On the Role of the Crystal Environment in Determining Protein Side Chain Conformations

Matthew P. Jacobson and Richard A. Friesner

Department of Chemistry, Columbia University, New York, New York 10027

Zhexin Xiang and Barry Honig

Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10027

Abstract

The role of crystal packing in determining the observed conformations of amino acid side chains in protein crystals is investigated by 1) analysis of a database of proteins that have been crystallized in different unit cells (space group or unit cell dimensions) and 2) theoretical predictions of side chain conformations with the crystal environment explicitly represented. Both of these approaches indicate that the crystal environment plays an important role in stabilizing the observed conformations of *polar* side chains on the *surfaces* of proteins. Inclusion of the crystal environment permits a more sensitive measurement of the achievable accuracy of side chain prediction programs, when validating against structures obtained by x-ray crystallography. Our side chain prediction program uses an all-atom force field and a Generalized Born model of solvation and is thus capable of modeling simple packing effects (i.e., van der Waals interactions), electrostatic effects, and desolvation, which are all important mechanisms by which the crystal environment impacts observed side chain conformations. These results are also relevant to the understanding of changes in side chain conformation due to "induced fit"

in the context of ligand docking and protein-protein association, insofar as the results reveal how side chain conformations change in response to their local environment.

[Keywords: Side chain prediction; crystal packing; protein structure; implicit solvation; all-atom force fields]

INTRODUCTION

It is well established that the overall folds of proteins are generally very similar in solution and in the crystalline environment (e.g., Refs. [1] and [2]). However, details of the structure may differ, especially the conformations of loops and surface side chains. The role of crystal packing in determining observed side chain conformations is of interest for several reasons:

1. Protein structures solved by x-ray crystallography are used nearly exclusively to validate methods of predicting side chain conformations. That is, the conformations of side chains in crystals are used to judge whether a prediction is correct, although the primary goal of side chain prediction algorithms is to aid in modeling of proteins in solution, e.g., in the context of comparative protein modeling. Thus, it is of interest to distinguish the errors caused by neglect of crystal packing from errors due to the potential function and optimization algorithms employed.

2. Structural changes associated with mutation can be difficult to distinguish from structural changes associated with crystal packing if, e.g., a mutant crystallizes in a different space group than the native protein [3,4,6]. For example, are the side chain rearrangements around a single point mutation due to the mutation itself or due to variation in the crystal packing forces?

3. Changes in side chain conformation due to crystal packing are likely to be related to changes in side chain conformation due to "induced fit" in the context of ligand docking or protein-protein association. In both contexts, side chain conformations are

determined by a combination of intramolecular (the local environment of the side chain in the protein) and intermolecular (interactions with ligands or surrounding proteins in the crystal) forces.

In this article we present two contributions to the understanding of crystal packing effects on side chain conformations. First, we exploit the fact that a number of proteins have multiple crystal structures deposited in the Protein Data Bank (PDB) [5], and the side chain conformations can be observed to vary among the structures (e.g., Refs. [6] and [7]). Prior studies of this type have attempted to quantify side chain "flexibility" through statistical analysis of the variability observed for different residue types; the derived flexibility was used as a point of comparison with 1) the results of side chain prediction programs [8,9], 2) changes in side chain conformation associated with induced fit effects caused by ligand docking [10] and protein-protein association [11], and 3) changes in side chain conformation of conserved residues in homologous pairs [12]. The physical mechanism underlying the conformational changes has received less attention. In particular, the data sets employed in previous studies contained two classes of structural pairs 1) those which differ in the nature of the crystal unit cell (space group or substantial changes in unit cell dimensions) and 2) those which do not, but which might differ in other environmental variables, such as pH or ionic strength. We demonstrate here that the variability of surface side chains is much larger in pairs of structures which differ in unit cell than pairs which do not; the discrepancy between the two types of pairs is a measure of the average effects of crystal packing.

Our second contribution is the development of a new side chain prediction program which can, as an option, explicitly represent the crystal environment. Although, in itself, the prediction of side chain conformations in the crystal environment is of limited utility to biological or pharmaceutical chemistry, studies of this type are valuable for several reasons. First, the effects of crystal packing on side chain conformations can be quantified by performing side chain prediction with and without crystal packing forces (i.e., those forces arising from interactions between asymmetric units in the crystal). These theoretical re-

sults complement the database studies, because variations in environmental conditions can be completely controlled, and variations in side chain and backbone conformation due to crystal packing can be studied independently. The two types of studies, however, provide remarkably similar results for the effects of crystal packing on side chain conformation. Second, with the explicit inclusion of the crystal environment, apparent errors in side chain prediction, when comparing to x-ray crystal structures, are greatly reduced and the remaining errors can be attributed solely to deficiencies in the potential function or search algorithm. The potential function that we employ, which is defined by an all-atom force field [13,14] and a Generalized Born model of solvation [15], performs quite well, and in fact is capable of reproducing changes in side chain conformation observed for proteins that crystallize in different space groups. Lastly, the ability to accurately reproduce side chain conformations in the crystal environment (and, especially, changes in conformation due to changes in crystal environment) leads to increased confidence in the ability of the methodology to reproduce conformational changes associated with other, more biologically relevant, modifications of the protein environment, such as those due to ligand docking.

RESULTS AND DISCUSSION

Database Analysis

Proteins which have been crystallized multiple times often have side chains in different rotameric states. This fact has been exploited by a number of groups to establish baseline side chain "flexibility" to compare with side chain prediction results or side chain conformational variation due to docking or mutation. The largest such study to date, carried out by Zhao *et al.* [9], included 123 pairs of structures from the PDB which contained chemically identical, uncomplexed proteins. The structures were screened to ensure that changes in backbone conformations among the different crystal structures were minimal (<0.5 Å RMSD), but the side chains, especially those on the surface, demonstrated substantial vari-

ability. For example, 90% of Ser residue pairs in the protein core had χ_1 side chain dihedral angles within 16.1° of each other, while the corresponding value for surface Ser residue pairs is 102.7°.

The changes in side chain conformation in different crystals of the same protein can be attributed to

1. Errors associated with structure determination from the diffraction pattern.

2. Changes in the crystallization conditions, which modify, e.g., the pH or ionic strength in the crystal.

3. Variation in the crystal packing forces.

The effect of the latter of these, which we wish to isolate, will only be significant if the arrangement of the proteins in the crystal varies, i.e., due to a different space group or substantially different unit cell dimensions. As mentioned in the Introduction, prior studies have not segregated pairs of structures with the same and different crystal unit cells, and we demonstrate here that pairs with different unit cells have substantially greater variability in side chain conformations.

Before turning to a statistical assessment of crystal packing effects on side chain conformation, we first provide an instructive example. Two crystal structures of human ubiquitin-conjugating enzyme Ubc9 were reported by Tong *et al.* [16]. The structure with PDB code 1u9a was determined in a monoclinic crystal with space group $P2_1$ (two proteins per unit cell), while that of 1u9b was determined in an orthorhombic crystal with space group $I222$ (eight proteins per unit cell).[1] The unit cells for each are represented at the top of Figure 1. The atoms in these space filling models are colored according to distinct protein chains, the

[1]The dimensions of the unit cell are as follows (edge lengths in Å and angles in degrees). 1u9a: $a = 52.0$, $b = 35.2$, $c = 58.1$, $\alpha = 90.0$, $\beta = 111.2$, $\gamma = 90.0$. 1u9b: $a = 35.4$, $b = 94.0$, $c = 115.9$, $\alpha = 90.0$, $\beta = 90.0$, $\gamma = 90.0$.

orientations of which are related to each other by space group symmetry operations. Note the prominent solvent channels in each crystal, as well as the extensive inter-chain contacts, including interdigitation of surface side chains. Typically, the solvent accessible surface area of proteins decreases by 20–40 % upon crystallization, with the precise number depending largely on protein size [18]. The different relative orientations among the protein chains in the two space groups cause different regions of the surfaces to be involved in inter-chain contacts in the two crystal structures.

The backbones of the two structures are very similar; the RMSD for the 158 C^{α} atoms is 0.44 Å [16]. The differences in the backbone are, as is typical, concentrated in the termini and surface loop residues. The most significant conformational changes between the two crystal structures, however, occur for certain surface side chains. The lower panels of Figure 1 provide a striking example of crystal packing effects on the conformation of the side chain of Lys 146. In 1u9a, this side chain is involved in an *inter*molecular salt bridge, specifically with Glu 122 on a neighboring protein chain. In 1u9b, on the other hand, Lys 146 is not involved in any intermolecular interactions, and sits instead in a solvent pocket formed among the various protein chains. Given these qualitatively different environments, it is unsurprising that the side chain adopts qualitatively different conformations, with the χ_1 angle, for example, differing by 113° in the two structures.

