

Protein sidechain conformer prediction: a test of the energy function

Robert J Petrella¹, Themis Lazaridis¹ and Martin Karplus^{1,2}

Background: Homology modeling is an important technique for making use of the rapidly increasing number of protein sequences in the absence of structural information. The major problems in such modeling, once the alignment has been made, concern the positions of loops and the orientations of sidechains.

Although progress has been made in recent years for sidechain prediction, current methods appear to have a limit on the order of 70% in their accuracy. It is important to have an understanding of this limitation, which for energy-based methods could arise from inaccuracies of the potential function.

Results: A test of the CHARMM function for sidechain prediction was performed. To eliminate the multiple-residue search problem, the minimum energy positions of individual sidechains in ten proteins were calculated in the presence of all other sidechains in their crystal orientations. This test provides a necessary condition that any energy function useful for sidechain placement must satisfy. For $\chi_1 \times \chi_2$ rotations, the accuracies were 77.4% and 89.5%, respectively, and in the presence of crystal waters were 86.5% and 94.9%, respectively. If there was an error, the crystal structure usually corresponded to an alternative local minimum on the calculated energy map. Prediction accuracy correlated with the size of the energy gap between primary and secondary minima.

Conclusions: The results indicate that the errors in current sidechain prediction schemes cannot be attributed to the potential energy function *per se*. The test used here establishes a necessary condition that any proposed energy-based sidechain prediction method, as well as many statistically based methods, must satisfy.

Introduction

Methods for the prediction of the native structure of a protein from the amino acid sequence make use of statistical data, empirical energy functions, or a combination of the two. The problem is usually separated into prediction of the mainchain structure and prediction of the sidechain conformations, given that the mainchain is known. Considerable progress in the prediction of sidechain conformations has been made in recent years through a variety of methods. These include general clustering approaches [1], hierarchical searches [2,3], simulated annealing [4–7], molecular dynamics refinement [8], genetic algorithms [9,10], neural network methods [11], systematic elimination of incompatible sidechains [12–17], homology modeling or detailed rotamer-library-based techniques [18–27], iterative, energy-based rotamer searches [28], and mean-field optimization [29–31]. Many of the prediction algorithms incorporate elements of more than one approach and utilize both energetic and statistical techniques.

Because these sidechain prediction methods have been used with different proteins and with different criteria for correct prediction, it is difficult to make quantitative

Addresses: ¹Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA. ²Laboratoire de Chimie Biophysique Institut le Bel, Université Louis Pasteur, 67000 Strasbourg, France.

Correspondence: Martin Karplus
E-mail: marci@tammy.harvard.edu

Key words: conformation, energy, prediction, protein, sidechain

Received: 09 July 1998
Accepted: 21 July 1998

Published: 30 September 1998
<http://biomednet.com/elecref/1359027800300353>

Folding & Design 30 September 1998, 3:353–377

© Current Biology Ltd ISSN 1359-0278

comparisons. The prediction accuracy for many different methods, however, as measured by the mean percentage of residues with χ_1 and χ_2 angles correctly predicted, ranges from 50% to 70% for all residues, and from 70% to 80% for the core, or buried, residues. These accuracy ranges apply to both energy-based methods and statistical approaches for predictions based on the exact mainchain geometry. Although it has been pointed out that these sidechain prediction results are better than those obtained for more general protein structure prediction problems [32], there is still a question as to why even better results have not been obtained. Use of a template from a related protein or predicted backbone coordinates has been shown to lead to a further reduction in prediction accuracy [6,15,18,20,27–29,33]. Part of the limited success could be a result of the presence of different sidechain conformers with essentially the same free energy. This is almost certainly a factor for the exposed sidechains, for which the crystal environment may select one of several possible conformers. All recent studies are based on isolated proteins, although the crystal environment was considered in the early work of Gelin and Karplus [34]. Solution NMR results [35–37] indicate that many of the exposed

sidechains are disordered. In addition, disorder is commonly observed for surface residues in protein crystals, and conformational heterogeneity or 'discrete disorder' in what are otherwise well-ordered regions of molecules has been shown to occur in 6% of sidechains or more [38]. It is also possible that there are multiple sidechain conformers at what are thought to be ordered sites, even in the interior of proteins, or that there are errors in the crystal structure.

For the methods that use energy calculations as part of the prediction algorithm, there are several possible explanations for the limitations in accuracy that involve errors in the calculations. First, the potential energy functions may not be accurate enough. Second, the conformational search may be inadequate. Given the large number of conformational possibilities for a set of protein sidechains [4,29], it is possible that the searches have not converged to the global energy minimum, even if the energy functions are sufficiently accurate. The use of rotamer libraries introduces possible limitations because large deviations from rotameric χ values have been shown to occur systematically in protein crystal structures [39]. Bower *et al.* [27] report that ~6% of the > 40,000 residues studied had either χ_1 or χ_2 values that were not within $\pm 40^\circ$ of any rotamer in their library. Third, other factors, such as entropic effects and solvation, which are often ignored in energy-based methods, may significantly influence prediction.

The current work focuses on the possible limitations of the potential energy function for structure prediction in vacuum. Although it is difficult to determine whether the accuracy of the energy function is sufficient to obtain results that are better than those cited above, it is possible to examine a necessary condition that an adequate energy function must satisfy. The test used in this study can be described as follows: given that all sidechains except the one being examined have their known (X-ray) conformation, does a full, rigid-rotation search for the remaining sidechain lead to an energy map with the correct minimum energy conformation? Such a minimal test was made for the bovine pancreatic trypsin inhibitor (BPTI) when the energy function used in an early version of what is now the CHARMM program [40] was first introduced [34]. It was found that varying one sidechain angle at a time, 27 out of a total of 36 sidechains were well-predicted (42 out of 58 dihedral angles, or 72.4%, correct in this group), and nine residues were poorly predicted (10 of 37 angles, or 27.0%, correct), where the criterion for a correct prediction was that the minimum energy position was within the same local minimum as that of the crystal position ($\Delta E < 0.5$ kcal/mol). All nine of the incorrectly predicted sidechains were found to be charged, to be located on the surface of the protein, and to have very flat dihedral angle energy maps in the isolated protein. When crystal neighbors and crystal solvent molecules were introduced into the calculations, there was an improvement in the prediction of

four of these nine residues, in addition to improvements for several other residues, all of which were polar and largely exposed. A similar test was carried out more recently by Wilson *et al.* [28] for α -lytic protease, a protein with 142 amino acid residues. They used the AMBER non-bonded energy plus a solvation term [41]. The results indicated an 89% correspondence of predicted sidechain positions with a 'lowest coordinate error structure', which was the structure closest to the native protein structure that could be created from the set of rotamers employed in the study. Although the work of Gelin and Karplus [34] and Wilson *et al.* [28] suggests a high accuracy for the energy function, both studies were limited to a single protein. Moreover, the use of a rotamer library by Wilson *et al.* makes the results less clear.

The potential functions have improved since the original Gelin and Karplus [34] investigation, and data for many proteins are now available. In this investigation, the same type of test performed in that study is applied to a series of proteins with the polar hydrogen (PARAM19) model [42] used in CHARMM. Sidechains are rotated individually, in a full, systematic conformational search about each χ_1 angle alone in the first set of calculations and then around χ_1 and χ_2 simultaneously for each sidechain in the second set of calculations. Absolute minima and relative minima are located for each sidechain and compared to the native crystal positions. For thermolysin, the calculations are repeated in the presence of crystal water molecules and for the quenched (energy-minimized) structure, and incorrectly predicted residues are analyzed individually.

Results

Here, we present the general sidechain prediction results for each of the following: X-ray coordinates in vacuum (10 proteins), X-ray coordinates in the presence of crystal waters (thermolysin only), and quenched structure in vacuum (thermolysin only). To obtain insights into the origin of the errors, an analysis of certain energetic factors is made. Finally the results for individual residue types are examined.

General results

Protein X-ray structures in vacuum

χ_1 rotations. Absolute minima in the energy surfaces mapped over χ_1 angular space predicted the experimental χ_1 angles in 90.4% of the sidechains for all proteins. For core residues, the prediction accuracy was 96.7% (555/574; see Table 1).

When the proteins were broken down into small (BPTI, crambin, 434cro and CTF) and large proteins, the two groups were predicted with similar accuracy. Overall, 158 out of 174 residues (90.8%) were correctly predicted in the small-protein group, compared with 874 out of 968 correct (90.3%) in the large-protein group. In the core, all 62

Table 1**Number of correctly predicted sidechain dihedral angles by protein (χ_1 rotations only).**

Protein PDB code	χ_1^\dagger	%	χ_1 core [†]	%
5pti	36/42	85.7	15/15	100.0
1crn	31/32	96.9	11/11	100.0
2cro	48/54	88.9	20/20	100.0
1ctf	43/46	93.5	16/16	100.0
4fxn	98/115	85.2	57/58	98.3
1hiv	140/154	90.9	80/84	95.2
1lz1	92/103	89.3	49/50	98.0
3app	226/247	91.5	131/136	96.3
3rn3	95/105	90.5	43/46	93.5
3tin	223/244	91.4	133/138	96.4
Totals	1032/1142	90.4	555/574	96.7

Results for χ_1 mappings, with all other geometric variables fixed.

[†]Numbers of correctly predicted χ_1 angles and the total χ_1 angles in the protein are given.

residues were correctly predicted in the small-protein group, compared with 493 out of 512 (96.3%) correct in the large-protein group.

χ_1 and χ_2 rotations. When the lowest energy minima were calculated for each sidechain by simultaneously varying both χ_1 and χ_2 , prediction accuracy fell slightly. The fraction of torsion angles correctly predicted averaged 86.8% for χ_1 (991 out of 1142), 77.1% for χ_2 (606 out of 786), 84.1% (564/671) for $\chi_2|\chi_1$, and 77.4% for χ_{1+2} (884 out of 1142), overall, where $\chi_2|\chi_1$ refers to the fraction of residues having χ_2 correct given that χ_1 is correct, and where χ_{1+2} refers to residues whose χ_1 and χ_2 angles are both correctly predicted or, in the case of residues with only one heavy-atom χ angle (such as valine), whose single angle is correct. For the core residues, the predictions were: 94.9% correct for χ_1 (545 out of 574), 89.4% correct for χ_2 (355 out

of 397), 91.8% for $\chi_2|\chi_1$ (347 out of 378), and 89.5% correct for χ_{1+2} (514 out of 574). Results for the individual proteins are given in Table 2.

Prediction accuracy for the $\chi_1 \times \chi_2$ rotations is better for the large proteins: 761/968 correct in large proteins (78.6%) compared with 123/174 correct in small proteins (70.7%), a statistically significant difference by χ^2 ($p = 0.02$). This is mainly a result of the fact that the ratio of core residues to total residues increases with the size of a globular protein. In the cores, the results are not significantly different.

In most cases, prediction errors resulted from the fact that, although there was a local minimum on the energy map at the X-ray value, there was another minimum in the map with a lower energy. A map of potential energy versus χ_1 for one of the seven incorrectly predicted core serine residues (Ser102 in thermolysin) reveals a pattern of three minima, with the crystal structure corresponding to a secondary minimum (Figure 1). As has been pointed out previously [24,34,43–46], in the protein structure a minimum close to one of the minima of the isolated sidechain is generally selected. This topic is explored further in the section Energy maps and relative minima.

The prediction accuracy by residue type is presented in Table 3. Across all residues (exposed as well as core), the least well-predicted residues were lysine, glutamate, asparagine, aspartate, histidine and serine. In this group, 81.4% (83/102) of errors in χ_1 and 81.7% (98/120) of errors in χ_2 were outside the core. These results are similar to those that have been obtained in other studies [21,24,25,29], with a few notable exceptions.

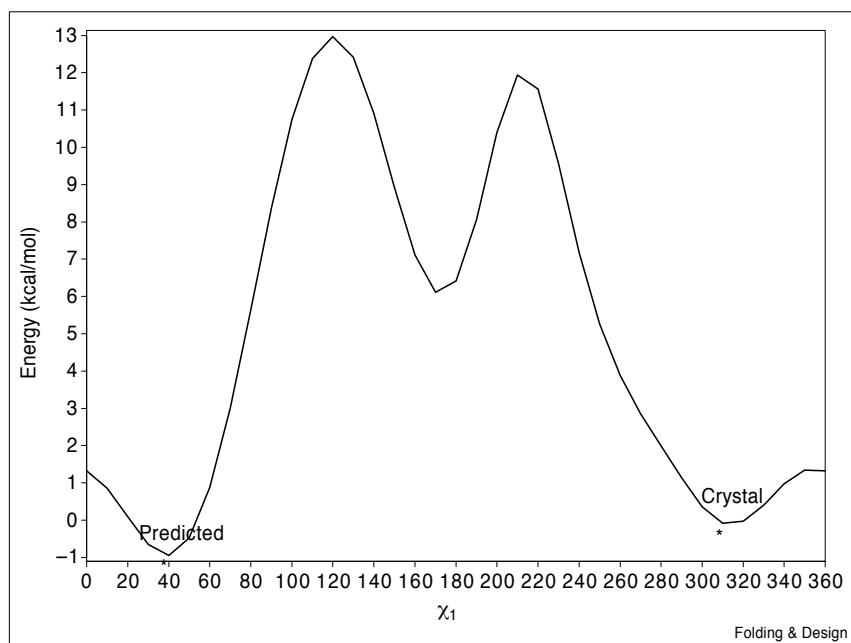
The accuracy of prediction of arginine is better in this study than in some others, possibly because in the current

Table 2**Number of correctly predicted sidechain dihedral angles for $\chi_1 \times \chi_2$ rotations, by protein.**

Protein PDB code	All residues			Core residues		
	χ_1	$\chi_2 \chi_1$	χ_{1+2}	χ_1	$\chi_2 \chi_1$	χ_{1+2}
5pti	33/42	18/24	27/42	14/15	9/10	13/15
1crn	32/32	9/16	25/32	11/11	4/4	11/11
2cro	43/54	26/32	37/54	20/20	13/15	18/20
1ctf	40/46	23/29	34/46	15/16	9/9	15/16
4fxn	87/115	49/65	71/115	53/58	33/39	47/58
1hiv	139/154	90/104	125/154	80/84	51/56	75/84
1lz1	88/103	55/62	81/103	49/50	31/31	49/50
3app	223/247	126/137	212/247	130/136	91/95	126/136
3rn3	90/105	43/51	82/105	43/46	23/25	41/46
3tin	216/244	125/151	190/244	130/138	83/94	119/138
All ratio	991/1142	564/671	884/1142	545/574	347/378	514/574
percentage	86.8	84.1	77.4	94.9	91.8	89.5

Results for $\chi_1 \times \chi_2$ mappings. $\chi_2|\chi_1$, the fraction of χ_2 angles correct given that χ_1 is correct. χ_{1+2} , the number of residues with both dihedral angles correct or, in the cases of valine, threonine, serine and cysteine, only the single angle correct, over the total number of rotatable residues.

Figure 1



Energy map for Ser102 in thermolysin. The three local minima are of note, as is the correspondence of the crystal position to the secondary minimum.

work, χ_3 and χ_4 are fixed at the 'correct' or X-ray values. In full prediction studies, χ_3 and χ_4 of the longer sidechains such as lysine and arginine should be determined without prior knowledge of the particular target X-ray structure. Lee and Subbiah ([4]; simulated annealing) and Hwang

and Liao ([11]; neural networks) allow these angles to vary more or less continuously; other investigators have set some or all of them arbitrarily to the *trans* configurations ([9,24,29]; R. Dunbrack, personal communication); and still others have allowed them to vary in accord with the

Table 3

Number of correctly predicted residues for $\chi_1 \times \chi_2$ rotations by residue type.

