fluctuations tend to capture both the directionality and the correlation structure of large conformational changes, as demonstrated by the applications of normal-mode analysis for prediction of functionally important transitions (Cheng et al., 2006; Ma and Karplus, 1997; Wang et al., 2005). The standard crystallographic model uses Debye-Waller factors to account for fluctuations about the mean structure, which describe the motion of an individual atom as an isotropic Gaussian distribution of displacements about an average position. For structures solved at ultra-high resolutions (<1.2 Å), at which a much larger number of independent observations are available, the isotropic temperature factor can be replaced with anisotropic displacement parameters that allow for varying magnitudes of atomic motion in different directions (Willis and Pryor, 1975). Another approach commonly used in addition to individual temperature factors involves dividing the protein into a set of rigid-body domains independently undergoing translational, librational, and coupled translational-librational vibrations (TLS) (Schomaker and Trueblood, 1968). Although limited to rigid-body motion, the TLS model has the advantage of requiring relatively few parameters. A fourth method, the use of alternate side chains, is employed frequently in high-resolution

structures in which two or more different conformers for a flexible side chain are clearly visible in the electron density. Each of the conformers is given a fractional occupancy, with the combined values typically adding to one, and interaction terms between atoms in the different alternate conformers are excluded from the potential energy function to allow them to coexist in the model.

Despite its widespread use, there are well-known limitations to the single-conformer model that relies on the Debye-Waller factor as the sole parameter for describing conformational variation. In addition to suggesting a misleading degree of accuracy (DePristo et al., 2004), conventional refinement techniques were demonstrated by Kuriyan et al. (1986) to lead to temperature factors that systematically underestimate root-mean-square (rms) deviations from the average coordinates, even when no restraints between neighboring atoms are used. Although temperature factors can model the magnitude and sometimes the direction of protein motion, they are limited to Gaussian distributions for describing the probability density function of each atom's position and cannot accurately capture anharmonic or multimodal motion. Furthermore, temperature factors provide no information on correlations between displacements of different atoms. A growing body of both theoretical (Elber and Karplus, 1987; Garcia et al., 1997b) and experimental (Ansari et al., 1985; Eisenmesser et al., 2005; Volkman et al., 2001) studies has provided evidence not only that anharmonic motion constitutes a significant portion of a protein's overall dynamics, but that those motions may play a vital role in protein function as well. Our current understanding of the protein energy landscape suggests that structure is best described as an ensemble of hierarchical conformational substates in constant exchange with each other (Austin et al., 1975; Frauenfelder et al., 2001). As a consequence, a more informative way to model the dynamics present in a crystallized protein may be to represent the structure as a set of overlapping, noninteracting conformers that each account for a fraction of the total electron density.

The concept of ensemble refinement for X-ray crystal structures is over a decade old (Burling and Brunger, 1994; Kuriyan et al., 1991); however, only a small number of structures containing complete multiple conformers have been reported in the literature (Burling et al., 1996; Gill et al., 2002; Pellegrini et al., 1997; Wall et al., 1997; Wilson and Brunger, 2000), and fewer still have been deposited in the Protein Data Bank (PDB) (Rader and Agard, 1997). Most of these previously refined structures were of ultra-high resolution, had atypically high R_{free} values when refined as a conventional single-conformer model, or were known to exhibit high degrees of conformational disorder. Two factors that may have limited the use of ensemble models in the past were the prohibitive computational expense of performing simulated annealing on systems containing large numbers of atoms and the lack of high-resolution data sets with sufficient observation-to-parameter ratios. Both of these obstacles have been made surmountable by increases in computer-processor speeds and improvements in crystallization, data collection, and phasing techniques, providing a greater number of high-resolution structures.

For these reasons it is now both practical and appropriate to conduct a large-scale assessment of the accuracy and usefulness of ensemble refinement for extracting quantitative descriptions of protein motion from X-ray crystallographic data (Furnham et al., 2006). In this paper, we describe the application of an automated ensemble refinement protocol to a sample of 50 crystal structures with a variety of sizes, resolutions, and degrees of conformational flexibility, as well as to 3 sets of simulated crystallographic data generated from molecular dynamics (MD) simulations. The refinement procedure used is similar to that described in work by Wilson and Brunger (2000), in which each atom is given an individual temperature factor, all conformers are given equal fixed occupancies, and the initial separation of the conformers is achieved by torsion dynamics simulated annealing. Our results suggest that refinement with an ensemble of conformers can substantially reduce the R_{free} values and improve the estimation of the magnitude and anharmonicity of motions of protein X-ray structures.

RESULTS

Validation with Simulated Data

Simulations 1 ns in length were carried out for three proteins by using the PDB entries 1XMT, 1Q4R, and 1VJH as starting structures, from which coordinates were sampled once per picosecond for the second 500 ps of the trajectories. The 500 coordinate sets from each simulation were aligned to the original structure and used to calculate structure factors, which were then averaged to produce a single set of reflections. Conventional single-conformer models with isotropic temperature factors were fitted to the simulated data. These models were then used as starting structures for the automated refinement of 1-, 2-, 4-, 8-, and 16-conformer models against the structure factors calculated from the simulations, by using a combination of torsion angle simulated annealing (Rice and Brunger, 1994) and standard maximum likelihood refinement.

Table 1 summarizes the results of the refinement of the ensemble models against the simulated data. All three simulations show a dramatic decrease in R_{free} values and modest phase improvements after ensemble refinement, although the optimum number of conformers varies. While the minima in the R_{free} and phase residuals over different conformer numbers do not coincide exactly for each protein, they follow the same general pattern. Little to none of the improvement appears to be driven by an ability to more accurately recover the true average coordinates, however, as the distance between the mean coordinates of the model and the true mean calculated from the simulations varies only slightly as the number of conformers increases. Instead, the drop in the R_{free} value of the ensemble models appears to arise, in part, from an improvement in the estimation of the average magnitude of displacements from the mean structure, as seen in Figure 1. The one-conformer models seem to exhibit a