We chose to present this particular example because the effects of crystal packing on the side chain conformation are particularly simple to understand and visualize. Not all surface side chains which show large conformational changes among different crystal forms are involved in direct inter-chain salt bridges or hydrogen bonds. Conformational differences may be caused by longer-range effects of the electrostatic field associated with neighboring protein chains in the crystal. Other surface side chains show little variability among different crystal forms. To quantify the effects of crystal packing on side chain conformation, we have carried out a statistical analysis of pairs of chemically identical structures crystallized in different unit cells; the construction of the data set is described in Methods.

Side chain variability is quantified by changes in the dihedral angles χ_n. Specifically,

a dihedral angle is considered to be the "same" in a residue pair if $\Delta\chi_n < 40°$ (note that a similar criterion is commonly used for assessing the accuracy of side chain prediction algorithms). To identify residues in or near interface regions, we define δ as the smallest distance from *any* atom on a particular residue to *any* atom on a different asymmetric unit (related by a space group symmetry operation). For pairs of structures, crystal packing forces will play an important role in conformational differences for a residue pair if the residue is close to an interface region in *either* of the crystal structures. Thus, for a pair of residues, we define $\delta_{\rm min}$ as the minimum value of δ for the residue pair.

In Figure 2, the percent same χ_1 and (for side chains with more than one heavy-atom dihedral angle) χ_{1+2} are plotted with residue pairs sorted according to $\delta_{\rm min}$ in 2 Å bins. [The crystal packing effects discussed below are also quite pronounced for long side chains, i.e., χ_{1+2+3} and $\chi_{1+2+3+4}$, but fewer data points are available.] Crystal packing forces are found to be maximal near the interfaces between asymmetric units in the crystals. The variability in side chain conformation is nearly the same for pairs with same/different unit cells at large $\delta_{\rm min}$ (i.e., residues far from the interface regions), but the results diverge significantly below 4.0 Å. Thus, a good working definition of the interface region is $\delta_{\rm min} < 4.0$ Å.[2] The discrepancy between the two sets of results provides a measure of the effects of crystal packing. That is, the variability in side chain conformations in pairs with the same unit cell is due to differences in the crystal preparation and possibly also to differing methods for solving the structures. The pairs of structures with different unit cells also have such differences, but the differences in crystal packing represent much larger effects.

Residue pairs with low values of $\delta_{\rm min}$ are of course located near the surface of the protein. Thus, it can be expected that side chain variability will be increase as $\delta_{\rm min}$ decreases both due to crystal packing effects and due to intrinsic increases in variability for surface side chains

---

[2]Note that this definition is essentially identical to the definition of interface regions in protein-protein complexes used by Betts and Sternberg [11].

(i.e., caused by dynamical effects/larger B-factors and/or increased sensitivity to variation in the solvent environment such as pH or ionic strength). Figure 3 investigates side chain variability as a function of solvent accessibility (defined as the ratio of the SASA in the protein—without the crystal environment included—versus the same residue in a dipeptide). For a pair of residues, the lower of the two solvent accessibility values is chosen. The solvent accessibility serves to identify how close the residues are to the surface of the protein. Clearly, the side chain variability, as measured by the percent same χ_1, increases strongly near the surface of the protein. However, in the interface region ($\delta_{\min} < 4$ Å), pairs with different crystal packing (dashed line) show much greater variability than pairs with the same crystal packing (solid line) near the protein surface. Very little difference in variability is observed between the two data sets for side chains that are not in the interface region ($\delta_{\min} > 4$ Å), regardless of solvent accessibility. Thus, although some portion of the increase in variability near the surface is clearly independent of crystal packing, a quantitative measure of the effects of crystal packing can be obtained by comparing the results for the two data sets (same/different unit cells) in the interface region ($\delta_{\min} < 4$ Å).

We have also analyzed the results as a function of amino acid type. The detailed results are presented in Supplementary Information, and we highlight only a few results here. Serine demonstrates the most substantial increase in variability in pairs with different crystal packing arrangements. In the interface region, 84% of the pairs with the same unit cell (367 total pairs) conserve χ_1, while only 64% of those in different unit cells (199 total pairs) do; the corresponding percentages outside the interface regions are 88% and 86%. Clearly, the small polar side chain of serine provides a sensitive measure of the local electrostatic environment. The long polar and charged side chains of Arg, Lys, Gln, and Glu also demonstrate striking sensitivity to the crystal environment. For example, 51% of Arg conformations near the interface regions are conserved out to χ_4 when the crystal unit cell remains the same (342 total pairs), but only 28% are conserved when the unit cell varies (202 total pairs). The side chains which are least sensitive to the changes in crystal environment are Tyr, Pro, and Phe, although even these show slight increases in variability (<2% for χ_1 and <3% for

χ_{1+2}) in pairs with different unit cells.

These results help to resolve a discrepancy in the literature. Dunbrack *et al.* compared the observed variability of the side chains in their data set of paired structures with their side chain prediction results and concluded that "SCWRL [the prediction program] is working up to its theoretical limits for most residue types, even without consideration of hydrogen bonding, solvent effects, and electrostatic interactions" [8]. In contrast, Olson *et al.*, using their larger data set, concluded that "SCWRL is still not able to predict side-chain conformation within the tolerances defined by the observed flexibility of each residue" [9]. Our results suggest that these differing conclusions are due largely to the composition of the data sets employed in each study. Specifically, only 16% of the structural pairs in the Olson study had different crystal unit cells, *versus* 77% of the pairs in the Dunbrack study.

Side Chain Prediction Incorporating Crystal Packing

Given the strong effect of the crystal environment on surface side chain conformation, a stringent test of a side chain prediction program would be to include crystal packing interactions, i.e., by explicitly reconstructing the environment around a protein using the known space group and unit cell dimensions.[3] That is, side chain variation in different unit cells should cease to be a source of prediction "error" if the crystal packing is explicitly included. Moreover, such calculations can provide a direct measure of the effects of crystal packing, although the results of course will also reflect the accuracy of the potential function and the efficiency of the optimization algorithm used.

Although the vast majority of side chain prediction studies have ignored the effects of crystal packing, there are precedents, as early as two decades ago, for its explicit inclusion in theoretical studies of side chain conformation and dynamics. Gelin and Karplus sampled the

[3]The unit cell dimensions could in principle be treated as variables, but in practice this may not be possible. We know of no practical way of "sampling" the many possible space groups.

potential energy of side chains along the χ_1 and χ_2 dimensions using an early force field and demonstrated that the crystal environment can play an important role in stabilizing observed conformations [21]. Molecular dynamics simulations of a trypsin inhibitor in solution and in the crystalline state were carried out by van Gunsteren and co-workers; these calculations highlighted substantial changes in both conformation and dynamics of polar surface side chains upon crystallization [2]. Somewhat more recently, Wilson *et al.* performed side chain prediction on a single protein, α-lytic protease, with symmetry related atoms included [22]

As described in Methods, we have developed a new side chain prediction program which can, as an option, include the crystal environment (more precisely, the simulation system consists of one asymmetric unit and all atoms from other, surrounding asymmetric units within 20 Å). Here we emphasize the effects of including/excluding the crystal environment on prediction accuracy. Our goal is to reproduce side chain conformations with the highest possible accuracy in a wide variety of environments (crystal, solvated, complexed) and for this reason we have chosen to evaluate side chain conformations with an energy function that is substantially more complex (and computationally expensive) than those which have been used in other prediction programs. Specifically, we use the all-atom OPLS force field [13,14] for the protein intramolecular energetics and the Surface Generalized Born (SGB) implicit model of solvation [15]. The SGB model can be understood as a relatively inexpensive, analytical approximation to the Poisson-Boltzmann description of continuum electrostatics, and the parameters have been calibrated against both Poisson-Boltzmann calculations and experimental solvation free energies for a wide range of small organic molecules. This is, to our knowledge, the first time that a Generalized Born model has been used for side chain prediction. Most prior studies have been performed in the gas phase, and many have used only simple packing potentials; the only prior treatment of solvation has been with the low accuracy but computationally convenient distance-dependent dielectric models [23–29]. Our choice of energy function is capable of modeling all of the important effects of the crystal environment, including simple short-range packing effects (i.e., van der Waals interactions), longer-range electrostatic effects, and desolvation.