Amino acid type	All residues					Core residues				
	Number of residues*	χ_1^\dagger	% [‡]	$\chi_2 \chi_1^{\dagger\$}$	% [‡]	Number of residues*	χ_1^\dagger	% [‡]	$\chi_2 \chi_1^{\dagger\$}$	% [‡]
Cys	8	8	100.0			6	6	100.0		
Cys-DS	30	26	86.7			24	22	91.7		
Val	101	98	97.0			76	75	98.7		
Thr	107	100	93.5			37	37	100.0		
Ser	110	88	80.0			34	27	79.4		
Leu	97	95	97.9	81	85.3	75	73	97.3	63	86.3
Ile	91	86	94.5	80	93.0	68	65	95.6	63	96.9
Asp	86	68	79.1	52	76.5	36	33	91.7	27	81.8
Asn	76	59	77.6	43	72.9	29	27	93.1	25	92.6
Glu	62	45	72.6	34	75.6	19	13	68.4	12	92.3
Gln	73	59	80.8	51	86.4	16	15	93.8	14	93.3
Met	21	20	95.2	18	90.0	15	15	100.0	14	93.3
Phe	52	52	100.0	52	100.0	40	40	100.0	40	100.0
Tyr	66	64	97.0	61	95.3	39	39	100.0	39	100.0
Lys	73	48	65.8	31	64.6	14	14	100.0	11	78.6
Arg	52	41	78.8	34	82.9	23	22	95.7	20	90.9
His	18	15	83.3	10	66.7	10	9	90.0	6	66.7
Trp	19	19	100.0	17	89.5	13	13	100.0	13	100.0
All	1142	991		564			574		545	347

*The total number of residues of the particular type in all of the proteins. $^\dagger\chi_1$ and χ_2 are the number of χ_1 and χ_2 angles correctly predicted. ‡ The percentage correct within that subset of angles. $^\S\chi_2|\chi_1$ is defined in Table 2.

rotamer library being used [7,17,25]. Because of the disorder commonly present in the more peripheral atoms of arginine and lysine, χ_3 and χ_4 entries for these residues were not made in the original rotamer library of Ponder and Richards [1], and arginine and lysine were excluded from the predictions in that work. Other investigators have set the χ_4 angles to the known X-ray values [15]. In the current work, the systematic search makes prohibitive the computational time required for explicit exploration of all heavy-atom dihedral angles. In evaluating a necessary condition that the energy function must satisfy, the introduction of possible errors by the use of arbitrary χ_3 and χ_4 angles would be inappropriate.

Histidine was more accurately predicted in rotamer library-based predictions [24], but χ_2 and $\chi_2 + 180^\circ$ conformers were considered equivalent (as in some other studies) whereas here they were not. Using this less stringent criterion, none (out of 10) of the χ_2 values in core histidines was incorrect. (If the criterion is applied to all histidine and asparagine residues in the study, the accuracies for all residues rise by $\sim 0.8\%$ both overall and in the core.) Threonine was poorly predicted in the simulated annealing studies of Lee and Subbiah [4], but the potential energy function used in that study (as in several others) did not include electrostatics.

For the core regions of the proteins, although the prediction accuracies are generally improved, the same residue types are the most poorly predicted, with the single exception that leucine is also badly predicted (Table 3). Of course, a smaller fraction of the charged and polar residues are in the core, but there are still enough for meaningful statistics. Of the core residues, 45.4% are polar or charged, a finding consistent with previous studies [47]. A more detailed treatment of specific error patterns is given in the section Analysis of individual residue types, below.

As shown in Table 4, although the cut-off for 'correct' dihedral angles was set at $\pm 40^\circ$, the overall errors ($\Delta\chi_1$ and $\Delta\chi_2$) of the correct dihedrals were generally much smaller than this value, with a mean \pm s.d. of $7.1^\circ \pm 7.5^\circ$ and a

median of 4.5° for all χ_1 and $8.2^\circ \pm 8.7^\circ$ and a median of 5.0° for all χ_2 . In the core, the errors were even smaller for both χ_1 and χ_2 ($p < 0.001$ by t-test for both). Hence, locally there is very good agreement between energy contours and crystal positions.

This raises the question of how far the prediction accuracy would be lowered using a stricter criterion for the 'correct' predictions of dihedral angles. Table 4 shows that the accuracy decreases more rapidly at lower cut-offs, consistent with a narrow spread in the angular deviations. As expected, the error in χ_2 is consistently greater than that for χ_1 , in terms of both accuracy and average deviation. This corresponds to what is generally found in prediction methods, and probably relates to the fact that χ_1 rotations involve more atoms (γ position) and larger displacements than do χ_2 rotations, leading to higher energy changes. As is discussed below in the section Energy maps and relative minima, more rugged energy maps are associated with better predictions.

Rmsd values between the sidechain atoms in the X-ray structures and the predicted structures are shown in Table 5. Across all residues (heavy-atom) deviations ranged from 0.83 Å for crambin to 2.39 Å for BPTI. The rmsd for all sidechain atoms in the entire set of proteins was 1.46 Å. In the structural core, the rmsd values were significantly lower.

The remainder of the results section will refer to the $\chi_1 \times \chi_2$ set, unless otherwise specified.

Inclusion of crystal waters

As in the work of Gelin and Karplus [34], prediction accuracy in the presence of crystal solvent was significantly improved over that for the vacuum calculations. For the thermolysin test system, the difference was ~ 8.6 percentage points, both overall and in the core ($p < 0.02$ by χ^2 for both). The results are summarized in Table 6. If twofold symmetry is assumed in χ_2 of asparagine and histidine, the (+)solvent predictions are 87.7% correct overall and 94.9% in the core.

Table 4

Accuracy of predictions in the set of 10 proteins by cut-off.

Cut-off	All residues				Core residues			
	χ_1		χ_2		χ_1		χ_2	
	Correct (%)	Mean Δ^*	Correct (%)	Mean Δ^*	Correct (%)	Mean Δ^*	Correct (%)	Mean Δ^*
40°	86.8	7.11	77.1	8.16	94.9	5.41	89.4	6.30
30°	84.2	6.27	73.8	6.93	92.9	4.79	88.4	5.97
20°	80.4	5.44	68.7	5.63	91.3	4.46	84.9	5.23
10°	67.8	3.85	56.4	3.85	83.6	3.58	73.3	3.87

Accuracy of vacuum predictions (in all proteins) computed with various cut-off values for the deviation allowed from the crystal positions. *The mean deviation for correctly predicted angles only.

Table 5**Rmsd values (Å) between predicted and crystal structures for $\chi_1 \times \chi_2$ rotations.**

Protein PDB code	All residues	Core residues
5pti	2.39	1.29
1crn	0.83	0.20
2cro	2.14	0.67
1ctf	1.34	0.82
4fxn	1.84	0.91
1hiv	1.51	0.71
1lz1	1.46	0.38
3app	0.89	0.50
3rn3	1.49	0.74
3tln	1.29	1.08
All	1.46	0.80

Rmsd values are for sidechain heavy atoms (not including C β). In each case, the rmsd is calculated for all atoms in the given category (i.e. no averaging over residues or structures).

Of the 19 core residues that were initially incorrect, only six remain incorrect upon addition of the crystal solvent, including three leucine residues and Val79, which appears to be a crystallographic error (see below in the section Atom position replacement error pattern). For these six residues, $\Delta E_{\text{pred-crys}} = -3.2 \pm 1.9$ kcal/mol (mean \pm s.d.). Three buried histidine residues whose predicted χ_2 orientations differed from the crystal structure by 180° (imidazole ring flip) in the vacuum calculations are correctly predicted in the presence of the waters, suggesting that the crystal structure orientations are the correct ones. Interestingly, addition of solvent generates a new prediction error in Leu144, also involving a conformer pair related by an exchange of atomic positions ($\Delta E_{\text{pred-crys}} = -2.0$ kcal/mol; see the section Atom position replacement error pattern, below).

Table 6**Prediction accuracy in thermolysin for various conditions.**

	X-ray (-)solv*	Prediction accuracy (%)	X-ray (+)solv†	Prediction accuracy (%)	Quenched (-)solv‡	Prediction accuracy (%)
All residues:						
χ_1	216/244	88.5	228/244	93.4	241/244	98.8
$\chi_2 \chi_1$	125/151	82.8	147/164	89.6	162/169	95.9
χ_{1+2}	190/244	77.9	211/244§	86.5	234/244	95.9
$\Delta\chi_1$	7.5 (8.5)		6.5 (7.5)		3.2 (2.9)	
$\Delta\chi_2$	7.4 (7.5)		6.5 (6.2)		3.6 (2.7)	
Core residues:						
χ_1	130/138	94.2	136/138	98.6	138/138	100.0
$\chi_2 \chi_1$	83/94	88.3	94/99	94.9	96/99	97.0
χ_{1+2}	119/138	86.2	131/138§	94.9	135/138	97.8
$\Delta\chi_1$	6.1 (7.2)		4.9 (5.8)		2.9 (2.9)	
$\Delta\chi_2$	6.5 (5.8)		5.3 (4.1)§		3.5 (2.7)	

*Vacuum calculations on the native structure. †Native structure in presence of available crystal solvent. ‡Quenched structure in vacuum. χ_1 , the number of correctly predicted χ_1 dihedral angles, expressed as the number of correct angles/total number of angles. $\chi_2|\chi_1$, the number of χ_2 angles correct, given that χ_1 is correct. χ_{1+2} , the number of residues having both dihedral angles correct or the single angle correct

It is interesting that the presence of crystal waters also affects polar residues in the core. For example, Ser102 and Glu166 are 100% buried, and yet they require the presence of solvent for accurate prediction.

Use of minimized coordinates

As expected, the predictions based on the quenched (900-step minimized) thermolysin coordinates were significantly more accurate (95.9%) than those based on the unrelaxed X-ray coordinates (77.9%; see Table 6). The three errors in the core of the quenched structure (two leucines and one histidine) involved pairs of conformers related by an exchange of atomic positions. Notably, in 61 out of 244 residues (25.0%) overall and in 37 out of 138 residues (26.8%) in the core, the predicted orientations for the sidechains were higher in energy in the quenched structure than in the native structure, indicating that minimization does not guarantee stabilization of every sidechain position. The sidechains that were higher in energy upon minimization were disproportionately non-polar (34 out of 61 or 55.7% overall and 26 out of 37 or 70.3% in the core; in 3tln, 29.1% of rotatable sidechains are non-polar overall, 40.6% in the core). This suggests that quenching of a protein favors the optimization of the energetics of polar or charged residues over that of non-polar residues.

A sorting of the prediction results by solvent exposure reveals that the bulk of the discrepancies between X-ray-structure-based and relaxed-structure-based predictions did indeed occur in exposed residues. Of the 48 residues that were well-predicted only upon relaxation of the structure, 32 were in the exposed regions (out of 106 total exposed residues) and 16 were in the buried or core

for small residues. $\Delta\chi_1$ and $\Delta\chi_2$, the error of the correct dihedral angles in degrees, expressed as a mean with standard deviation in parentheses. §A statistically significant difference ($p < 0.05$) between the (+)solv and (-)solv results (first and second sets). The differences between the results for the quenched structure and those for the (-)solv vacuum structure (first and third sets) are all significant.

regions (out of 138). This result indicates that, in the absence of solvent and intermolecular contacts, the exposed sidechains in crystal structures are often distant from both their energy-minimized and predicted conformations. Consequently, the local minimum obtained by energetic relaxation significantly alters the sidechain conformation in the exposed regions in many cases (to a much greater extent than it does in the core regions). In the exposed regions, energetic relaxation mainly involves an optimization of intramolecular hydrogen bonding. Because the predictability of charged and polar sidechains relates directly to the extent of such hydrogen bonding (see the section Analysis of individual residue types), the prediction is more accurate in the quenched structure.

When sidechains were grouped by whether they had changed (40° or more in at least one dihedral angle) or had not changed upon quenching of the crystal, all the sidechains that had changed were found to have been subsequently well-predicted (56 overall and 25 in the core). Of the unchanged residues, the prediction accuracy was 178/188 or 94.7% (in comparison with 77.4% for the unrelaxed crystal structure). Three of the errors in this group were new.

It is clear from these findings that in the great majority of cases, energy minimization, at least in vacuum, results in a significantly higher correspondence of individual sidechain positions to the absolute minima of their single-residue energy landscapes than in the native crystals. When the native sidechain position is close to the absolute energy minimum, minimization simply stabilizes the native orientation; when it is not, as in many exposed residues, minimization results in significant shifts, which in most cases align it with the absolute minimum.

Effect of dielectric factor

In a set of preliminary calculations varying the dielectric factor, the best results were obtained with the distance-dependent dielectric (rdie), and although differences were not statistically significant, rdie was selected for the main set of calculations. Using the whole thermolysin molecule in vacuum as a test system, however, a more extensive set of calculations was performed to examine the importance of the dielectric parameter. Even with this size sample (244 residues) there were no statistically significant differences between any of the results. The dielectric form, constant, and associated accuracies (percentage correct overall and in the core) were: rdie 1.0, 77.9, 86.2; cdie 1.0, 75.8, 84.1; cdie 2.5, 78.3, 85.5; cdie 5.0, 78.7, 85.5; cdie 10, 78.3, 85.5; cdie infinity (no electrostatic contribution) 76.2, 85.5. Hence, in the exposed regions there is a slight trend towards worse prediction with a dielectric constant of 1 or infinity, which have associated accuracies of 65.1% and 64.2%, respectively, compared with 69.8% for a dielectric constant of 5. The findings indicate that although the

electrostatic term is probably important for sidechain prediction, a global adjustment of the dielectric by itself is inadequate to compensate for solvent effects.

Use of original crystal coordinates

Exploration of a possible biasing effect of the initial 25-step minimization of the structures was carried out by repeating the $\chi_1 \times \chi_2$ vacuum calculations for thermolysin with the original (unaltered) crystal coordinates. The prediction accuracy was very similar to that in the calculations using the 25-step minimized structure. Overall, 186 out of 244 sidechains, or 76.2%, were well-predicted using the original coordinates, and in the core 119 out of 138, or 86.2%, were well-predicted. This is to be compared to 77.9% correct overall and 86.2% for the core in the 25-step minimized case. The average angular deviations for correct χ_1 angles was 8.1° overall and 6.4° in the core, which compare to 7.1° and 5.4° , respectively, for the 25-step minimized coordinates. Thus, the effect on the predictions of the slight initial minimization appears to have been very small.

As a separate test, the effect of changing the standard of comparison on the same prediction results was explored. The predictions based on the 25-step minimized structure were reanalyzed using the original crystal coordinates (rather than the minimized coordinates) as the standard. (In the main set of vacuum calculations, the 25-step minimized starting structures were used as the standard of comparison.) This substitution changed the results by $<1\%$. Using the X-ray structure as a standard, 76.8% of all sidechains were correct and 77.4% were correct using the relaxed structure as a standard. In the core, the results were 89.4% and 89.5%, respectively. The average angular deviations of the correctly predicted χ_1 angles from the correct positions were 8.7° and 7.1° , respectively.

Energy maps and relative minima

The following results refer to the vacuum calculations for the crystal structures (25-step minimized), unless otherwise specified.

Correspondence of crystal structures with relative energy minima

As described in the Materials and methods section, a quantitative analysis of the relative minima was carried out for thermolysin in vacuum. The pattern for most residues was to have a few minima within ~ 5 kcal/mol of the primary minimum, and for exposed residues to have more numerous low-energy minima than core residues. Table 7 lists the energies of all of the calculated local minima for four representative residues in thermolysin (buried-polar, buried-non-polar, exposed-polar and exposed-non-polar).

As previously noted, in both core and exposed regions, when residues were incorrectly predicted the crystal

structure generally corresponded to one of the relative minima. In 13 out of the 19 incorrectly predicted residues in the core of thermolysin, the crystal structure was located within $\pm 17.5^\circ$ of the χ_1 value and 22.8° of the χ_2 value of either the secondary minimum (11 cases) or the tertiary minimum (two cases). In five additional cases, the crystal structure is located within 17.5° of the χ_1 value and within 49° of the χ_2 value of one of the four lowest relative minima. The only exception, Val79, is discussed below.