As a first test, we predicted the conformations of *single* side chains while holding the rest of the protein fixed at the native configuration (similar tests in other contexts have been performed by Gelin and Karplus [21], Wilson *et al.* [22], Petrella *et al.* [32], Xiang and Honig [33], and Liang and Grishin [34]). This test has the advantage that sampling error is not a serious problem, since the combinatorial optimization problem is entirely avoided. As will be shown below, the effects of crystal packing are revealed quite clearly by this test.

The single side chain prediction results are depicted in Figure 4. The solid/dashed lines represent the results with/without crystal packing forces included. The thin dotted lines are the results of the database analysis, included for ease of comparison (the top/bottom dotted lines are the results with the same/different unit cells). The agreement between the two analyses of crystal packing effects is remarkable. The calculated results with the crystal environment included achieve just slightly less accuracy than would be obtained by guessing the side chain conformations of a protein in one crystal from the conformations observed in another crystal with the same unit cell. Neglect of the crystal environment leads to prediction accuracy that is roughly comparable to guessing the side chain conformations of a protein in one crystal from the conformations observed in another crystal with a different unit cell. The single side chain results are analyzed as a function of residue type in the Supplementary Information. These results are generally consistent with those from the paired protein database study; that is, polar and charged side chains demonstrate the largest improvement in prediction accuracy when the crystal environment is represented.

The single side chain prediction results provide strict upper limits on the accuracy that a given model can achieve for full side chain prediction. Thus, the neglect of the crystal environment places severe limitations on the achievable accuracy of *surface* side chain prediction. We have also performed *full* side chain prediction in the crystal environment, by, in essence, optimizing the side chains on all symmetry-related copies of the asymmetric unit simultaneously. See Methods for further details.

The full side chain prediction results are shown in Figure 5. The decrease in accuracy relative to the single side chain results is likely due to both incomplete sampling (i.e., not

locating the global minimum) and inadequacy of the potential function, which is tested more rigorously when all side chains are free to move. Nevertheless, the effects of crystal packing can still be observed very clearly. A comparison of these results with other side chain prediction programs is presented in the Conclusion below; the overall results are 88% correct χ_1, 77% correct χ_{1+2} with the crystal environment *included*, and 85% correct χ_1, 73% correct χ_{1+2} with the crystal environment *excluded*. [Note that the effect of including/excluding crystal packing on the overall statistics is relatively minor, due to the relatively small number of residues involved in intermolecular crystal contacts. More generally, such overall statistics can mask substantial errors in surface side chain prediction. However, as Figures 3 and 5 emphasize, the effect of crystal packing on surface side chain conformations, and particularly those in the interface regions, is substantial.]

CONCLUSION

Given the results presented here, what is the ultimate limit on the accuracy that could potentially be achieved by side chain prediction programs, and how close does the current generation of programs actually come to this limit? A definitive answer to this question is complicated by variation in the protein data sets chosen in different studies and inconsistencies in the criteria chosen to evaluate accuracy. Nonetheless, several conclusions can be drawn.

First, care must be taken, when attempting to establish "instrinsic" side chain variability, to distinguish between pairs of protein structures with the same and with different crystal unit cells. We do not believe that side chain conformational variability observed in different crystal environments should be considered "intrinsic" and used as a point of comparison with side chain prediction programs, as has been done implicitly in other work (i.e., no distinctions made between same and different unit cells). In order for a side chain prediction program to be of maximal utility to biological and pharmaceutical chemistry, it must be capable of predicting side chain conformations in a variety of environments, including both

unimolecular solvated proteins, proteins in complexes with ligands and other proteins, and proteins in membrane environments. The variability observed in pairs of structures with the same crystal unit cell provides a more meaningful and stringent point of comparison with side chain prediction programs and with side chain conformational variation due to docking or mutation. Ultimately, however, even the side chain conformational variability observed in structural pairs with the same crystal unit cell should not be considered intrinsic, as it is likely to be related to pH and the identity and concentration of ions (both those that can be imaged by the x-ray diffraction experiment [20] and labile ions in the solvent channels); information concerning crystallization conditions could potentially be used for further studies of side chain variability.

Neglecting the variability in crystallization conditions, the results here provide definite targets for achievable side chain prediction accuracy when comparing to protein structures solved by x-ray crystallography (see Figure 3): >95% correct χ_1 for core residues, >80% correct for surface side chains that are involved in crystal contacts, and 60–80% correct for surface side chains not involved in crystal contacts (the precise value depends on the precise level of solvent accessibility). Of course, these results can only be achieved if validation studies for side chain prediction programs are performed with crystal packing included, as we have done here, in order to adequately represent the physical environment of the proteins in the test set. [Alternatively, one could propose to use protein structures solved by solution phase NMR for validation, but *side chain* conformations, particularly beyond χ_1, are frequently poorly restrained by the experimental data.]

Encouragingly, there have been several reports of side chain prediction accuracy approaching or even slightly exceeding the 95% correct χ_1 level for core residues [29,33,34]. The prediction methods that achieve this level of accuracy for the core residues are all relatively recent and utilize somewhat more complicated energy/scoring functions than in many prior studies, including at least some electrostatic effects. The side chain prediction program introduced here utilizes arguably the most sophisticated energy function yet employed for side chain prediction: the all-atom OPLS force field with a Generalized Born implicit solvent

model. The results for core residues (<20% solvent exposure) are excellent, 95.6% correct χ_1. However, even methods that use only steric clash and/or statistical preferences to place side chains perform reasonably well in the core, with χ_1 prediction accuracies approaching 90% in some cases, including results reported for the SCWRL program [8]. Although gratifying, these excellent results for core residues are unsurprising given the densely packed core environment. [Prediction results for dihedral angles beyond χ_1, and particularly beyond χ_2, are less commonly reported, but the available evidence suggests that prediction accuracy degrades rapidly further from the backbone, and more rapidly than does the reproducibility of the side chain dihedral angles, as studied here.]

Most active sites in proteins involve significant solvent exposure, and thus the prediction accuracy for surface and partially buried side chains is of greater biological importance than the prediction accuracy for buried side chains. As we have demonstrated here, assessing the performance of side chain prediction programs for the surface side chains is complicated by the effects of crystal packing, because few validation studies have explicitly incorporated the crystal environment (the only studies to do so, to our knowledge, are the very early work of Gelin and Karplus [21], Wilson *et al.* [22], and this study). We encourage other groups to do so, or at least to divide surface side chains into those that participate in intermolecular crystal contacts and those that do not, in order to assess more sensitively the adequacy of currently employed energy functions for surface side chain prediction. The results presented here (Figure 5) indicate that substantial improvement in accuracy should still be possible. Much of the remaining error appears to be due largely to inadequacies of the potential function rather than inadequacies of the sampling algorithm, because the predicted protein structures nearly always have substantially lower energies than those of the native (or minimized native). Efforts are currently underway to refine the force field (specifically the torsional parameters [41]) and solvent model using side chain prediction as a measure of accuracy.

METHODS

Database Analysis

Of the 123 pairs of structures in the Zhao *et al.* study [9], only 20 corresponded to different crystal unit cells.[4] To provide better statistics, we augmented this data set with certain pairs of structures used in a study by Bower *et al.*, in which the authors were concerned with evaluating the SCWRL side chain prediction program [8]. In particular, that study relied largely on the large number of crystal structures of lysozyme (both hen egg white lysozyme, with 10 uncomplexed structures in 4 space groups, and bacteriophage T4 lysozyme, with 3 structures in 2 space groups).[5] We augmented the Zhao *et al.* data set by 20 representative pairs of hen egg white lysozyme structures (10 pairs with the same space group[6] and 10 with different space groups[7]). We did not, however, use any of the T4 lysozyme structures, due to their relatively low resolution (2.7 Å) and high backbone RMSD among the pairs (up to 3.4 Å), which is associated with large-amplitude hinge bending [6].

When we analyzed the combined data set for the first time, we noticed that several of the pairs of structures with the same unit cell had extremely similar side chain confor-

---

[4]These pairs are 193l/1aki, 194l/1aki, 1bxa/1aaj, 2rac/1aaj, 1yme/1cpx, 1arl/1cpx, 1tld/1tgn, 1mpb/1mpc, 1pgb/1pga, 1svn/1jea, 1mku/1mks, 1cub/1cuc, 1une/1mkt, 1une/2bpp, 5pti/6pti, 4pti/6pti, 1fib/1fid, 3fib/1fid.

[5]The Bower *et al.* study also used structures of bovine pancreatic trypsin inhibitor, but representative pairs are included in the Zhao *et al.* data set.

[6]Specifically, all pairs of 5 structures in the $P4_32_12$ space group: 6lyt, 1hel, 1lza, 1lse, and 2lym.

[7]All possible pairs of structures among 1lza ($P4_32_12$), 2lzt ($P1$), 1lma ($P2_1$), 5lym ($P2_1$, Chain A), and 5lym ($P2_1$, Chain B). Note that although 1lma and 5lym have the same space group, the unit cells are different. Specifically, the 5lym asymmetric unit contains two non-identical copies of the lysozyme chain; each of these is used independently, because they have different environments.

mations, raising suspicions that the pairs did not represent truly independent structural determinations. (By definition, no such issue arises for the pairs with different unit cells.) A small study by Flores *et al.* included only pairs of structures that were solved by different research groups, in an effort to ensure that all structure determinations were independent [12]; no such precaution was taken in the much larger Zhao *et al.* study [9]. Indeed, some of the highly similar pairs were solved by the same research groups and differed only trivially in crystallization conditions, such as 193l and 194l, which were crystallized and analyzed identically, except that one of the structures was grown under zero-gravity conditions [19]. Other extremely similar pairs were solved by molecular replacement using the same starting models. However, it should be emphasized that not all pairs solved by molecular replacement showed this behavior, including pairs in which one structure was the starting model for the other structure, presumably reflecting extensive refinement of the initial models.

For our purposes, the data set of pairs with the same unit cell would ideally encompass the same diversity of crystallization conditions and methods for solving the structures as the set of pairs with different unit cells. This is a difficult criterion to quantify, and in practice we used a simple but effective criterion to remove those pairs of structures with the same unit cell which evidenced much higher similarity than other pairs, to the extent that they can be considered outliers. In the left column of Figure 6, we present histograms which illustrate the range of structural similarity observed for the pairs with the same crystal unit cells. The criterion we use for comparing side chain conformations, here and below, is one that is commonly used for assessing side chain prediction algorithms: two side chain dihedral angles are the same if they are with $\pm 40°$ of each other. In each of the histograms there is a clear "spike" very near 100% similarity, which represents the structures with extremely high similarity.

Our criterion for pruning the set of structures with the same unit cell is to reject all structures which have over 90% of all relevant side chains with the same values of χ_1, χ_2, and χ_3. We abbreviate this criterion as χ_{1+2+3}. This criterion is used for a cutoff rather than χ_{1+2} or χ_1 because its use makes it easier to distinguish the outliers. The precise cutoff

is of course somewhat arbitrary, but the histograms for the pruned data set (right column of Figure 6) no longer have obvious outliers. The final sets of pairs used in this study are listed in Tables I and II. Note that the average resolutions of the structures in the two data sets are extremely similar (and quite low), 1.75 Å in both cases. We wish to emphasize that, although the pruning of the data set changes the precise statistics that we obtain, by <10% percent in most cases, the overall conclusions listed below remain unchanged. We have also repeated our analyses using the Zhao *et al.* and Bower *et al.* data sets independently, again with no significant changes to the overall results. Thus, although the precise data set chosen is, inevitably, somewhat arbitrary, the differences that we observe between pairs of structures with and without the same unit cell are robust.

One final issue concerning the construction of the data sets is the extent of backbone variation among the pairs. The average backbone RMSD is low in both data sets, but the pairs with different unit cells (0.58 Å average RMSD), expectedly, demonstrate greater backbone variation than the pairs with the same unit cells (0.29 Å average RMSD). Although the backbone variation is fairly small and arises largely from short disordered terminal regions, some of the observed crystal packing effects on side chain conformation can potentially arise indirectly from modifications to the backbone structure. No rigorous decomposition between the direct and indirect mechanisms can be made in the database analysis, and for many applications the distinction is not critical. However, as described below, we also perform calculations in which the backbone is held fixed at the native configuration, and crystal packing forces are turned on and off. These calculations demonstrate that crystal packing forces have a strong direct effect on side chain conformations, roughly equal to that observed in the database analysis.

Side Chain Prediction

A diverse set of high resolution protein structures solved by X-ray crystallography were chosen for use in this study. Specifically, 36 proteins were selected from a "Culled PDB"

list compiled by Dunbrack [30,31] which consists of 909 protein structures solved to 2.0 Å resolution or better (R-value < 0.2) with maximum pairwise sequence identity of 30% or less. Proteins with non-peptide ligands or nonstandard (chemically modified) residues were excluded from study, as were proteins with large disordered regions. The largest protein contained 285 residues; the total number of residues represented is 4808. The PDB codes for the proteins included are 1ew4, 1u9a, 5icb, 2pth, 1bk7, 1dvo, 3vub, 1et1, 1aie, 1ej8, 2fcb, 1nps, 1whi, 1aho, 1bv1, 1c44, 1edm, 2igd, 1d4t, 1dhn, 1qto, 1ay7, 5hpg, 1f94, 3ezm, 1pbv, 1qtw, 1bue, 2btc, 1sur, 1b2p, 1a8l, 1byi, 1ako, 1tvd, 2plc, 1qts. This data set encompasses considerably greater diversity of structure than the paired protein database.

Monoatomic ions reported in the PDB files were included in the calculations, to avoid gross errors in the electrostatic environment. The protonation states of titratable side chains were assigned using pH information in the PDB files and the assumption that the pKa's of the side chains in the protein environment are unmodified from those in the isolated amino acids. That is, the side chain was considered to be protonated if pH $<$ pKa. The positions of hydrogen atoms and all other atoms not reported in the PDB files were determined as follows. First, all unreported atoms were placed in "standard" geometries, as defined by the OPLS force field. Next, the positions of polar hydrogen atoms were optimized using the energetic function described below, OPLS with SGB/NP solvation, by scanning the hydrogen dihedral angles at 10° intervals. Finally, the positions of all unreported atoms were energy minimized using the algorithm described below.

For most proteins, the crystal unit cell contains too many atoms for explicit lattice summation techniques (e.g., Ewald summation) to be computationally feasible. Instead, the simulation system that we employ consists of one asymmetric unit (which may contain more than one protein chain) and all atoms from other, surrounding asymmetric units that are within 20 Å. Every copy of the asymmetric unit is identical at every stage of the calculation (e.g., if the conformation of a side chain is modified, it is changed on all copies of the asymmetric unit simultaneously).

The all-atom OPLS force field [13,14] was used to describe the protein intramolecular energetics. All nonbonded interactions between pairs of atoms within 20 Å of each other were calculated. The solvation free energy was estimated using an implicit solvent model consisting of the Surface Generalized Born (SGB) model of polar solvation [15], and a non-polar estimator developed by Levy and co-workers [40]. Specifically, the SGB solvation free energy (in kcal/mol) is calculated according to

$$-166\left(1-\frac{1}{\epsilon}\right)\sum_{i}^{N}\sum_{j}^{N}\frac{q_i q_j}{\sqrt{(r_{ij}^2+\alpha_i\alpha_j\exp -D_{ij}}} \tag{1}$$

where $D_{ij}=r_{ij}^2/4\alpha_i\alpha_j$, q represents partial atomic charges, r_{ij} is the distance between two atoms, and ϵ is the solvent dielectric constant. The "Born" α parameters are given by an integral over the surface of the protein

$$\frac{4\pi}{\alpha_k}=\int_S\frac{(\vec{R}-\vec{r}_k)\cdot\vec{n}(\vec{r})}{|\vec{R}-\vec{r}_k|^4}d^2\vec{R} \tag{2}$$

where $\vec{n}$ is the surface normal, $\vec{R}$ is the vector of integration over the whole surface, and $\vec{r}_k$ is an atomic position. As described in Ref. [15], correction terms have also been developed to improve the agreement between SGB and Poisson-Boltzmann solvation free energy calculations.

In this work, a new version of the SGB/NP solvation free energy code was written to account for the crystallographic symmetry. That is, the Born α's are calculated with the full crystal packing. Two surfaces (extending over the entire simulation region, including the symmetry copies) were calculated to perform the SGB surface integrals: one high resolution (330 points per sphere) and one low resolution (only 10 points per sphere). The distributions of points on the spheres used to construct these surfaces were determined using the spiral points algorithm [39]. Because the magnitude of the integrand of the surface integral decreases rapidly with distance, the integration was performed with the high resolution surface for all points within 7.5 Å of the charge in question, and the lower resolution surface at longer distances, up to an absolute cutoff of 20 Å.