It is notable that this correspondence of the crystal structure to a local minimum on the energy maps in the vacuum calculations occurred even for the disulfide-bonded cysteine residues, which were treated without the disulfide bond, and for residues bound to metal ions, which were treated without the ions. In the two instances of incorrect prediction of core cysteine residues (22 out of 24 were correct), the crystal orientation corresponded to a secondary minimum in the energy map ($\Delta E_{\text{pred-crys}} = -0.2$ kcal/mol and -2.4 kcal/mol). In the five cases of residues binding to metal ions in the core of thermolysin (four Ca^{2+} and one Zn^{2+}), the crystal conformers all lie within 41° of either the first or second minimum on the energy maps ($\Delta E_{\text{pred-crys}} = -2.4 \pm 0.4$ kcal/mol, mean \pm s.d.).

In the exposed regions of thermolysin, the incorrectly predicted residues (35 out of 106) show a pattern that is similar to that in the core regions, with the crystal orientations corresponding to subordinate minima on the energy maps. Compared to those in the core regions, however, the orientations in the exposed regions tend to correspond to lower-ranking minima on the maps, and the correspondence is less precise. Out of the 35 incorrectly predicted residues in the exposed regions, 21 (60.0%) have relative minima within 34.3° of the crystal orientation. These minima rank anywhere from second to seventh on the maps, except for the case of Asn280, which has a very flat and complex energy map, and where the crystal corresponds to the twelfth minimum ($\Delta E_{\text{pred-crys}} = -0.9$ kcal/mol). These data suggest that even in the exposed regions of the protein and even in the absence of intermolecular crystal contacts and solvent, there is some effect of low energy sidechain orientations. The observed crystal orientation is likely to be selected from these by interactions with particular crystal contacts and crystal water molecules.

Energy maps of the residues that were incorrectly predicted using the structure without solvent were compared with the maps of the same residues in the presence of crystal water molecules. In the presence of the waters, the local energy wells generally become narrower and steeper, and in some cases of polar residues multiple wells collapse to a single one (see Figure 2). In other instances, there is a deepening of a single well (the correct one) in a cluster of several. This occurs in three histidine residues, for which the correct χ_2 orientation is selected out of two possible

Table 7

Energies (kcal/mol) of all relative minima for four thermolysin residues.

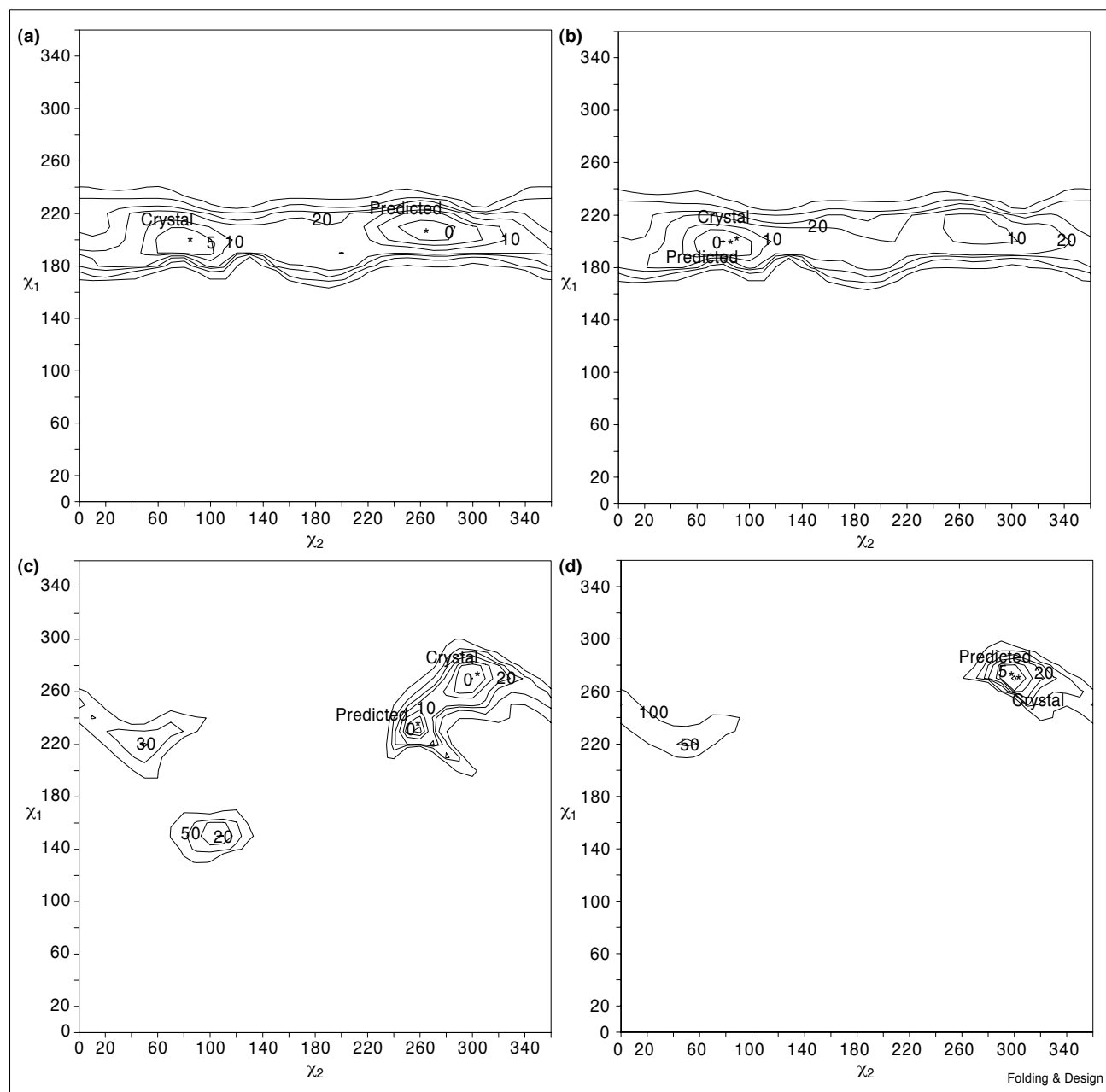
	Leu14 (E)	Leu30 (I)	Asn19 (E)	Asn33 (I)
1°	-7.8	-12.7	-14.6	-19.3
2°	-7.1	-8.4	-12.7	-15.8
3°	-6.2	-2.5	-9.9	-12.2
4°	-5.7	-0.2	-8.7	-10.8
5°	-4.2	0.4	-8.1	-9.9
6°	-3.3	2.7	-5.4	-6.8
7°	-3.1	449.5	-4.9	0.6
8°	5.4	481.8	-4.0	3.5
9°	52.7	738.1	93.7	36.0
10°	123.1	789.1	107.2	311.9
11°	163.3	1108.0	843.9	335.6
12°	167.6	1201.0	1061.5	
13°	182.2	1501.4	1118.3	
14°	241.6	1948.1	1149.0	
15°	251.9	1975.1		
16°	370.7	2081.7		
17°		2230.5		
18°		2413.7		
19°		2701.2		
20°		3189.2		
21°		8127.6		

List of all local minima, sorted by energy, occurring in the calculated energy maps of four representative residues in thermolysin. The first column simply designates the rank of the minimum. I, an interior or buried residue; E, an exposed residue. In each residue, note the abrupt transition from low-energy to high-energy conformers.

minima by the presence of solvent (see Figure 2a). For the incorrectly predicted leucine residues, the presence of crystal solvent does not substantially alter the maps.

The general correspondence of the calculated minima with the observed crystal structure positions implies a high accuracy of the latter, because it is unlikely that the coordinates would be incorrect and correspond by chance to local minima on the calculated energy surfaces. Furthermore, because the refinements of most of the X-ray structures in this study did not involve energy-based procedures (see the Materials and methods section), it is improbable that the correspondence between energy minima and crystal positions is an artifact of the refinement process. Hence, although the protein crystal structure is an 'average' structure, and there can be both static and dynamic disorder, the crystallographic conformers clearly do not represent averages of two or more alternative low-energy dihedral angle rotamers, but rather single low-energy rotamers in the vast majority of cases. It has been shown previously that where multiple conformers are present, the refinement method usually selects the one of highest probability rather than an average of different conformer positions [48,49]. Such an average position would lead to an intermediate position on the energy map rather than one near a minimum as the crystal conformers did throughout this study. These findings are relevant to

Figure 2



Energy maps for His216 and Glu166 of thermolysin. His216 (a) in vacuum and (b) in the presence of solvent molecules. Glu166 (c) in vacuum and (d) in the presence of solvent molecules. The crystal positions are set to an energy of zero.

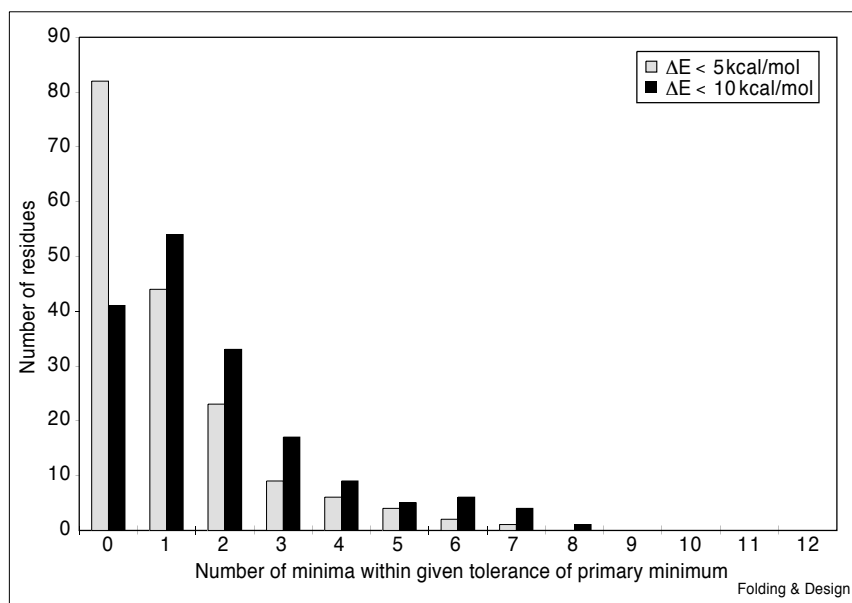
crystallographic data interpretation, because they generally support the feasibility of modeling a crystal structure as a single molecule.

'Ruggedness' of energy surfaces

To assess the 'ruggedness' of the energy maps, the number of relative minima that were within a given energy of the absolute minimum and the difference

between the primary and secondary minimum, ΔM_{2-1} , were determined. In general, these two measures were found to be in good agreement; the larger ΔM_{2-1} , the smaller the number of low-energy minima. In order for comparisons to be meaningful, residues having only one heavy-atom dihedral (serine, valine, threonine and cysteine) were excluded from this part of the analysis, leaving 171 residues.

Figure 3



Histogram of the distribution of energies in the relative minima occurring in sidechain dihedral energy maps of thermolysin. The plot counts minima of rank two and lower for sidechains having at least two heavy-atom dihedral angles. Degenerate χ_2 conformers in phenylalanine and aspartate are excluded. 52% and 76% of all residues have at least one local minimum within 5 kcal/mol and 10 kcal/mol, respectively, of their absolute minima.

Figure 3 shows the distribution of residues by the number of relative minima within a tolerance of either 5 kcal/mol or 10 kcal/mol of the primary minimum. The counts do not include the degenerate minima in phenylalanine and aspartate. Most residues (130 out of 171, or 76.0%) have at least one alternative (subordinate) minimum within 10 kcal/mol, and most (145 out of 171 or 84.8%) have no more than three. A similar pattern holds when the cut-off is set at 5 kcal/mol or when degenerate phenylalanine and aspartate conformers are included. Even in the core, there are significant numbers of alternative minima, although they are less numerous than in the exposed regions. 50 out of 72 core residues, or 69.4%, have at least one alternative minimum within a 10 kcal/mol cutoff; at a cutoff of 5 kcal/mol, the number is 29 out of 72, or 40.3%.

It is significant that for many sidechains, in both the exposed and buried regions, there are multiple low-energy conformers even in the fixed protein environment, because this suggests that such a multiplicity of states exists in solution as well; that is, that torsional transitions can occur in all regions. Although the energy cut-offs have been set to relatively large values (5 kcal/mol and 10 kcal/mol) in this rigid-rotation study, the actual $|\Delta E|$ values are expected to be significantly smaller, because relaxation can occur around alternative conformers. These findings are consistent with molecular dynamics results suggesting that there are alternative conformers in sidechains for both buried and surface residues of most types [49,50], with findings of orientational heterogeneity for sidechain sites within the same crystal [38] or orientational variation for sidechains in different crystals of the same protein [27,51,52], and with

other results in the current investigation showing sidechain orientational differences in all regions of the nearly superposable subunits of the HIV-1 protease (see Comparison of HIV-PR-A and HIV-PR-B).

An analysis of the distribution of the relative (subordinate) minima with respect to residue characteristics is shown in Table 8. The median energy differences (ΔM_{2-1}) are given because the means are skewed by large values for residues having only one low-energy minimum. As in the work of Gelin and Karplus [34], core residues have significantly fewer low-energy minima than exposed residues ($p < 0.001$ by t-test), and the energy differences are much larger. The pattern is the same for correct and incorrect residues, with correct residues having significantly fewer minima and much larger energy differences. The 'flatter' dihedral maps for the exposed residues are expected and because most of the incorrect residues are also exposed, it is not surprising that these two groups share the same energy map characteristics.

But when finer subdivisions of the data are made into four groups consisting of correct core, correct exposed, incorrect core, and incorrect exposed, the correctly predicted residues in the exposed regions have more 'rugged' energy maps than the incorrectly predicted residues in the core; the median ΔM_{2-1} values are 4.3 kcal/mol and 2.7 kcal/mol, respectively. Thus, while it is true that, overall, the exposed regions have flatter energy maps than the core regions, these results suggest that predictability is more strongly linked to the separation of the relative minima in energy than to solvent accessibility by itself.

This is further demonstrated in Figure 4, which shows the percentage of correctly predicted residues as a function of ΔM_{2-1} . The bars indicate the percentage correct in each group and the lines above the bars indicate the cumulative percentage correct as each additional group is added (moving from higher energy to lower energy, or left to right). The numbers above the bars indicate the number of residues in each group. For all residues with $\Delta M_{2-1} > 6.0$ kcal/mol, the prediction accuracy is 93.2% (69 out of 74), and for core residues with $\Delta M_{2-1} > 5.0$ kcal/mol the accuracy is 96.6% (57 out of 59). Thus, regardless of solvent exposure, a large energy difference between primary and secondary minima correlates well with predictability.

Analysis of individual residue types

Most of the errors in the core predictions relate to either or both of two factors. One involves hydrogen bonding (polar and charged residues) and the other involves an interchange or replacement of atomic positions (primarily leucine, but also isoleucine, methionine, asparagine and histidine).