The sampling of *single* side chain conformations was accomplished primarily by using a

highly detailed (10° resolution) rotamer library constructed by Xiang and Honig [33] from a database of 297 proteins. This library contains, for example, 2086 rotamers for lysine. The additional computational expense of such a detailed library was tolerated in order to ensure adequate sampling. In addition, the expense was mitigated by pre-screening the rotamers using only hard sphere overlap as a criterion, allowing many rotamers to be excluded before performing any energy evaluations. This step can be made very rapid with the use of a cell list (1 Å grid size) to identify nearby atoms. In addition, for side chains with multiple dihedral angles, the discovery of a single steric clash can eliminate many rotamers. For example, if a steric clash is found for the C_γ atom of Lys with $\chi_1 = 120°$, then all rotamer states with that value of the first dihedral angle can be eliminated.

After choosing the lowest energy rotamer, the side chain is completely energy-minimized (<0.001 kcal/mol/Å final root-mean-square gradient) in Cartesian space (i.e., all side chain atoms are free to move) using a novel multi-scale minimization algorithm [35]. This algorithm is a variant of the Truncated Newton (TN) method, specifically the TNPACK implementation of Schlick and co-workers [36]. The multi-scale implementation of TNPACK by Jacobson and Friesner [35] is based upon a division of the molecular mechanics forces into short- and long-range components, in analogy to multi-scale molecular dynamics methods such as RESPA [38]. Short-range forces include all bond, angle, and torsion terms in the force field, as well as all nonbonded interactions between atoms separated by <10 Å. The remaining nonbonded interactions constitute the long-range forces. The long-range forces are *never* evaluated during the "inner" TN cycles (which determine the line search direction), and only periodically updated in the outer TN cycles (in this work, once every 5 Newton cycles). The division of the nonbonded interactions into short- and long-range is also updated every 5 Newton cycles.

The Generalized Born solvation model is well suited for performing rapid minimizations because the pair screening term is analytical and thus differentiable. However, the resultant expression for the gradient involves derivatives of the Born α's with respect to the atomic coordinates, which must be determined numerically. A simple solution to this prob-

lem, however, can be obtained by simply holding the Born α's fixed during the course of the minimization, then updating them, performing another minimization, and so on until self-consistency is achieved (defined by the energy varying by less than 1 kcal/mol). In practice, self-consistency rarely requires more than 2 cycles of TN minimization, and the second minimization is generally extremely rapid (i.e., only a very small number of Newton cycles, with the energy typically changing by only 0.01–0.1 kcal/mol). This self-consistent minimization with GB solvent requires only $\sim$50% greater computational expense than vacuum minimizations.

The method we use for the combinatorial optimization is adopted from the method described by Xiang and Honig [33], which is similar in spirit to earlier work by Bruccoleri and Karplus [37]. In brief, 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, using the single side chain procedure described above, holding the others fixed. The procedure is iterated to convergence (no side chains changing rotamer states) After convergence is achieved, all side chains are completely energy minimized simultaneously in Cartesian coordinates to remove any remaining clashes. This complete procedure is repeated several times (5 in our case; greater numbers of iterations did little to improve accuracy), because different initial (random) rotamer states lead to somewhat different optimized structures. The lowest energy structure is chosen for comparison with the experimental data.

ACKNOWLEDGMENTS

MPJ wishes to acknowledge support from an NSF Postdoctoral Fellowship in Bioinformatics. This work was supported in part by a grant to RAF from the NIH (GM-52018).

REFERENCES

[1] Wagner, G., Hyberts, S. G. & Havel, T. F. (1992). NMR structure determination in solution: a critique and comparison with X-Ray crystallography. *Annu. Rev. Biophys. Biomol. Struct.*, **21**, 167–198.

[2] van Gunsteren, W. F. & Berendsen, H. J. (1984) Computer simulation as a tool for tracing the conformational differences between proteins in solution and in the crystalline state. *J. Mol. Biol.*, **176**, 559–564.

[3] Kossiakoff, A. A., Randal, M., Guenot, J. & Eigenbrot, C. (1992) Variability of conformations at crystal contacts in BPTI represent true low-energy structures: correspondence among lattice packing and molecular dynamics simulations. *Proteins*, **14**, 65–74.

[4] Eigenbrot, C., Randal, M. & Kossiakoff, A. A. (1992). Structural effects induced by mutagenesis affected by crystal packing factors: the structure of a 30–51 disulfide mutant of basic pancreatic trypsin inhibitor. *Proteins*, **14**, 75–87.

[5] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Research* **2000**, *28*, 235.

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

[7] Kishan, K. V. R., Zeelen, J. Ph., Noble, M. E. M, Borchert, T. V. & Wierenga, R. K. (1994). Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms. *Prot. Sci.*, **3**, 779–787.

[8] Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.*, **267**, 1268–1282.

[9] Zhao, S., Goodsell, D. S. & Olson, A. J. (2001). Analysis of a data set of paired uncom-

plexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins*, **43**, 271–279.

[10] Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins*, **39**, 261–268.

[11] Betts, M. J. & Sternberg, M. J. E. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Prot. Eng.*, **12**, 271–283.

[12] Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Prot. Sci.*, **2**, 1811–1826.

[13] Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **118**, 11225–11236.

[14] Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., Jorgensen, W. J. (2001). Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, **105**, 6474–6487.

[15] Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998). Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B*, **102**, 10983–10990.

[16] Tong, H., Hateboer, G., Perrakis, A., Bernards, R. & Sixma, T. K. (1997). Crystal structure of murine/human Ubc9 provides insight into the variability of the ubiquitin-conjugating system. *J. Biol. Chem.*, **272**, 21381–21387.

[17] Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD—Visual Molecular Dynamics. *J. Mol. Graph.*, **14**, 33–38.

[18] Islam, S. A. & Weaver, D. L. (1990). Molecular interactions in protein crystals: solvent accessible surface and stability. *Proteins*, **8**, 1–5.

[19] Vaney, M. C., Maignan, S., Riès-Kautt, M. & Ducruix, A. (1996). High-resolution structure (1.33 Å) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Cryst.*, **D52**, 505–517.

[20] Vaney, M. C., Broutin, I., Retailleau, P., Douangamath, A., Lafont, S., Hamiaux, C., Prangé, T., Ducruix, A. & Riès-Kautt, M. (2001). Structural effects of monovalent anions on polymorphic crystals. *Acta Cryst.*, **D57**, 929–940.

[21] Gelin, B. R. & Karplus, M. (1979). Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochem.*, **18**, 1256–1268.

[22] Wilson, C., Gregoret, L. M. & Agard, D. A. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.*, **229**, 996–1006.

[23] Roitberg, A. & Elber, R. (1991). Modeling side chains in peptides and proteins: applications of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.*, **91**, 9277–9287.

[24] Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**, 539–542.

[25] Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1993). A critical comparison of search algorithms applied to the optimization of protein side-chain conformations. *J. Comput. Chem.*, **14**, 790–798.

[26] Tanimura, R., Kidera, A., & Nakamura, H. (1994). Determinants of protein side-chain packing. *Prot. Sci.*, **3**, 2358–2365.

[27] Tufféry, P., Etchebest, C., & Hazout, S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformational stability in the rotamer space. *Prot. Eng.*, **10**, 361–372.

[28] Leach, A. R., & Lemon, A. P. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, **33**, 227–239.

[29] Mendes, J., Baptista, A. M., Carrondo, M. A. & Soares, C. M. (1999). Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins*, **37**, 530–543.

[30] Dunbrack, R. L., Jr. Culling the PDB by resolution and sequence identity. http://www.fccc.edu/research/labs/dunbrack/culledpdb.html

[31] Hobohm, U., Scharf, M. & Schneider, R. (1993). Selection of representative protein data sets. *Prot. Sci.*, **1**, 409–417.

[32] Petrella, R. J., Lazardis, T. & Karplus, M. (1998). Protein sidechain conformer prediction: a test of the energy function. *Fold. Des.*, **3**, 353–377.

[33] Xiang, Z. & Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.*, **311**, 421–430.

[34] Liang, S. & Grishin, N. V. (2002). Side-chain modeling with an optimized scoring function. *Prot. Sci.*, **11**, 322–331.

[35] Jacobson, M. P. & Friesner, R. A. Protein Local Optimization Program (PLOP): a new software platform for all-atom protein structure refinement. In preparation.

[36] Xie, D. & Schlick, T. (1999). Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications. *SIAM J. Optim.*, **10**, 132–154.

[37] Bruccoleri, R. E. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, **26**, 137–168.

[38] Tuckerman, M., Berne, B. J. & Martyna, G. J. (1993). Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, **97**, 1990–2001.

[39] Rakhmanov, E. A., Saff, E. B. & Zhou, Y. M. (1994). Minimal discrete energy on the sphere. *Math. Res. Lett.*, **1**, 647–662.

[40] Gallicchio, E., Zhang, L. Y. & Levy, R. M. (2002). The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.*, **23**, 517–529.

[41] Jacobson, M. P., Kaminski, G. A., Friesner, R. A. & Rapp, C. S. (2002). Force field validation using protein side chain prediction. *J. Phys. Chem. A*, submitted.

1bxa/2rac	1ak2/2ak2	1ald/2ald	1axn/1aii
3rn3/1rat	3rn3/1rhb	1rat/1rhb	1brf/1bq8
1brf/1caa	1bq8/1caa	1lhm/2bqa	2lzm/3lzm
2lzm/4lzm	3lzm/4lzm	1lz1/1rex	1npk/1nsp
1rro/1omd	1paz/3paz	1paz/1pza	3paz/1pza
4pnp/1pbn	1pnc/1pnd	1pnc/2pcy	1pnd/2pcy
2pkc/2prk	1sbh/1yja	1sbh/1yjb	1yja/1yjb
2st1/1st2	1tcy/1wqr	2tgi/1tfg	1amf/1wod
1a58/1a33	1bkr/1aa2	1zia/1zib	1osa/1clm
1scs/2ctv	1enr/2ctv	1crm/2cab	1ede/2had
1ede/2dhd	1ert/1eru	1ert/1auc	1eru/1auc
1amm/1gcs	1i1b/4i1b	2ilk/1ilk	1jcv/2jcw
1jcv/1yso	2jcw/1yso	1top/1ncx	1top/1ncz
1xac/1xad	1mkt/2bpp	2che/2chf	1gvp/1vqb
2lhm/3lhm	5icb/6icb	3rnt/8rnt	5pti/4pti
4rxn/5rxn	1djc/1djb	3lip/2lip	1loz/1oua
1a3z/1rcy	1ame/1gzi	4paz/5paz	6paz/7paz
6lyt/1hel	6lyt/2lym	6lyt/1lza	6lyt/1lse
1hel/2lym	1hel/1lza	1hel/1lse	2lym/1lza
2lym/1lse	1lse/2lym		

TABLE I. Pairs of protein structures with same unit cell (77).

193l/1aki	194l/1aki	1bxa/1aaj	2rac/1aaj
1yme/1cpx	1arl/1cpx	1tld/1tgn	2ptn/1tgn
1mpb/1mpc	1pgb/1pga	1svn/1jea	1mku/1mks
1cub/1cuc	1une/1mkt	1une/2bpp	5pti/6pti
4pti/6pti	1fib/1fid	3fib/1fid	1u9b/1u9a
1lza/2lzt	1lza/1lma	1lza/5lym	1lza/6lym
2lzt/1lma	2lzt/5lym_A	2lzt/5lym_B	1lma/5lym_A
1lma/5lym_B	5lym_A/5lym_B		

TABLE II. Pairs of protein structures with different unit cells (30).

FIGURES

FIG. 1. Top: Crystal unit cells for 1u9a (right; $P2_1$) and 1u9b (left; $I222$). The proteins are chemically identical (human ubiquitin-conjugating enzyme Ubc9). The atoms in these space filling models are colored according to distinct protein chains, the positions and orientations of which are related to each other by space group symmetry operations. Note both the prominent solvent channels in each crystal, as well as the extensive inter-chain contacts, including interdigitation of surface side chains. Bottom: Conformations of Lys 146 (red space filling) in the two structures. In 1u9a, this side chain participates in an *inter*molecular salt bridge with Glu 122 on a different protein chain; in 1u9b, the side chain makes no intermolecular contacts and sits in a solvent pocket. Only residues within 12 Å of Lys 146 are depicted. All images were generated with VMD [17], and rendered with POV-Ray.

FIG. 2. Fraction of side chains that have identical, using the $\pm 40°$ criterion, χ_1 dihedral angles (top) and identical χ_1 *and* χ_2 values (bottom); the latter criterion is abbreviated as χ_{1+2}. The data are averaged over 2 Å bins of $\delta_{\rm min}$, which provides a measure of whether *either* of a pair of residues is close to an interface region between asymmetric units in the crystal. The solid/dashed lines represents pairs with the same/different unit cells, respectively.

FIG. 3. Side chain variability (as measured by % "same" χ_1, i.e., $\Delta\chi_1 < 40°$) as a function of solvent accessibility (ratio of solvent accessible surface area in the protein to that in a dipeptide). The solid/dashed lines represents pairs with the same/different unit cells, respectively. Top: All residue pairs with $\delta_{\rm min} < 4$ Å, i.e., those residues that reside in "interface regions" in the crystal. Bottom: All residue pairs with $\delta_{\rm min} > 4$ Å. The final point for the dashed line in the bottom panel is not represented due to inadequate statistics. Specifically, there are only 2 residue pairs with $\delta_{\rm min} > 4$ Åand solvent accessibility of greater than 0.8. Side chain variability increases strongly near the surface of the protein, but the effects of crystal packing can be isolated by comparing the results with same/different unit cells for residues in the interface regions.

FIG. 4. Prediction accuracy for single side chains (keeping the remainder of the protein fixed at the native), with (solid line) and without (dashed line) inclusion of the crystal environment. The axes are defined as in Figure 2. The dotted lines are taken from Figure 2, for ease of comparison.

FIG. 5. Prediction accuracy for full side chain addition (keeping the backbone fixed at the native), with (solid line) and without (dashed line) inclusion of the crystal environment. The axes are defined as in Figure 2. The dotted lines are taken from Figure 2, for ease of comparison.

FIG. 6. Histograms representing the distribution of side chain variability in the data set of paired structures with the same crystal unit cells, before (left column) and after (right column) pruning. The pruning criterion was chosen to be $\chi_{1+2+3} > 90\%$.

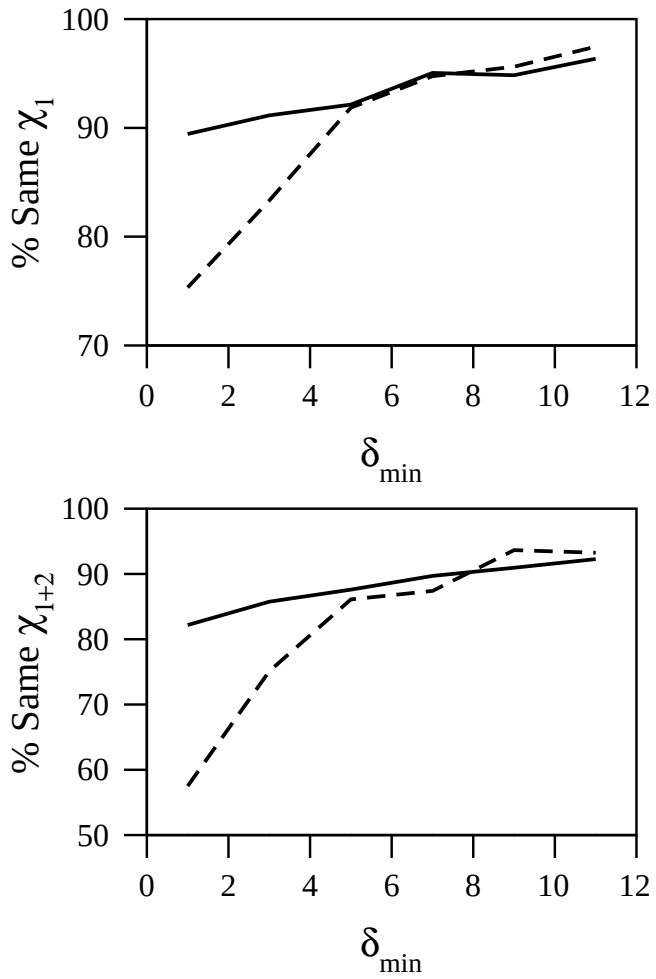
100
90
80
70
% Same χ_1
0
2
4
6
8
10
12
δ_{min}
100
90
80
70
60
50
% Same χ_{1+2}
0
2
4
6
8
10
12
δ_{min}