Hydrogen bonding error pattern: donor–acceptor collapse

For the polar and charged sidechains, the errors in χ angle prediction generally related to errors in hydrogen bonding. In some cases, the predictions involved hydrogen-bonded atom pairings that were different from those in the X-ray structures, as also observed by Wilson *et al.* [28]. In other cases, predictions involved intramolecular hydrogen-bonding distances that were shorter than those in the X-ray structures. For example, in the crystal structures, the hydroxyl groups in the seven incorrectly predicted core serine residues are uniformly oriented toward the periphery and/or crystal solvent molecules (six out of the seven are within 3.1 Å of crystal waters). In the vacuum predicted structures, the hydroxyl groups of these residues tend to be oriented away from solvent waters and towards

hydrogen-bond acceptors or donors in the protein. In some cases, what would have been a ‘weak’ or ‘near’ hydrogen bond between two protein atoms in the crystal structure becomes a shorter, more stable hydrogen bond in the vacuum calculations. This error pattern was seen in all of the rotatable polar (or charged) sidechains, both in the core and in the exposed regions. Generally, it appears to be a result of the absence in the calculations of crystal solvent molecules that would otherwise interact with these residues. The pattern is most pronounced in the exposed regions, where the absence of crystal waters sometimes leads to the prediction of intramolecular hydrogen bonds, between neighboring surface polar groups, that do not exist in the native crystal. For example, in the 1hiv crystal structure, Glu164 is surrounded by three positively charged or partially charged groups, namely the $-\text{NH}_3^+$ groups of Lys169 and Lys113 and the backbone amide $-\text{NH}$ group at Gly167, with the following O–N interatomic distances: $\text{O}\epsilon 1-\text{NH}_3^+(\text{Lys113}) = 4.71$ Å, $\text{O}\epsilon 1-\text{NH}_3^+(\text{Lys169}) = 3.56$ Å and $\text{O}\epsilon 2-\text{NH}(\text{Gly167}) = 3.86$ Å, and hydrogen-bond angles of 124.1°, 150.7° and 113.4°, respectively. Thus, there is a central charge surrounded by what are essentially three opposite charges, at interatomic distances (from the central atom) that are much greater than the corresponding van der Waals radii (in PARAM19, the sum of heavy-atom van der Waals radii in hydrogen-bonding pairs is modeled as $2\sigma = 2.85$ Å). On the basis of the internal electrostatics of the protein alone, one would expect the central charge to position itself as close as possible to one of the surrounding opposite charges. This is exactly the case in the predicted structure, where the glutamate carboxyl is shifted toward the Lys169 amino group, at a distance of 2.82 Å, and slightly closer to that of Lys113, at 4.27 Å, while the hydrogen bond with Gly167 has been lost. Thus, in cases such as these, what is a relatively unstable arrangement in vacuum for a cluster of charges has been stabilized in the crystal by water molecules (or protein–protein crystal contacts).

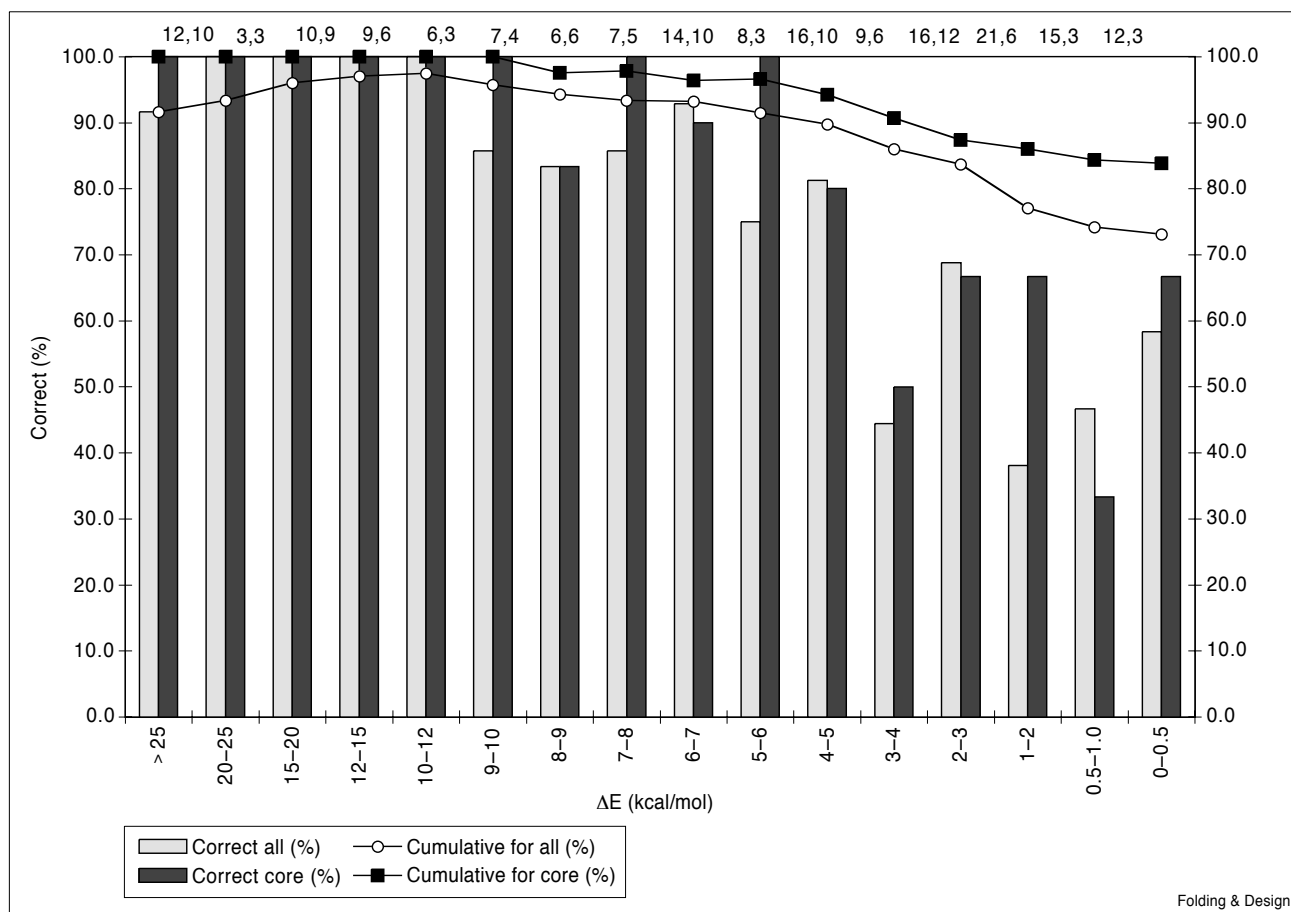
Table 8

Analysis of relative minima with respect to residue characteristics.

	Number of residues	Average number of minima*		Median	ΔM_{2-1} (kcal/mol) [†]	
		5 kcal/mol	10 kcal/mol		Mean	s.d.
All	171	1.1	1.8	4.6	12.7	38.2
Correct	125	0.6	1.3	6.6	13.0	23.9
Incorrect	46	2.3	3.4	1.7	11.7	62.6
Core	99	0.6	1.3	6.7	14.5	26.4
Exposed	72	1.6	2.7	2.0	10.1	50.1
Correct						
Core	83	0.4	1.0	7.7	16.7	28.3
Exposed	42	1.0	1.9	4.3	5.7	6.9
Incorrect						
Core	16	1.7	2.6	2.7	3.1	2.2
Exposed	30	2.6	3.8	1.4	16.3	77.6

Data include only residues with at least two heavy atom sidechain dihedral angles. *Average number of minima within given tolerance of the absolute energy minimum. [†]Energy difference between primary and secondary minima on the energy maps.

Figure 4



Prediction accuracy as a function of the energy gap between primary and secondary minima in thermolysin vacuum calculations. The bars represent the percentage correct within each energy category, and the lines represent the cumulative percentage correct for the union of that

category and all those preceding it (those with larger gaps). The numbers above the bars indicate absolute numbers of sidechains in each category. For gaps of ≥ 6 kcal/mol, accuracy is 93.2% overall and 96.4% for the core.

Predictions in several such cases are generally corrected in the presence of crystal solvent molecules (data not shown). The current findings clearly support the proposition [34] that for some polar or charged sidechains it is not possible to predict orientation solely on the basis of the internal energetics of the protein. In these cases, unless solvent effects are accounted for, calculations will generally lead to the prediction of nearest-neighbor donor-acceptor distances that are shorter than those in the crystal.

An analysis of hydrogen bonding in thermolysin corroborates these findings. Table 9 shows that the predictability of polar sidechains is strongly linked to the number of intramolecular hydrogen bonds between a given sidechain and its surroundings in the crystal structure. This holds true for both core and exposed regions. Because more extensive intramolecular hydrogen bonding of sidechains generally means less extensive hydrogen bonding of the

sidechains with crystal water, this result suggests that vacuum predictability is inversely related to the interaction with solvent.

As in other prediction studies [4,24,25,29], serine residues were poorly predicted. Seven of 34 (20.6%) core serines were incorrect. Given the much higher prediction accuracy for the other single-dihedral sidechain types, cysteine, valine and threonine (7.3%, 3.0% and 0.0% incorrect in the core, respectively), it appears that both the polarity and the small size of the serine sidechain contribute to its poor predictability.

Atom position replacement (spatial pseudosymmetry) error pattern

Leucine had the highest total number of errors in the core of any residue type, two in χ_1 and 12 in χ_2 . It has been noted previously that for leucine residues, very different

Table 9**Prediction accuracy versus number of intramolecular hydrogen bonds for polar sidechains in the thermolysin crystal structure.**

Residues	Number of hydrogen bonds						All
	0	1	2	3	4	5	
All polar							
Total	57	62	35	11	7	1	173
Incorrect	30	10	5	2	0	0	47
Incorrect (%)	52.6	16.1	14.3	18.2	0.0	0.0	27.2
Core polar							
Total	12	35	22	7	6	0	82
Incorrect	5	5	4	1	0	0	15
Incorrect (%)	41.7	14.3	18.2	14.3	0.0	(-)	18.3

Total, the total number of polar residues having the specified number of hydrogen bonds. For each residue, only hydrogen bonds between the sidechain (not backbone N or O) and its surroundings (any protein atom) are counted.

sets of χ angles can result in nearly coincident atom positions [4,24]. In this study, 14 out of the 16 poorly predicted Leu residues had near-superposition of exchanged atomic positions. In all 14 cases, the predicted structure varied from the X-ray structure in such a way that the $C\delta_1$ and the $C\delta_2$ had inverted positions, usually with a small shift in the $C\gamma$ atoms. A typical example is Leu243 in thermolysin, shown in Figure 5. There is a small shift of 33° in χ_1 and a 153° rotation of χ_2 between predicted and crystal orientation, such that $C\delta_{2[\text{pred}]}$ and $C\delta_{1[\text{cryst}]}$ are adjacent

($d = 0.62 \text{ \AA}$), $C\delta_{1[\text{pred}]}$ and $C\delta_{2[\text{cryst}]}$ are adjacent ($d = 0.40 \text{ \AA}$), and there is a shift in $C\gamma$ (0.74 \AA). The χ_2 values are 16.7° and 170° for the crystal and predicted orientations, respectively. The two-dimensional energy map for Leu243 is shown in Figure 6. There are two local minima, and the crystal corresponds to the conformation at the secondary minimum ($\Delta E = -4.6 \text{ kcal/mol}$). Most of the leucine energy maps showed a similar positioning of minima, separated by $\Delta\chi_1$ of $\sim 30\text{--}40^\circ$ and $\Delta\chi_2$ of $\sim 130\text{--}160^\circ$. This kind of pairing of superposable conformers in leucine has been noted previously [4,24].

An additional characteristic of the leucine conformer pairs, however, is that the ability to superimpose the reversed $C\delta$ positions of leucine depends on the starting position of χ_2 . The superposability of reversed $C\delta$ positions was calculated for different starting positions of χ_2 ($\chi_{2[\text{initial}]}$) as described in the Materials and methods section. The results are shown in Figure 7. For each starting value of χ_2 , the best superposition, D_{min} , that was found over the full conformational search is plotted (dark circles). In this 5° search grid, D_{min} is at a minimum ($\approx 0.12 \text{ \AA}$) when $\chi_{2[\text{initial}]} = 40^\circ$ or 195° and at a maximum (2.5 \AA) at $\chi_{2[\text{initial}]} = 110^\circ$ or 290° . This implies that at values of $\chi_{2[\text{initial}]}$ around 110° and 290° , no superposition of reversed atom positions is possible, whereas at $\chi_{2[\text{initial}]}$ values around 40° and 195° , very close superposition is possible. For reasons of inherent symmetry, the curve should, in principle, be symmetric with respect to the axes

Figure 5

Stereoview of superposable leucine rotamers. Predicted and crystal $C\gamma$ positions are denoted.

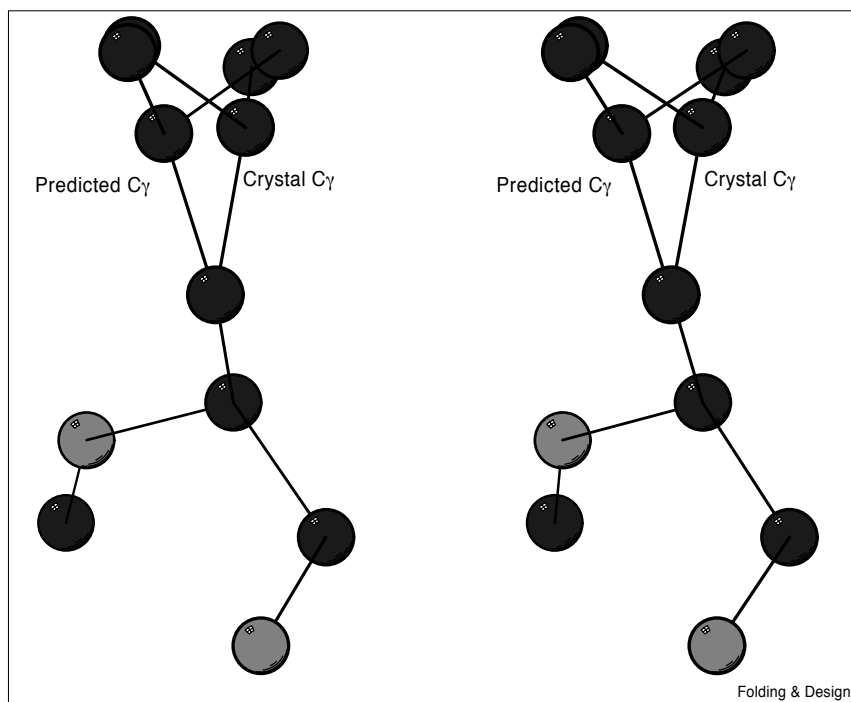
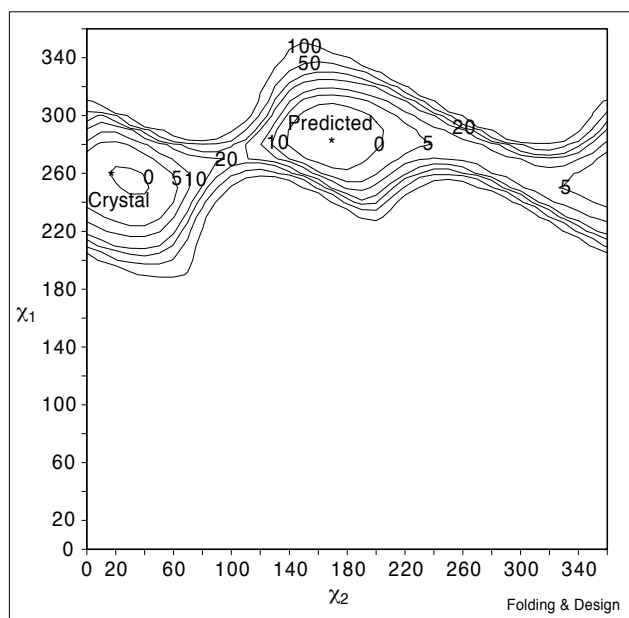


Figure 6



Contour map of energy (kcal/mol) as a function of χ_1 and χ_2 for Leu243 in thermolysin.

of $\chi_{2[\text{initial}]} = 120^\circ$ and $\chi_{2[\text{initial}]} = 300^\circ$; the small shifts result from minor asymmetries in the crystal coordinates. The squares and the triangles on the graph represent the change in χ_1 and the final value of χ_2 required to achieve D_{min} for a given $\chi_{2[\text{initial}]}$, respectively. Investigation using a finer grid of 1° increments similarly revealed that the minimum, D_{min} , is 0.11 \AA , occurring at $\chi_{2[\text{initial}]} = 40^\circ$ or 196° . The results indicate that the conformers around $\chi_1 = 40^\circ$ or 45° and $\chi_1 + 45^\circ$, 195° or 200° are the most nearly symmetric or pseudosymmetric pairs. These optimal χ_2 values are dependent on the bond angle $C\beta-C\gamma-C\delta_1$, and will, therefore, vary to some degree with this angle in other cases.