Figure 2

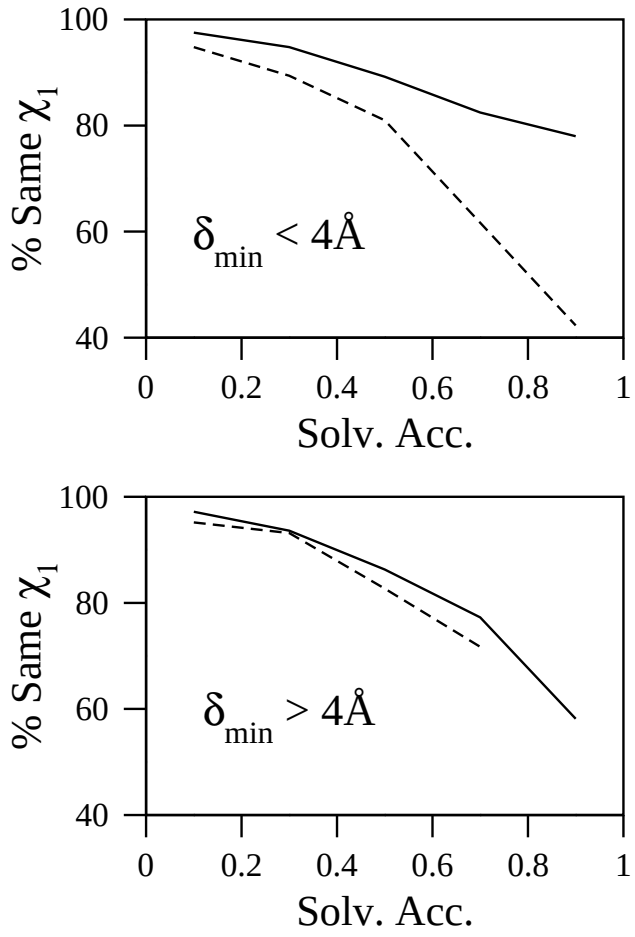
100
80
60
40
% Same χ_1
δ_{min} < 4Å
0
0.2
0.4
0.6
0.8
1
Solv. Acc.
100
80
60
40
% Same χ_1
δ_{min} > 4Å
0
0.2
0.4
0.6
0.8
1
Solv. Acc.

Figure 3

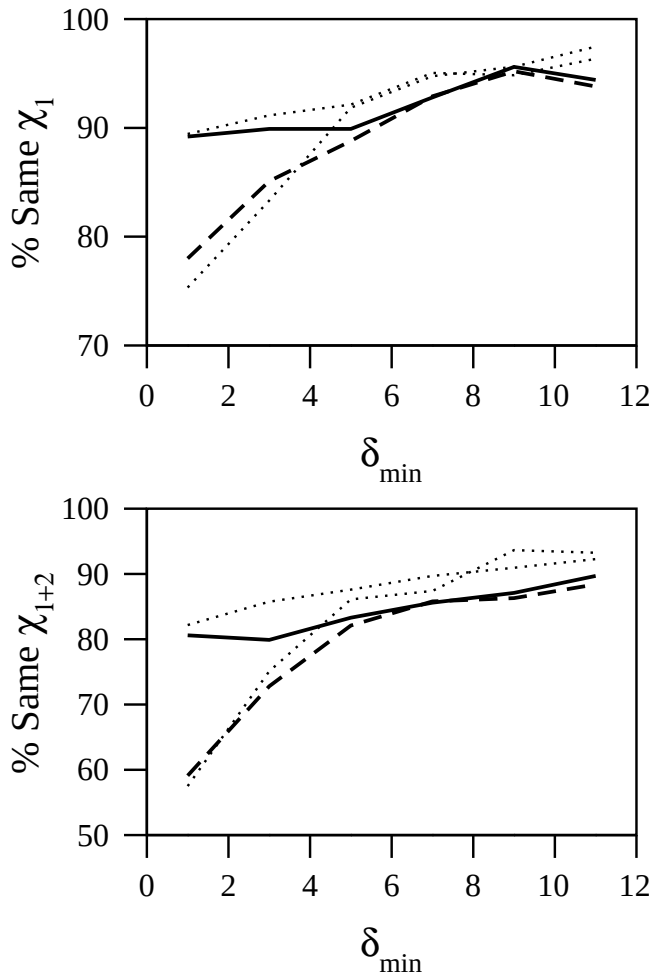
100
90
80
70
% Same χ_1
0
2
4
6
8
10
12
δ_{min}
100
90
80
70
60
50
% Same χ_{1+2}
0
2
4
6
8
10
12
δ_{min}

Figure 4

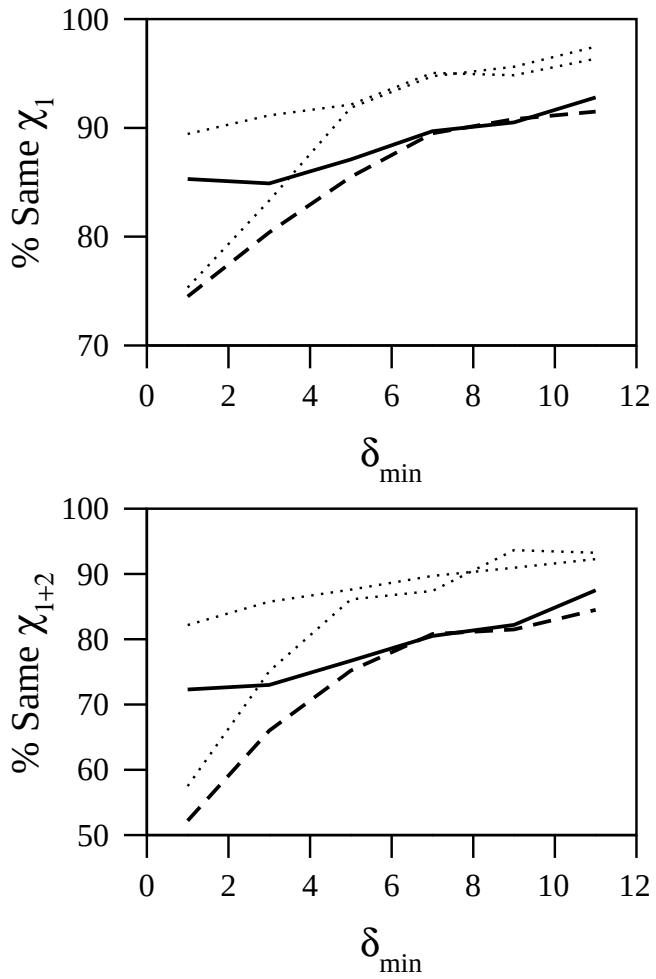
100
90
80
70
% Same χ_1
0
2
4
6
8
10
12
δ_{min}
100
90
80
70
60
50
% Same χ_{1+2}
0
2
4
6
8
10
12
δ_{min}

Figure 5

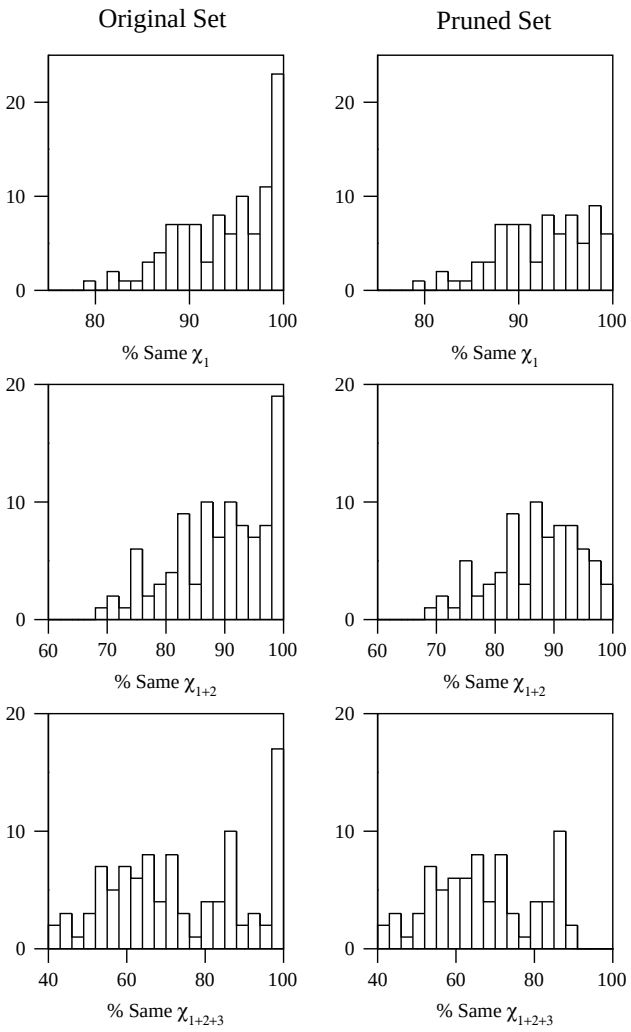
Original Set
Pruned Set
20
10
0
80
90
100
% Same χ_1
60
70
% Same χ_{1+2}
40
% Same χ_{1+2+3}

Figure 6

SUPPLEMENTARY INFORMATION

	Same Unit Cell			Diff. Unit Cell		
	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$
ARG	495	332	163	220	191	29
LYS	753	441	312	276	185	91
GLN	383	222	161	139	89	50
GLU	653	347	306	166	89	77
MET	279	89	190	73	10	63
ASN	701	348	353	375	257	118
ASP	737	348	389	260	149	111
ILE	610	164	446	250	50	200
LEU	866	259	607	311	112	199
TRP	172	65	107	135	53	82
HIS	223	93	130	75	29	46
PHE	464	157	307	168	66	102
TYR	397	163	234	191	114	77
SER	877	364	513	361	189	172
CYS	276	84	192	196	70	126
THR	660	277	383	290	171	119
VAL	892	221	671	250	70	180
PRO	527	284	243	170	99	71

TABLE III. Number of residue pairs in the paired protein data sets.