The dependence of leucine superposability on χ_2 is shown diagrammatically in Figure 8. From a superposable starting orientation (in this case $\chi_2 = 195^\circ$), shown as conformation 1 in Figure 8a, a rotation of χ_2 by 210° (or -150°) does not, by itself, result in close superposition of inverted δ -carbon positions. (Conformation 2 does not superpose closely on conformation 1.) A compensatory shift in χ_1 of -45° , however, results in near superposition. (Conformation 3 superposes closely on conformation 1.) In a similar manner, a starting position of $\chi_2 = 45^\circ$ (conformer 2) would require $\Delta\chi_1, \Delta\chi_2 = +45^\circ, +150^\circ$ for superposition. In contrast, from the non-superposable starting conformation of $\chi_2 = 300^\circ$ (or alternatively, $\chi_2 = 120^\circ$), shown as conformation 1 in Figure 8b, a rotation of χ_2 to invert the δ carbons, followed by any change in χ_1 , does not result in superposition.

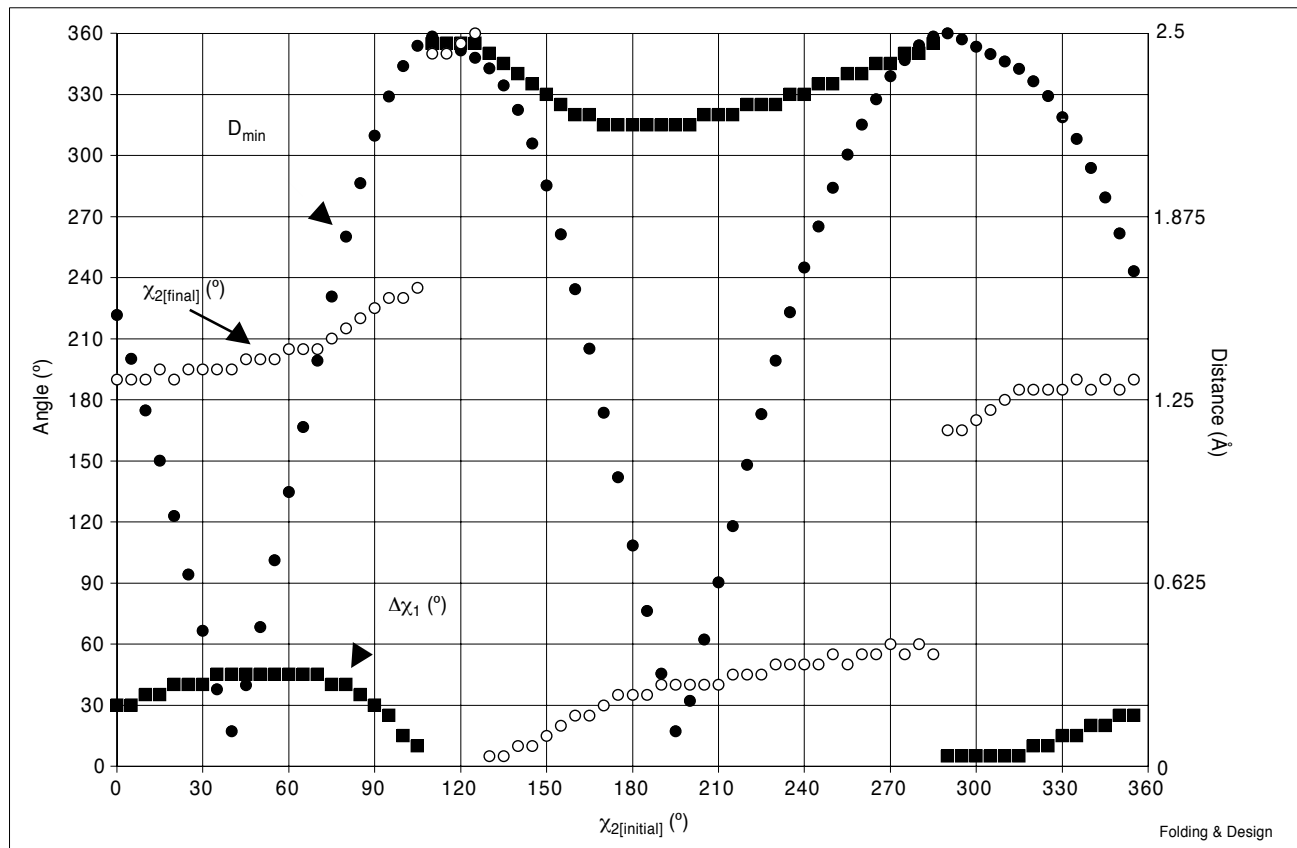
It is interesting that in Figure 7 most of the $\chi_{2[\text{initial}]}$ angles map to $\chi_{2[\text{final}]}$ values around either 40° or 195° (with accompanying shifts in χ_1). This leads to the paradoxical conclusion that, except for conformers with χ_2 values in the regions of 120° or 300° , which have no complementary conformer, the complementary conformer of any χ_2 value in leucine is in the region of $40-45^\circ$ or $195-200^\circ$. If $\chi_{2[\text{final}]}$ is substituted for $\chi_{2[\text{initial}]}$ recursively, the process always converges to the values of 40° and 195° , which map to one another.

In the present set of investigations, there were 16 incorrectly predicted leucine residues. For 13 of these leucines, the predicted conformation was the spatially pseudosymmetric conformer of the crystal orientation, whereas the crystal structure generally corresponded to the 2° or 3° minimum on the energy map (all four cases in thermolysin). In each case, the location of at least one conformer in the pair can be predicted from the location of the other (using Figure 7) on the basis of geometric considerations alone. In thermolysin, the mean energy difference between the primary minimum and the complementary conformers in incorrectly predicted pairs is $1.9 \pm 1.5 \text{ kcal/mol}$. Thus, the spatial pseudosymmetry of the leucine conformer pairs seems to give rise to a pseudo-degeneracy of angular states in terms of the potential energy. (It is not a degeneracy in the strict sense because the conformers are not chemically equivalent.) Furthermore, in the case of the superposable χ_2 positions in leucine ($\chi_2 = 40-45^\circ$ and $195-200^\circ$) there is an entropic contribution to the free energy, as a result of the presence of an additional local minimum of low energy. In the limiting case, this contribution is $-kT \ln(2/1) \sim -0.4 \text{ kcal/mol}$.

Particularly in cases of disagreement where the energies of the calculated and experimental conformations are very close, it is unclear whether one or the other represents the true crystal position, or whether both are in fact correct because the structure is conformationally heterogeneous at those sidechain positions. A plot of B factors, normalized by protein, versus χ_2 values for the crystal leucine positions revealed no clear trends.

Errors related to spatial pseudosymmetry, or atom position replacement, were also found in other residue types, namely isoleucine (two out of five errors in the core were spatially pseudosymmetric), histidine (three out of four), methionine (one out of one), and asparagine (one out of four). The symmetry errors involve the following changes in atomic positions: in isoleucine, one γ carbon replaces another through a χ_1 rotation of $\pm 120^\circ$; in histidine, the imidazole ring is inverted through a 180° rotation in χ_2 ; in methionine, compensatory changes in χ_1 and χ_2 result in nearly coincident sulfur atom positions and small shifts in carbon atom positions ($< 1 \text{ \AA}$); and in asparagine, a 180° rotation in χ_2 results in an inversion of the carbon and

Figure 7



Plot of best superposability of alternative conformers, D_{\min} , as a function of starting χ_2 angle ($\chi_{2[\text{initial}]}$) in leucine residues. The clear circles and the dark squares represent the final χ_2 angle and the χ_1 shift necessary to achieve optimal superposition for each starting χ_2 . See text for more details.

nitrogen atom positions of the carboxylamide group. Energetically, these errors fall into two groups: those in which the crystal conformer corresponds to a local minimum on the energy surface and in which the energy difference between the predicted and crystal conformer is small ($< \sim 4$ kcal/mol, four cases); and those in which the crystal orientation is not near a minimum and the energy differences are large (> 4 kcal/mol, three cases). It appears likely that at least some of the errors in the latter group involve errors in the crystal structure.

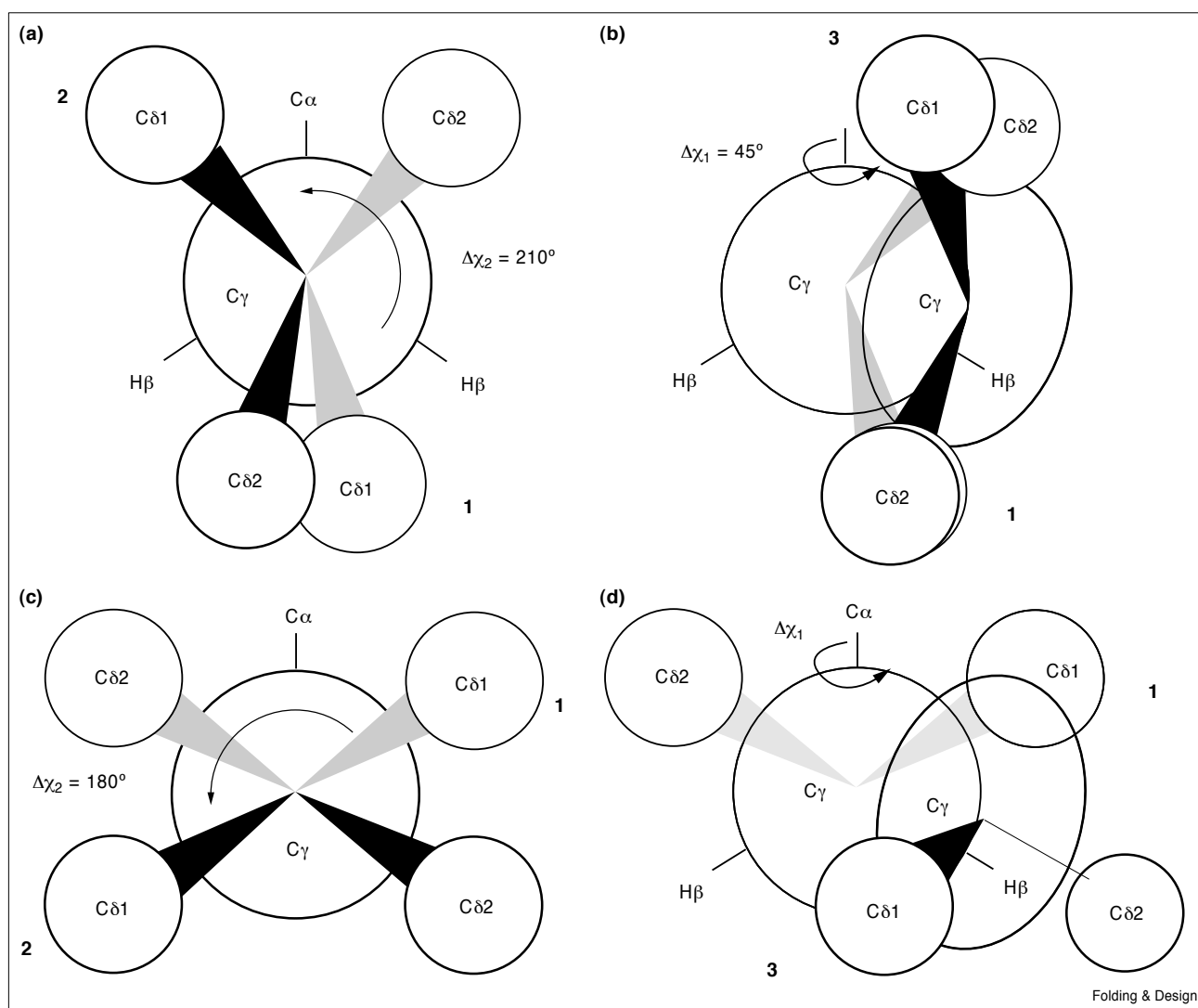
Notably, there were no instances of atom position replacement errors in core valine residues, despite the fact that a 180° rotation in χ_1 of valine with a slight shift in the $C\alpha$ and $C\beta$ atoms can result in near-superposition of inverted $C\gamma$ positions. The explanation for the absence of this error pattern in the predictions is suggested by the energy map for Val79, the only incorrectly predicted core valine (out of 76 total), which is shown in Figure 9. Note that the minima in the energy curve for valine occur around the expected (staggered) positions corresponding to $\chi_1 = 60^\circ$,

180° and 300° and that maxima in the curve occur around the (eclipsed) positions corresponding to $\chi_1 = 0^\circ$, 120° and 240° . Thus, a 180° χ_1 transition from any low-energy conformer in a valine residue results in a high-energy conformer. In contrast to the case with leucine, the spatially pseudosymmetric sidechain conformers in valine do not appear to be pseudodegenerate, at least when the backbone and $C\beta$ atoms are fixed. The crystal orientation in this particular valine is $\chi_1 = 0^\circ$, which is at an energy maximum, suggesting an error in the crystal coordinates. This energy map, in light of the inherent geometry of valine, suggests that the actual crystal conformer is near 180° (0° crystal position + 180° transition) and that the backbone has been shifted slightly.

Comparison of HIV-PR-A and HIV-PR-B

To investigate whether errors in the predictions may relate to sidechain variability in the proteins, the subunits A and B of the 1hiv structure were superposed and compared as described in the Materials and methods section. The backbones were found to vary little

Figure 8



Modified Newman diagrams showing the angular dependence of leucine pseudosymmetry. (a) and (b) depict the case of a superposable starting conformation. (a) Starting conformation 1 (lighter lines) has $\chi_2 = 195^\circ$. A χ_2 rotation of 210° (or -150°) results in conformation 2. (b) A shift in χ_1 of -45° results in conformation 3 (heavy lines), which nearly superposes onto conformation 1. In

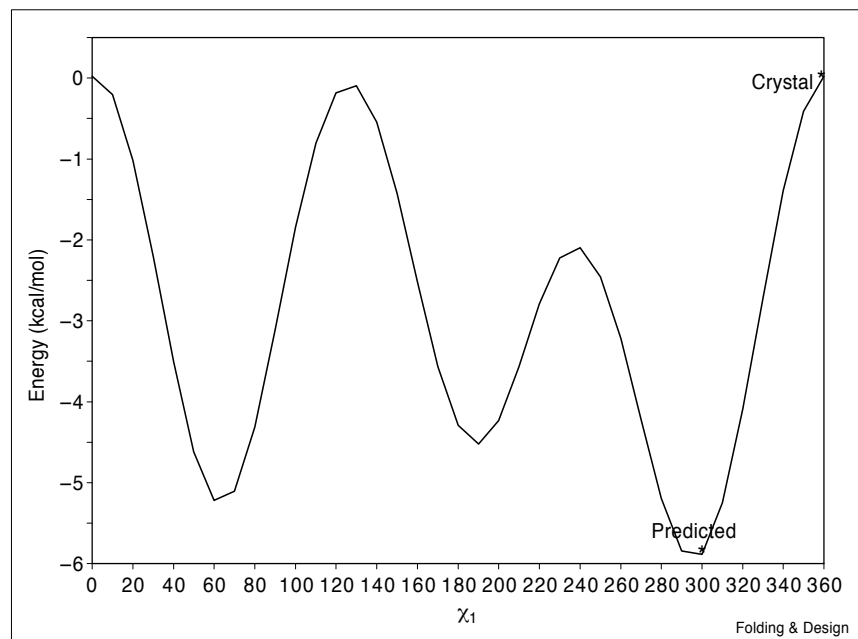
contrast, (c) and (d) depict the case of a non-superposable starting conformation. (c) The starting conformation has $\chi_2 = 300^\circ$ (conformation 1, lighter lines). A χ_2 rotation (arbitrarily chosen to be 180°) results in conformation 2 (heavy lines). (d) Changes in χ_1 do not lead to superposition. See text for more details.

compared with the sidechains. The rmsd values were as follows: all atoms 1.28 Å, core atoms 0.85 Å, backbone atoms 0.73 Å, core backbone atoms 0.48 Å, sidechain atoms (including C β atoms) 1.57 Å, and core sidechain atoms 1.04 Å. For two different crystal preparations of BPTI, Wlodawer *et al.* [51] found rmsd values of 0.40 Å for the mainchain atoms (not including the deviating carboxyl terminus) and 1.53 Å for sidechain atoms. Flores *et al.* [52] found rmsd values of 0.30–1.00 Å (mean = 0.40 Å) for C α coordinates in 16 pairs of identical proteins in different crystal preparations.