	Same Unit Cell			Diff. Unit Cell		
	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$
ARG	86.5	84.6	90.2	74.1	71.2	93.1
LYS	85.3	87.1	82.7	83.3	78.9	92.3
GLN	86.2	80.2	94.4	77.7	66.3	98.0
GLU	85.5	85.0	86.0	86.8	78.7	96.1
MET	91.4	87.6	93.2	93.2	80.0	95.2
ASN	95.4	96.0	94.9	83.7	79.8	92.4
ASP	96.6	96.5	96.7	88.5	85.2	92.8
ILE	97.4	97.0	97.5	91.2	84.0	93.0
LEU	94.3	91.1	95.7	89.7	87.5	91.0
TRP	100.0	100.0	100.0	97.8	94.3	100.0
HIS	98.7	97.8	99.2	97.3	93.1	100.0
PHE	99.3	98.7	99.7	100.0	100.0	100.0
TYR	99.8	99.4	100.0	98.4	98.2	98.7
SER	85.9	83.5	87.5	74.8	64.0	86.6
CYS	99.6	100.0	99.5	99.5	100.0	99.2
THR	94.8	93.1	96.1	87.9	84.8	92.4
VAL	94.0	91.4	94.8	84.0	72.9	88.3
PRO	90.7	91.5	89.7	94.1	90.9	98.6

TABLE IV. Percent identical χ_1 in the paired protein data sets, using the $\pm 40°$ criterion.

	Same Unit Cell			Diff. Unit Cell		
	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$
ARG	78.6	75.0	85.9	54.5	50.3	82.8
LYS	71.8	72.6	70.8	68.5	62.7	80.2
GLN	81.5	74.8	90.7	64.8	47.2	96.0
GLU	80.7	81.3	80.1	80.1	71.9	89.6
MET	88.5	85.4	90.0	93.2	80.0	95.2
ASN	91.9	92.0	91.8	76.5	72.0	86.4
ASP	91.6	92.2	91.0	79.2	74.5	85.6
ILE	92.8	92.1	93.0	81.6	62.0	86.5
LEU	84.0	78.4	86.3	73.6	68.8	76.4
TRP	98.3	95.4	100.0	92.6	81.1	100.0
HIS	98.2	96.8	99.2	94.7	93.1	95.7
PHE	98.9	98.7	99.0	99.4	98.5	100.0
TYR	98.7	98.2	99.2	96.9	95.6	98.7
PRO	87.7	88.7	86.4	86.5	85.9	87.3

TABLE V. Percent identical χ_{1+2} (i.e., both χ_1 and χ_2 identical) in the paired protein data sets, using the $\pm 40°$ criterion.

	Same Unit Cell			Diff. Unit Cell		
	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$
ARG	63.0	56.9	75.5	39.5	34.0	75.9
LYS	59.0	57.8	60.6	53.3	46.5	67.0
GLN	73.9	64.9	86.3	53.2	34.8	86.0
GLU	74.0	76.1	71.6	74.1	65.2	84.4
MET	81.0	67.4	87.4	89.0	70.0	92.1

TABLE VI. Percent identical χ_{1+2+3} in the paired protein data sets.

	Same Unit Cell			Diff. Unit Cell		
	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$	All	$\delta_{\min} < 4$	$\delta_{\min} > 4$
ARG	58.6	51.2	73.6	32.3	26.2	72.4
LYS	42.6	41.5	44.2	40.2	34.0	52.8

TABLE VII. Percent identical $\chi_{1+2+3+4}$ in the paired protein data sets.

	All	$\delta < 4$	$\delta > 4$
ARG	233	128	105
LYS	309	169	140
GLN	203	105	98
GLU	311	167	144
MET	76	27	49
ASN	262	120	142
ASP	297	130	167
ILE	294	80	214
LEU	435	128	307
TRP	78	31	47
HIS	107	58	49
PHE	205	64	141
TYR	175	75	100
SER	287	122	165
CYS	30	3	27
THR	320	149	171
VAL	333	98	235
PRO	235	112	123

TABLE VIII. Number of residues in the side chain prediction data set.

	Crystal			Unimolecular		
	All	$\delta < 4$	$\delta > 4$	All	$\delta < 4$	$\delta > 4$
ARG	91.8	88.3	96.2	86.7	79.7	95.2
LYS	87.4	84.6	90.7	87.7	83.4	92.9
GLN	87.7	85.7	89.8	81.8	75.2	88.8
GLU	82.3	87.4	76.4	68.2	66.5	70.1
MET	88.2	74.1	95.9	92.1	85.2	95.9
ASN	87.0	83.3	90.1	80.9	70.8	89.4
ASP	87.2	90.8	84.4	80.1	78.5	81.4
ILE	99.0	97.5	99.5	98.0	93.8	99.5
LEU	98.2	96.1	99.0	97.5	93.8	99.0
TRP	100.0	100.0	100.0	98.7	96.8	100.0
HIS	94.4	94.8	93.9	88.8	89.7	87.8
PHE	99.5	98.4	100.0	98.0	93.8	100.0
TYR	97.7	97.3	98.0	96.6	94.7	98.0
SER	74.6	74.6	74.5	73.9	70.5	76.4
CYS	96.7	66.7	100.0	96.7	66.7	100.0
THR	90.6	89.9	91.2	90.9	89.9	91.8
VAL	96.4	95.9	96.6	95.2	91.8	96.6
PRO	97.0	95.5	98.4	96.6	94.6	98.4

TABLE IX. Percent correct χ_1 in the side chain prediction data set, using the $\pm 40°$ criterion, with and without inclusion of the crystal environment.

	Crystal			Unimolecular		
	All	$\delta < 4$	$\delta > 4$	All	$\delta < 4$	$\delta > 4$
ARG	82.0	79.7	84.8	76.8	68.8	86.7
LYS	76.4	75.2	77.9	74.1	68.6	80.7
GLN	73.9	66.7	81.6	68.0	57.1	79.6
GLU	70.1	77.8	61.1	51.1	47.9	54.9
MET	75.0	66.7	79.6	76.3	70.4	79.6
ASN	67.2	65.0	69.0	59.2	47.5	69.0
ASP	71.0	66.9	74.2	64.7	56.1	71.3
ILE	94.6	90.0	96.3	93.5	86.2	96.3
LEU	93.1	89.1	94.8	92.0	85.2	94.8
TRP	100.0	100.0	100.0	94.9	87.1	100.0
HIS	83.2	84.5	81.6	72.9	70.7	75.5
PHE	98.5	98.4	98.6	97.6	93.8	99.3
TYR	97.7	97.3	98.0	96.6	94.7	98.0
PRO	91.5	88.4	94.3	91.5	88.4	94.3

TABLE X. Percent correct χ_{1+2} (i.e., both χ_1 and χ_2 identical) in the side chain prediction data set.

	Crystal			Unimolecular		
	All	$\delta < 4$	$\delta > 4$	All	$\delta < 4$	$\delta > 4$
ARG	62.7	60.2	65.7	54.1	46.9	62.9
LYS	63.8	59.2	69.3	61.5	52.7	72.1
GLN	58.1	55.2	61.2	50.2	39.0	62.2
GLU	57.2	65.9	47.2	43.1	41.3	45.1
MET	59.2	51.9	63.3	61.8	59.3	63.3

TABLE XI. Percent correct χ_{1+2+3} in the side chain prediction data set.

	Crystal			Unimolecular		
	All	$\delta < 4$	$\delta > 4$	All	$\delta < 4$	$\delta > 4$
ARG	56.6	53.1	61.0	46.8	37.5	58.1
LYS	38.5	34.9	42.9	36.6	30.8	43.6

TABLE XII. Percent correct $\chi_{1+2+3+4}$ in the side chain prediction data set.