When the dihedral angles of the two subunits were compared, a total of 52 out of 154 or 33.8% varied by $\geq 40^\circ$ in χ_1 , χ_2 , or both. This is consistent with the results of previous authors [27,52]. In the core, 24 out of 86, or 27.9% varied. Thus, even in the core, where the backbone varied only slightly, there were significant deviations in the sidechains of the subunit structures. The structural predictions were much more accurate in the residues that did not vary from one subunit to the other. 91.2% (93/102) of the unvarying residues were correctly predicted, whereas only 61.5% (32/52) of the varying residues were correctly

Figure 9

Energy map for Val79 in thermolysin. Note the minima near 60°, 180° and 300° and that the crystal conformer is at a local maximum.



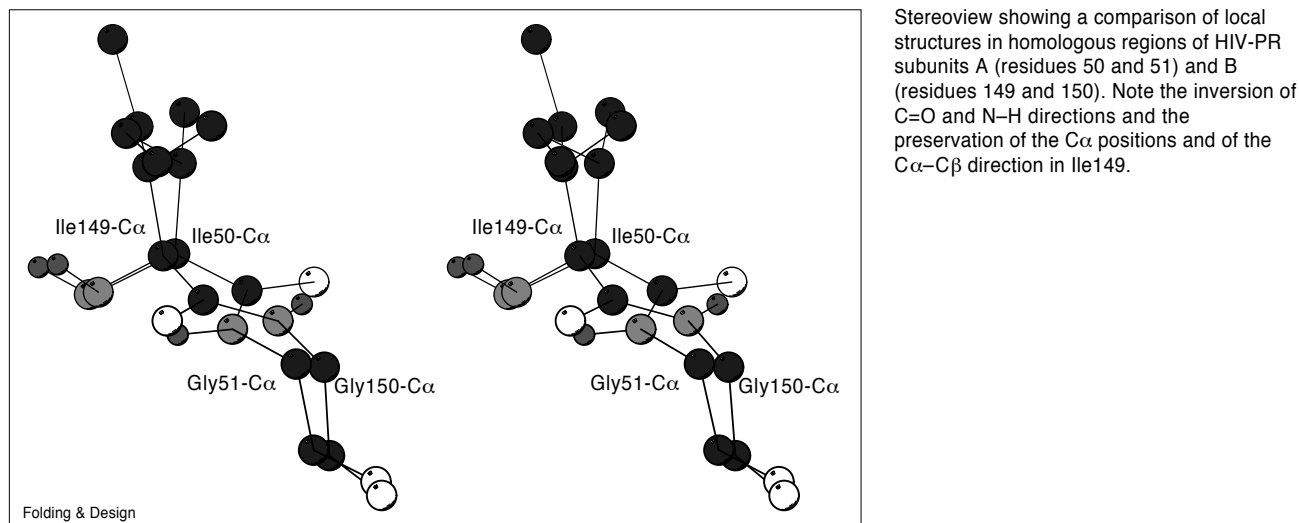
predicted. In the core, the accuracies were 96.7% (58/60) for the unvarying residues and 70.8% (17/24) for the varying residues. The differences are significant by χ^2 ($p < 0.001$) in the core and overall. Out of the total of 29 incorrectly predicted residues in the entire structure, nine occurred in the prediction of the orientation observed in the equivalent residue of the opposite subunit (four out of nine in the core). These results indicate that prediction errors in this study are often related to alternative conformations of the sidechains, and suggest that the predictability of the sidechains is inversely related to their mobility in solution. This is consistent with low-temperature Monte Carlo results showing an inverse relationship between sidechain entropy and predictability [7].

The patterns in the angle variations between homologous residues in the two subunits largely mirror the kinds of atom position replacement errors that were observed in the prediction studies. For example, in two leucine residues (89 and 38), the change between the A and B subunits involves spatially pseudosymmetric conformer pairs. There are pseudosymmetric changes in asparagine (180° rotation of χ_2), methionine (compensatory χ_1 and χ_2 shifts) and isoleucine (replacement of C γ 1 by C γ 2 or vice versa). In addition, there are other instances of pseudosymmetric differences that were not observed in the prediction studies, including a change in a glutamate residue that required rotations out to χ_3 . In one instance, around residues Gly51 and Ile50, which are situated at a turn, there is an inversion in the backbone structures (Figure 10). In subunit A, $\psi_{50} = 124.8^\circ$ and $\phi_{51} = 99.3^\circ$,

whereas in subunit B, $\psi_{50} = -25.1^\circ$ and $\phi_{51} = -91.1^\circ$, the consequence being rotation of the peptide bond plane by $\approx 180^\circ$ and preservation of the direction of the C α -C β bond. The calculated energy difference between the two backbone conformers is 6.1 kcal/mol, 1hiv B < 1hiv A (taking into account the local environments in each case). But because the overlap of inverted atomic positions is only partial, it is not clear that this conformational variation is an artifact of crystallographic fitting. Significantly, there is variation between the subunits in a core threonine residue (Thr80), even though there were no errors in the core threonine in the prediction studies. This finding suggests that the differences in sidechain orientations of the 1hiv A and the 1hiv B X-ray structures may result, in part, from subtle shifts in the backbone positions.

The correspondence of prediction error patterns to crystal structure variation raises the question of whether any of the prediction errors in the study are related directly to crystallographic disorder. There are several regions of disorder in the various crystal structures, mostly in the exposed regions, but alternative sidechain positions are specified in only two instances. Met52 and Glu7 each occupy two distinct sites, A and B, in the 5pti structure. The sites are non-overlapping and hence do not represent spatially pseudosymmetric disorder. The B sites, or conformers, were chosen arbitrarily for starting structures in the prediction studies. In both residues, the predictions were incorrect but corresponded to the A conformers. (Fortuitously, the χ_3 values are roughly equal for the A and B conformers in both residues.) The results suggest that it

Figure 10



may be possible to model this kind of discrete disorder, or conformational heterogeneity, in crystals by rigid rotation mapping. In other cases of disorder, although the position of alternative sites were not specified in the structures, their (χ_1 , χ_2) positions are predicted to be as follows: Ile7 and Ile25 in 1crn have positions (-70° , -60°) and (-60° , 170°), respectively; His119 in 3rn3 has position (180° , 90°); and Asp279 in 3app has position (-170° , -150°).

Discussion

At this time, the prediction of protein sidechain conformations given the correct backbone structure seems to have an upper limit in accuracy of the order of 70% overall (as measured by having both χ_1 and χ_2 correct), and 75–80% for the buried regions (Table 10). It is important to understand the origin of this limitation. In energy-based prediction methods, there are three broad categories of errors (Figure 11). The present study focuses on the accuracy of the potential function, in this case the CHARMM energy function. The results show that for the minimal problem of predicting individual sidechain orientations by rigid-rotation mapping of χ_1 and χ_2 , given the rest of the protein crystal structure, the accuracy of the potential function yields considerably better predictions than those obtained in full prediction studies. The fraction of residues correctly predicted averaged 77.4% (884 out of 1142) overall and 89.5% (514 out of 574) for the core in the vacuum calculations, and 86.5% overall and 94.9% in the core in the presence of crystal waters. These results, taken together with the detailed analysis of the incorrectly predicted residues, strongly suggest that the potential energy function *per se* is sufficiently accurate for sidechain prediction, and that prediction errors originate largely from other sources. In this study, the errors are caused predominantly by the lack of crystal water molecules in the vacuum calculations; this

affects both the exposed and interior regions of the proteins. The findings indicate that inclusion of water molecules (or their effects) in energy-based prediction significantly improves results in all regions. This is consistent with studies showing improved prediction accuracy with the use of implicit solvation models ([53]; R.J.P., T.L. and M.K., unpublished observations). Calculations in vacuum that do not account for solvent effects therefore appear to have inherent limitations for predicting crystal sidechain positions, even for sidechains in the core. Because the observed orientations of exposed residues can depend upon particular interactions with specific water molecules and with specific protein crystal contacts [34], the comparison with crystal coordinates may be less relevant for the structure in solution. It appears likely that there is a limit to the predictability of crystal structures in the absence of information about the particular crystal environment, although even in vacuum there is some preselection of conformers in the exposed regions of the protein.

The current results demonstrate that the absence of solvent in the calculations introduces an error in the hydrogen-bonding patterns of charged and polar residues: nearest-neighbor intramolecular donor-acceptor distances in the calculations are often shorter than those in the crystal structure. The elimination of electrostatic terms from the potential energy function, as in some crystal refinement methods and many sidechain prediction methods [4,6,15,20,26,27,29–31], removes this source of error and hence may not result in as large a reduction in accuracy as might be expected. Nonetheless, the current results indicate that accurate energy-based predictions require the inclusion of both electrostatic forces and solvation effects.

Table 10

Accuracy of various sidechain prediction methods.

Reference	All residues				Core residues*				
	χ_1	χ_2	χ_{1+2}^\dagger	Rmsd (Å)	χ_1	χ_2	χ_{1+2}^\dagger	Rmsd (Å)	
Lee and Subbiah (1991)	[4]	57–89	50–69		1.97	81–82	74–81		1.25
Desmet <i>et al.</i> (1992)	[12]			72 [‡]				93 [‡]	
Holm and Sander (1992)	[6]	72 [§]			1.8	81 [§]			1.4
Wendoloski and Salemme (1992)[23]		53–70 [#]		39–59 [#]	2.04 [#]				
Dunbrack and Karplus (1993)	[24]	78	74	69		88	83		
Eisenmenger <i>et al.</i> (1993)	[25]	74		54	1.7	87		74	1.1
Tanimura <i>et al.</i> (1994) [¶]	[69]			68–70	1.82			77–85	1.24
Koehl and Delarue (1994)	[29]	72	75	62	1.89	82	79	72	1.38
Hwang and Liao (1995)	[11]	82	72	68	1.68				1.27
Keller <i>et al.</i> (1995)	[15]	73	66	60					
Vasquez (1995)	[31]	80		68	1.55	88		79	0.99
Shenkin <i>et al.</i> (1996)	[7]	74		63					
Bower <i>et al.</i> (1997) [¶]	[27]	78		66	1.97				

*All authors define the core of a protein as that set of residues in which each residue falls below a given cut-off value for surface accessibility, which is usually expressed as a percentage of maximum surface accessibility for each amino acid type. The cut-off values vary from author to author, however: Lee and Subbiah, not available; Desmet *et al.*, 10%; Holm and Sander, 20%; Dunbrack and Karplus, 10%; Eisenmenger *et al.*, 25%; Tanimura *et al.*, 30%; Koehl and Delarue,

30%; Hwang and Liao, 20%; and Vasquez, 40%. [†] χ_{1+2} refers to both χ angles being correct. [‡]Data for single insulin dimer only. [§]30° scoring criterion. For structures determined at 2.5 Å or better. [#]Given C α coordinates only. [¶]Results only for 'independent model', which had best results for core. Rmsd values are averages of three sets of calculations for this model. [¶]For set of highest resolution structures (2.0 Å or better).

The effect of crystal waters on the predictability of buried residues arises in part from the fact that a polar residue can be mostly buried, and hence be categorized as a core residue [47], but still have its polar or charged sidechain functional group exposed to solvent (see Figure 12). Even residues that were 100% buried in some cases required the presence of solvent for accurate prediction, however. The observation that buried protein functional groups are not precluded from solvent effects has been made by previous investigators. Lazaridis *et al.* [54] have demonstrated that carbonyl groups on protein residues with a solvent accessibility of zero by a Lee and Richard's surface calculation interact significantly with solvent water atoms. The interposition of neighboring chemical groups between a particular protein atom and a solvent atom (or a contact atom on

an adjacent crystal molecule) does not preclude significant electrostatic interactions.

In a small number of cases, there was evidence for an error in the backbone structure. Because the backbone is taken as given in the present study and in many sidechain prediction methods, calculation of the correct energy surface (and, therefore, correct prediction) is predicated upon the backbone coordinates being correct. From both statistical and energetic analyses [24,46,55] and from sidechain prediction studies using non-self backbones [6,15,27–29,33], it is clear that local shifts in backbone structure can significantly alter sidechain orientation. The conformational variability of 25–30% for sidechains, or more, found in different crystal preparations of the same protein [27,52] and

Figure 11

General breakdown of possible sources of error in energy-based structure prediction.

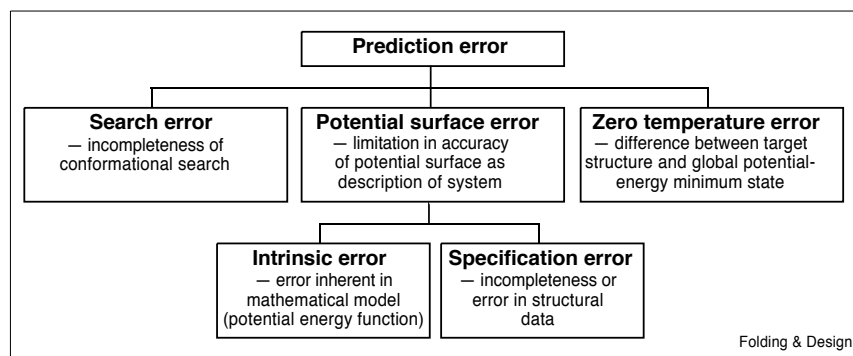
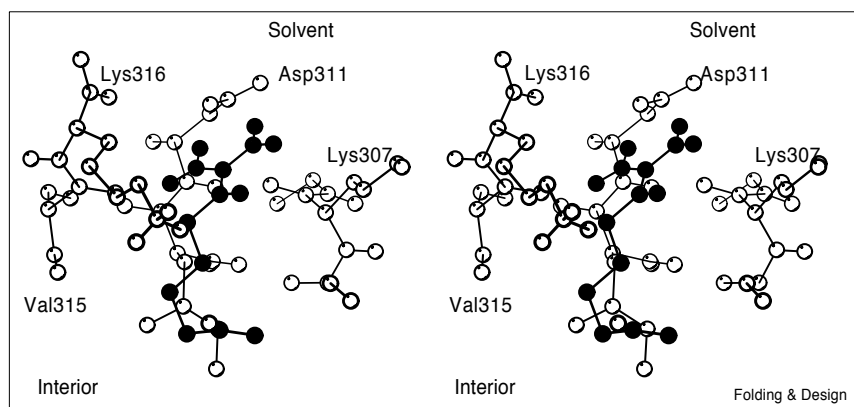


Figure 12

Stereoview showing Arg285 in thermolysin. Although the residue is mostly buried (< 25% solvent accessibility), the polar end groups are largely exposed to solvent.

in the comparison of the two nearly identical subunits of the HIV-1 protease in this study, compared to the conformational heterogeneity of only 6–13% of sidechains within single crystal structures [38], indicates that local backbone shifts of $\leq 0.5 \text{ \AA}$ may be sufficient to alter a significant fraction of sidechain conformations. Hence, the use of any backbone geometry other than that of the actual target crystal structure is likely to limit sidechain prediction (although this was not a problem in the current study because the exact mainchain structure was used).

For a few residues (perhaps 2–3% of the total number) with incorrectly predicted conformers and native conformers having a similar energy, it is difficult to rule out any of several possible sources of prediction error, including intrinsic CHARMM error, crystal disorder (with pseudodegeneracy of alternative conformers), or entropic contributions. In at least two cases, the error related to the prediction of the alternative site in sidechains that occupied two separately defined sites in the crystal. In other cases, it involved the prediction of conformers having both similar energies and also a high degree of spatial overlap with the crystal conformer (spatial pseudosymmetry). In these instances, it is difficult to distinguish between the above-mentioned sources of error in the predictions and possible errors in the crystallographic model.

The results also indicate that, in the presence of the full protein, rigid-rotation energy mapping for an individual sidechain generally results in a calculated absolute energy minimum that corresponds to the native crystal conformer, particularly in the core. Hence, the inclusion of local entropic effects (i.e. thermal averaging over slightly differing conformations about each dihedral angle position) is not required for correct prediction in the current test. Even when the prediction is incorrect, there is a high correlation between the crystal positions and relative, or local, minima on the calculated energy maps. Usually, prediction

errors are not a result of a disagreement between crystal positions and local minima, but rather a result of the existence of one or more calculated minima having lower energies than that of the native conformer. These results are consistent with previous studies performed on more limited regions of conformational space [34]. Furthermore, because the structures in the study were refined almost entirely with stereochemically restrained refinement techniques, not energy-based ones, the findings are unlikely to represent an artifact of the refinement process. The correspondence of the crystal orientation with one of the minima on the energy map is higher in the core than in the exposed regions, particularly in the vacuum calculations (99.3% versus 86.8% for thermolysin in vacuum), because the energy surfaces of the exposed sidechains are flatter than those in the core and the positions of local minima are more easily shifted by external forces (such as those of solvent). Still, even in the exposed regions, the protein environment generally preselects a set of possible sidechain conformers.

Most residues have alternative conformers (of higher energy) even in the fixed protein environment. This is true for buried residues as well as those that are exposed and suggests that, in most cases, multiple conformers exist for sidechains in solution and the process of crystallization selects one out of the several local minima. Orientational differences in corresponding sidechains of the two subunits of the HIV-PR crystal structure (1hiv), in the context of an almost identical backbone structure, also suggest conformational heterogeneity of the sidechains in solution. This is supported by similar findings in other studies [27,38,49–52,56,57]. The current results indicate that the conformational variation commonly involves replacement or exchange of atomic positions. Sidechain predictability was found to be directly related to the ruggedness of the energy maps (and inversely related to sidechain variability). The size of the calculated energy gap between alternative

potential energy minima correlates well with predictability in both buried and exposed regions (i.e. for gaps > 6.0 kcal/mol, the prediction accuracy is 93.2% overall). This may serve as a gauge of reliability for the energy-based prediction of sidechain structure.

In summary, the present results suggest that the potential energy function is sufficiently accurate for sidechain predictions and that there are other factors limiting the absolute accuracy of current energy-based sidechain prediction methods. These factors include inadequate representation of solvent and crystal contact effects, sidechain orientational heterogeneity (accompanied by pseudodegeneracy of conformers), inadequate conformational searches and rotamer approximations, and a small number of inaccuracies in crystal structures. The necessary condition on the energy function explored in this study could be applied usefully to many sidechain prediction methods to establish upper limits of accuracy.

Materials and methods

Selection and preparation of protein structures

Ten protein crystal structures (Table 11) were chosen from the Brookhaven Protein Data Bank [58]. They were selected because they have all been used in sidechain prediction studies, with the exception of the HIV-1 protease. All are at a resolution of 2 Å or better with the exception of the Cro repressor structure, and all were refined with stereochemically restrained refinement techniques such as the least squares refinement methods of Konnert [59], Morffew and Moss [60], and Sussman *et al.* [61]. In only three cases did some part of the refinement involve energy-based procedures: X-PLOR [62] in 1hiv; EREF [63] in 5pti; and an early empirical function of Levitt [64] in 4fxn.

To reflect the design of most previous energy-based sidechain prediction studies, the proteins were treated in vacuum (but see below), and solvent molecules, metal ions, and ligands were removed from the coordinate files of the PDB structures. The ligands removed were PO_4^{3-} from 5pti, SO_4^{2-} from 1ctf and 3rn3, the U75875 ligand from 1hiv, and four Ca^{2+} and one Zn^{2+} from 3tln. Disulfide bonds were broken, and cysteine sidechain residues were allowed to rotate freely when they were being tested. Polar hydrogens were added to the structures using the empirical energy placement protocol H-BUILD [65] in the CHARMM program. For consistency, in the case of the neutron structure of bovine pancreatic trypsin inhibitor (5pti), which defines hydrogen positions in the crystal, all hydrogen coordinates were first omitted and then rebuilt in the same manner as for the other nine proteins. In two proteins, residues were renumbered: $N_{\text{study}} = N_{\text{PDB}} - 52$ for 1ctf, and $N_{\text{study}} = N_{\text{PDB}} + 2$ for 2cro.

All the protein structures were subjected to a very limited energy minimization comprising 25 steps Steepest Descent to eliminate bad contacts and justify bond angles and lengths in accord with the CHARMM potential energy function while preserving the structure as much as possible. This is important not for the predictions but for meaningful energy calculations. Bad contacts that are present in many crystal structures, even at high resolution, can introduce artifacts into the energy values. Also, because the rotations are rigid, justifying the bond lengths and angles in the native structure to be consistent with the potential energy function is necessary to obtain more meaningful torsional-energy contours. As shown in Table 12, the 25-step minimization changed the structures only slightly, with rmsd values of 0.082 Å for all residues and 0.079 Å for buried residues. For comparison, the limiting precision for refined atomic positions of protein X-ray structures at 2.0 Å has been estimated to be 0.15–0.25 Å [66]. The mean χ angle

shifts are 4.2° and 3.6°, respectively, for all residues, with a maximum (in the entire set of proteins) of 34.2°. To verify that the minimizations did not bias the predictions, a portion of the calculations were repeated using the original crystal coordinates. The difference was $\leq 1.7\%$ or less, as described in the Results section. The terms 'X-ray structures' or 'crystal structures' refer to the slightly minimized coordinates unless otherwise indicated.

Rigid rotation and torsional potentials

The sampling of conformational space was carried out by rigid-geometry mapping for individual sidechains, as described by Gelin and Karplus [34]. Sidechains for all protein residues, with the exceptions of proline, alanine and glycine, were rotated rigidly about their heavy-atom dihedral angles. Bond lengths and bond angles were not varied.

In the first set of calculations, each sidechain was rotated around χ_1 through 360° at 5° intervals; all other χ angles of the sidechain were fixed at the X-ray values. In the second set of calculations, each sidechain was rotated through its entire $\chi_1 \times \chi_2$ angular space at 10° intervals (1296 conformers per sidechain for those having both χ_1 and χ_2 angles); all other χ angles of the sidechain were fixed at the X-ray values. A 10° angle was chosen to save time, because comparison of results from computations using 5° and 10° increments in thermolysin showed no significant change in prediction accuracy. Furthermore, an analysis of the energy minima for all proteins confirmed that the 10° grid was fine enough to locate the absolute minimum in the calculated energy maps in almost every case. In only 1 out of 258 incorrectly predicted residues (0.4%) was the energy of the predicted orientation higher than that of the crystal orientation, implying that the crystal position was 'missed' in the grid search. Out of 308 cases in which $E_{\text{cryst}} < E_{\text{pred}}$, 307 were correct, indicating that if the crystal orientation corresponded to the absolute minimum of the complete (i.e. continuous) energy map for a given residue, a 10° grid was successful in locating it 99.7% of the time.

Calculations of the torsional energies were carried out using a standard CHARMM empirical potential energy function [40]. Version 19 of parameter and topology files was used [42], which defines a polar-hydrogen protein model, with non-polar hydrogens represented by extended atoms. Because all except the chosen dihedral angles of one sidechain remain unchanged during the rigid-geometry mapping procedure, the energy computations are rapid. For an individual dihedral angle rotation, they involve only the dihedral angle term of the rotated angle and the non-bonded interaction energy between the rotated portion of the sidechain and the union of the unrotated portion of the sidechain and the rest of the protein. For simultaneous changes in two dihedral angles, changes in the internal energy of the rotated portion of the sidechain must also be taken into account.

Non-bond terms were truncated at 9 Å, with a shifted smoothing function and a distance-dependent dielectric function for electrostatics and a switched smoothing function for van der Waals terms [40]. The distance-dependent dielectric (which varies with $1/r$) was chosen for these *in vacuo* calculations to approximate the shielding effects of solvent water; version 19 is designed to be used with such a dielectric function. The non-bonded list, which defines the groups of atoms included in the calculation of non-bond energies (van der Waals and electrostatics) was not updated throughout the calculations for a given sidechain to save computer time. For the first 150 residues of thermolysin, using either no non-bond update, an update at every 25 steps, or an update at every step, the accuracies of prediction were 91.5%, 90.6% and 90.6%, respectively, for χ_1 and 80.5%, 80.5% and 80.5%, respectively, for χ_2 . Average angular deviations of predicted positions from crystal positions and the results for the core were also unchanged. The computer time, however, increased from 1.04 s/orientation with no update to 5.11 s/orientation for an update at every step.

Calculations were performed on a Hewlett Packard-9000/735. CPU time across all calculations averaged ~ 1.0 s/orientation. In the first set

Table 11

Characteristics of the proteins used in this study.

Protein PDB code	Name	Resolution* (Å)	R value	Number of residues in protein†	Number of core residues‡	Number of χ_1 §	Number of χ_2 §
5pti	BPTI	1.80	0.200	58	17	42	31
1crn	Crambin	1.50	0.114	46	13	32	16
2cro	434cro	2.35	0.195	65	26	54	42
1ctf	L7/L12	1.70	0.174	68	23	46	35
4fxn	Flavodoxin	1.80	0.200	138	70	115	89
1hiv	HIV-1 protease	2.00	0.169	198	99	154	118
1lz1	Lysozyme	1.50	0.187	130	61	103	75
3app	Penicillopepsin	1.80	0.136	323	173	247	146
3rn3	Ribonuclease A	1.45	0.223	124	50	105	63
3tln	Thermolysin	1.60	0.213	316	179	244	171
All				1466	711	1142	786

*Crystallographic resolution (X-ray or neutron scatter). †Total number of residues in protein. ‡Total number of core residues, as defined in the text. §Total number of sidechain heavy-atom torsional angles (not including prolines). 5pti, bovine pancreatic trypsin inhibitor, Wlodawer *et al.* [73]; 1crn, Hendrickson and Teeter [74]; 2cro, Mondragon *et al.*

[75]; 1ctf, C-terminal domain of the ribosomal protein L7/L12, Leijonmarck and Liljas [76]; 4fxn, Smith *et al.* [77]; 1hiv, Thanki *et al.* [78]; 1lz1, Artymiuk and Blake [79]; 3app, James and Sielecki [80]; 3rn3, Borkakoti *et al.* [81]; and 3tln, Holmes and Matthews [82].

of calculations, this yielded total run times ranging from 20 min for crambin (32 χ_1 angles \times 36 = 1152 orientations) to 2.5 h for penicillopepsin (247 χ_1 angles \times 36 = 8892 orientations). In the second set of calculations, which covered both χ_1 and χ_2 , run times were much longer, ranging from 6 h for crambin (32 χ_1 , 16 χ_2 , 21,312 orientations) to 2.6 days for thermolysin (244 χ_1 , 171 χ_2 , 2.2×10^5 orientations).

Definition of the core

The structural core for each protein was defined as the set of residues (sidechain and mainchain atoms included) having <25% solvent accessibility as measured by the solvent-accessible surface area [67] compared to extended peptides [68]; a solvent probe size of 1.4 Å was used. All residues not in the core are defined as exposed residues. Previous works have set this solvent accessibility cut-off for the core in the range 10% [12] to 40% [31].

Analysis of calculations

Predicted dihedral angles were defined as 'correct' if they were within $\pm 40^\circ$ of the known value in the starting structure. This choice means

that the angle is within the same local minimum and in many cases would minimize to a closer value [21]. Many workers (but not all) have used a corresponding criterion [4,11,15,21,24,25,29,69]. Because chemical equivalence at the δ position creates a twofold symmetry in torsional states about χ_2 in phenylalanine and aspartate residues, values equal to $180 \pm 40^\circ$ were allowed in these cases. In addition, actual angle deviations and root mean square atomic positional deviations (rmsds; heavy atoms only) were calculated and compared. The twofold rotational symmetry axes in χ_2 of phenylalanine and aspartate were taken into account in the rmsd calculations as well. Because C β atom positions are not altered in χ angle torsions, rmsds were calculated for sidechain atoms including only the γ positions and beyond; in some other sidechain prediction studies, C β atoms were included in sidechain rmsd calculations, although they are essentially fixed by the mainchain.

The energy as a function of χ_1 or ($\chi_1 \times \chi_2$) was plotted for all the residues of thermolysin and for selected residues in other proteins. Conformers corresponding to relative energy minima were then compared with the X-ray positions. This was of particular interest when the X-ray result deviated from the lowest calculated energy minimum.

Table 12

Deviations of 25-step relaxed structures from original PDB coordinates.

Protein PDB code	Number of χ^*	Mean $\Delta\chi^\dagger$ (°)	s.d. $\Delta\chi^\dagger$ (°)	$\Delta\chi \text{ max}^\ddagger$ (°)	Rmsd § (Å)	Rmsd core § (Å)	Initial E $^\#$ (kcal/mol)	Final E $^\#$ (kcal/mol)
5pti	103	5.3	5.3	32.2	0.11	0.11	245.3	-502.0
1crn	55	2.8	2.9	13.7	0.09	0.11	8.3	-405.5
2cro	136	3.9	3.8	19.0	0.08	0.08	-305.6	-717.7
1ctf	113	5.1	5.2	30.5	0.08	0.07	-80.3	-595.2
4fxn	256	4.3	4.0	26.4	0.09	0.08	-188.2	-1068.6
1hiv	344	4.1	3.2	18.0	0.09	0.08	-463.8	-1876.5
1lz1	241	4.9	5.2	34.2	0.09	0.09	-328.2	-1703.9
3app	432	3.9	3.5	31.5	0.08	0.07	-1451.5	-3171.4
3rn3	216	5.3	4.5	30.3	0.10	0.10	-60.8	-1293.5
3tln	490	3.2	2.7	18.8	0.07	0.07	-2002.5	-3658.1
All		4.2	3.6	34.2	0.08	0.08		

*Number of heavy atom sidechain dihedral angles in entire protein. †Mean and standard deviation of the change in dihedral angles from original X-ray values. ‡Maximum deviation for any dihedral angle in the protein. §For all heavy (non-hydrogen) atoms in the protein. #Calculated energies of the protein before (initial) and after (final) 25-step relaxation.

Analyses for statistical significance were performed by student's t-test and chi-square (χ^2).

Studies of thermolysin

Thermolysin was selected for more detailed investigation. Its size (316 residues, 244 χ_1 and 171 χ_2 dihedral angles) means that it has a significant number of core residues and, more generally, provides a large enough number of residues to yield meaningful results.

Investigation of solvent effects. With the TIP3 water model [70], 173 water molecules were built from the water oxygen coordinates in the 3tn crystal structure and added to the protein. Polar hydrogens were built onto both crystal solvent and protein molecules as described above. Limited energy minimization (25-step steepest descent) was performed as above for all protein atoms and water hydrogen atoms, with water oxygen atoms constrained to their initial positions. Rigid geometry energy mapping for the protein was then carried out as described earlier.

Use of quenched structure. To assess the effects of relaxing the starting structure, the same potential energy mapping was repeated for the thermolysin structure using minimized coordinates, in the absence of crystal solvent. The original PDB coordinates (mainchain and sidechains) were subjected to 300 steps Steepest Descent and 600 steps Newton–Raphson minimization. The final rmsd values were 1.08 Å overall, 0.83 Å for the backbone atoms and 1.29 Å for the sidechain atoms (including C β). 56 out of 244 (22.9%) sidechains shifted by $\pm 40^\circ$ or more in χ_1 , χ_2 or both, including 25 out of 138 (18.1%) sidechains in the core. It was expected that predictions would be more accurate in the quenched structure than in the X-ray structure, because the quenching process should stabilize the native positions but not necessarily the alternative conformers. Of interest, however, are the magnitudes and mechanisms of this effect.

Relative minima. An automated procedure was developed that locates all relative minima on the discrete two-dimensional χ_1 , χ_2 surface. A relative minimum was defined as a point on the energy surface from which proceeding in any direction results in a higher energy.

Multiple directions were explored about each point. This was necessary because, unlike the case in continuous space, in discrete space the attainment of a minimum with respect to both orthogonal axes is not always sufficient to demonstrate a local minimum at a given point on the surface. It is only sufficient if the energy contour about the minimum is parallel to one of the axes. In other cases, where the contour around a local minimum runs obliquely to the axes, exploration of only the $-\chi_1$, $+\chi_1$, $-\chi_2$, and $+\chi_2$ directions about points in the region will often result in the generation of a set of false 'minima' situated along the contour rather than the true minimum. This discretization effect was observed for at least one relative minimum in the majority of energy maps in this study (10° grid).

Hence, if two minima located in this procedure were within $\pm 30^\circ$ for both χ_1 and χ_2 (i.e. if the difference in the angles was $< 30^\circ$) they were considered as part of the same minimum, and the higher-energy member of the pair was discarded. All relative minima were located in the energy maps of every (rotatable) sidechain of the crystal structure of thermolysin. In addition, relative minima were located for selected sidechains in the quenched structure of thermolysin and the crystal structures of the other proteins.

Superposability of leucine conformers. It has been noted previously that for leucine residues, very different sets of χ angles can result in nearly coincident atom positions [4,24]. Specifically, there are conformer pairs that appear to be related by an inversion of C δ_1 and C δ_2 , and a small shift in C γ [4]. To investigate this spatial pseudosymmetry of leucine conformer pairs, a model system was designed to search for the χ_1 , χ_2 combination that most nearly superimposes reversed C δ

positions. A single leucine residue was rotated about its entire $\chi_1 \times \chi_2$ angular space at 5° intervals, and the following quantity was evaluated:

$$D = |\mathbf{C}\delta_{1[\text{final}]} - \mathbf{C}\delta_{2[\text{initial}]}| + |\mathbf{C}\delta_{2[\text{final}]} - \mathbf{C}\delta_{1[\text{initial}]}| \quad (1)$$

where the **Cs** are the coordinate vectors for the respective atoms. $\mathbf{C}\delta_{1[\text{final}]}$ and $\mathbf{C}\delta_{2[\text{final}]}$ are the coordinates resulting from a particular ($\Delta\chi_1$, $\Delta\chi_2$) transformation of the coordinates in the initial conformation, $\mathbf{C}\delta_{2[\text{initial}]}$ and $\mathbf{C}\delta_{1[\text{initial}]}$. D is thus a measure of the deviation from exact superposition of reversed C δ positions, where $D = 0$ would imply perfect superposition. On the 5° grid, D was evaluated for all possible ($\Delta\chi_1$, $\Delta\chi_2$) transformations on all possible $\chi_{2[\text{initial}]}$ positions, and the values of $\Delta\chi_1$, $\chi_{2[\text{initial}]}$, and $\chi_{2[\text{final}]}$ for which D was at a minimum were recorded. Leu14 from thermolysin (3tn) was chosen arbitrarily for this study, but the results are expected to hold for all leucine residues, with minor adjustments as a result of small differences in bond lengths and bond angles.

Effect of dielectric factor. To test the effect of the dielectric factor on predictions in vacuum, the vacuum calculations were repeated for thermolysin varying the dielectric function in both magnitude and form (constant versus distance-dependent). This is of interest because many structure prediction methods have employed only van der Waals non-bonded energy terms (i.e. no electrostatics or hydrogen-bonding terms; [4,6,7,15,20,26,27,29–31]). Also, most X-ray refinements at higher resolution do not employ restraint terms based on hydrogen bonding [49,71,72]. In X-PLOR positional refinements, charges on the longer sidechains (glutamate, lysine and arginine) are typically removed. Physically, improved prediction would be expected with the inclusion of electrostatics, although in the absence of solvent this is not necessarily true.

Possible biasing effect of initial minimization. To test for a possible biasing effect on the predictions of the initial 25-step minimization, the $\chi_1 \times \chi_2$ vacuum predictions on thermolysin were repeated using the original (unaltered) crystal coordinates both for the calculations and also as the basis for comparison. All other parameters and conditions were left unchanged.

Measurement of sidechain conformer variability by use of the HIV-1 protease dimer (HIV-PR)

To investigate a possible relationship between prediction errors in this study and alternative sidechain positions occurring in the protein structures themselves, the HIV-1 protease dimer was used as a measure of sidechain orientational variability. This is similar in spirit to the early work of Tulinsky *et al.* [56] that compared the subunits of dimeric α -chymotrypsin. More recent studies [27,51,52,57] have compared the sidechain structure of different crystal refinements of the same protein. HIV-PR is composed of two subunits, A and B, which are refined separately in the X-ray structure 1hiv (they are both included in the asymmetric unit of refinement). But because the subunits are identical in sequence and because they are arranged in a twofold symmetric fashion about a rotational axis that intersects the geometric center of the bimolecular complex, their backbones are almost exactly superposable. The total rmsd for the backbone is 0.73 Å, and for the core backbone it is 0.48 Å. In the context of this near-identity of backbone structures, differences in sidechain positioning give an indication of sidechain conformational variability in proteins. To measure the difference between the subunits, their structures were superposed by least-squares fit. Because the backbone superposability is not exact, comparison of atomic positions in individual residues required an additional least squares fit of the local backbone and C β atoms at each amino acid position. The rmsds and χ angle deviations were determined as described above.

Calculation of hydrogen bonds

Intramolecular hydrogen bonds were calculated using the Hydrogen Bonding facility in the molecular graphics program QUANTA (© Molecular Simulations, Inc., Version 4.1.1). Only hydrogen bonds and 'near'

hydrogen bonds with heavy-atom distances of 4.0 Å or less and donor–hydrogen–acceptor bond angles of 90° or more were included. The number of intramolecular hydrogen bonds for each sidechain was then tabulated, including only the bonds involving any sidechain atom on the residue being examined (not backbone N or O) and any atom of the surrounding residues.

Acknowledgements

We thank Bruce Gelin, Michael Schaefer, John-Marc Chandonia, Sung-Sau So, Matthias Buck, Herman van Vlijmen, Diane Joseph-McCarthy, Carla Mattos, and Adrian Mulholland for contributing with helpful discussions and technical expertise. We thank Roland Dunbrack for his careful reading of the paper and for his many useful suggestions, in particular that of using Newman diagrams to help depict the leucine superpositions. This work was supported in part by a grant from the National Science Foundation.

References

- Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.
- Moult, J. & James, M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* **1**, 146-163.
- Bruccoleri, R.E. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26**, 137-168.
- Lee, C. & Subbiah, S. (1991). Prediction of protein sidechain conformation by packing optimization. *J. Mol. Biol.* **217**, 373-388.
- Rotiberg, A. & Elber, R. (1991). Modeling sidechains in peptides and proteins: application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **95**, 9277-9287.
- Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for sidechain optimization in proteins: application to model-building by homology. *Proteins* **14**, 213-223.
- Shenkin, P.S., Farid, H. & Fetrow, J.S. (1996). Prediction and evaluation of sidechain conformations for protein backbone structures. *Proteins* **26**, 323-352.
- Correa, P.E. (1990). The building of protein structures from α -carbon coordinates. *Proteins* **7**, 366-377.
- Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein sidechain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267-1289.
- Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1993). A critical comparison of search algorithms applied to the optimization of protein sidechain conformations. *J. Comput. Chem.* **14**, 790-798.
- Hwang, J. & Liao, W. (1995). Sidechain prediction by neural networks and simulated annealing optimization. *Protein Eng.* **8**, 363-370.
- Desmet, J., DeMaeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein sidechain positioning. *Nature* **356**, 539-542.
- Lasters, I. & Desmet, J. (1993). The fuzzy-end elimination theorem: correctly implementing the sidechain placement algorithm based on the dead-end elimination theorem. *Protein Eng.* **6**, 717-722.
- Goldstein, R.F. (1994). Efficient rotamer elimination applied to protein sidechains and related spin glasses. *Biophys. J.* **66**, 1335-1340.
- Keller, K.A., Shibata, M., Marcus, E., Ornstein, R.L. & Rein, R. (1995). Finding the global minimum: a fuzzy end elimination implementation. *Protein Eng.* **8**, 893-904.
- Lasters, I., DeMaeyer, M. & Desmet, J. (1995). Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein sidechains. *Protein Eng.* **8**, 815-822.
- DeMaeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improved both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* **2**, 53-66.
- Sutcliffe, M.J., Hayes, F.R.F. & Blundell, T.L. (1987). Knowledge-based modeling of homologous proteins, part II: rules for the conformations of substituted sidechains. *Protein Eng.* **1**, 385-392.
- Summers, N.L., Carlson, W.D. & Karplus, M. (1987). An analysis of sidechain orientations in homologous proteins. *J. Mol. Biol.* **196**, 175-198.
- Reid, L.S. & Thornton, J.M. (1989). Rebuilding flavodoxin from $C\alpha$ coordinates: a test study. *Proteins* **5**, 170-182.
- Summers, N.L. & Karplus, M. (1989). Construction of sidechains in homology modeling: application to the C-terminal lobe of rhizopuspepsin. *J. Mol. Biol.* **210**, 785-811.
- Levitt, M. (1992). Accurate modeling of protein conformation by automatic sequence matching. *J. Mol. Biol.* **226**, 507-533.
- Wendoloski, J.J. & Salemme, F.R. (1992). PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries. *J. Mol. Graphics* **10**, 124-127.
- Dunbrack, R. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to sidechain prediction. *J. Mol. Biol.* **230**, 543-574.
- Eisenmenger, F., Argos, P. & Abagyan, R. (1993). A method to configure protein sidechains from the main-chain trace in homology modeling. *J. Mol. Biol.* **231**, 849-860.
- Laughton, C.A. (1994). Prediction of protein sidechain conformations from local three-dimensional homology relationships. *J. Mol. Biol.* **235**, 1088-1097.
- Bower, M.J., Cohen, F.E. & Dunbrack, R.L. (1997). Prediction of protein sidechain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**, 1268-1282.
- Wilson, C., Gregoret, L.M. & Agard, D.A. (1993). Modeling sidechain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996-1006.
- Koehl, P. & Delarue, M. (1994). Application of self-consistent mean field theory to predict protein sidechain conformations and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249-275.
- Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918-939.
- Vasquez, M. (1995). An evaluation of discrete and continuum search techniques for conformational analysis of sidechains in proteins. *Biopolymers* **36**, 53-70.
- Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* **66**, 549-579.
- Cheng, B., Nayeem, A. & Scheraga, H.A. (1996). From secondary structure to three-dimensional structure: improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins. *J. Comput. Chem.* **17**, 1453-1480.
- Gelin, B.R. & Karplus, M. (1979). Sidechain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry* **18**, 1256-1268.
- Kline, A.D., Braun, W. & Wuthrich, K. (1986). Studies by ^1H nuclear magnetic resonance and distance geometry of the solution conformation of the α -amylase inhibitor tendamistat. *J. Mol. Biol.* **189**, 377-382.
- Clore, G.M., et al., & Poulsen, F.M. (1987). Comparison of the solution and X-ray structures of barley serine protease inhibitor 2. *Protein Eng.* **1**, 313-318.
- Wagner, G., et al., & Wuthrich, K. (1987). Protein structures in solution by nuclear magnetic resonance and distance geometry: the polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *J. Mol. Biol.* **196**, 611-639.
- Smith, J.L., Hendrickson, W.A., Honzatko, R.B. & Sheriff, S. (1986). Structural heterogeneity in protein crystals. *Biochemistry* **25**, 5018-5027.
- Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: to be or not to be? An analysis of amino acid and sidechain configuration in globular proteins. *J. Mol. Biol.* **230**, 592-612.
- Brooks, B.R., et al., & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187.
- Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.
- Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902-1921.
- Gelin, B.R. & Karplus, M. (1975). Sidechain torsional potentials and motion of amino acid in proteins: bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA* **72**, 2002-2006.
- Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformations of amino acid sidechains in proteins. *J. Mol. Biol.* **125**, 357-386.
- Bhat, T.N., Sasisekheran, V. & Vijayan, M. (1979). An analysis of sidechain conformations in proteins. *Int. J. Pept. Protein Res.* **13**, 170-184.
- Dunbrack, R.L. & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Struct. Biol.* **1**, 334-340.
- Miller, S., Janin, J., Lesk, A.M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.
- Ichiye, T. & Karplus, M. (1988). Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for X-ray analysis. *Biochemistry* **27**, 3487-3497.

49. Kuriyan, J., *et al.*, & Karplus, M. (1991). Exploration of disorder in protein structures by X-ray restrained molecular dynamics. *Proteins* **10**, 340-358.
50. Kuriyan, J., Petsko, G.A., Levy, R.M. & Karplus, M. (1986). Effect of anisotropy and anharmonicity on protein crystallographic refinement. *J. Mol. Biol.* **190**, 227-254.
51. Wlodawer, A., Deisenhofer, J. & Huber, R. (1987). Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **193**, 145-156.
52. Flores, T.P., Orengo, C.A., Moss, D.S. & Thornton, J.M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811-1826.
53. Cregut, D., Liautard, J. & Chiche, L. (1994). Homology modelling of annexin I: implicit solvation improves sidechain prediction and combination of evaluation criteria allows recognition of different types of conformational error. *Protein Eng.* **7**, 1333-1344.
54. Lazaridis, T., Archontis, G. & Karplus, M. (1995). Enthalpic contribution to protein stability: atom-based calculations and statistical mechanics. *Adv. Protein Chem.* **47**, 231-306.
55. McGregor, M.J., Islam, S.A. & Sternberg, M.J.E. (1987). Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.
56. Tulinsky, A., Vandlen, R.L., Morimoto, C.N., Venkit Mani, N. & Wright, L.H. (1973). Variability in the tertiary structure of α -chymotrypsin at 2.8Å resolution. *Biochemistry* **12**, 4185-4192.
57. Moul, J., *et al.*, & Saya, A. (1976). The structure of triclinic lysozyme at 2.5Å resolution. *J. Mol. Biol.* **100**, 179-195.
58. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
59. Konner, J.H. (1976). A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Crystallogr. A* **32**, 614-617.
60. Morfiew, A.J. & Moss, D.S. (1982). RESTRAIN: a restrained least squares refinement program for use in protein crystallography. *Comput. Chem.* **6**, 1-3.
61. Sussman, J.L., Holbrook, S.R., Church, G.M. & Kim, S. (1977). A structure-factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. *Acta Crystallogr. A* **33**, 800-804.
62. Brunger, A.T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R-factor refinement by molecular dynamics. *Science* **235**, 458-460.
63. Jack, A. & Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallogr. A* **34**, 931-935.
64. Levitt, M. (1974). Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.* **82**, 393-420.
65. Brunger, A.T. & Karplus, M. (1988). Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins* **4**, 148-156.
66. Janin, J. (1990). Errors in three dimensions. *Biochimie* **72**, 705-709.
67. Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379.
68. Baumann, G., Frommel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329-334.
69. Tanimura, R., Kidera, A. & Nakamura, H. (1994). Determinants of protein sidechain packing. *Protein Sci.* **3**, 2358-2365.
70. Jorgensen, W.L., Chandrasekhar, J. & Madura, J.D. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935.
71. Hendrickson, W.A. (1985). Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol.* **115**, 252-270.
72. Tronrud, D.E., Ten Eyck, L.F. & Matthews, B.W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr. A* **43**, 489-501.
73. Wlodawer, A., Walter, J., Huber, R. & Sjolín, L. (1984). Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.* **180**, 301-329.
74. Hendrickson, W.A. & Teeter, M.M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulfur. *Nature* **290**, 107-113.
75. Mondragon, A., Wolberger, C. & Harrison, S.C. (1989). Structure of phage 434 cro protein at 2.35Å resolution. *J. Mol. Biol.* **205**, 179-188.
76. Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 Å. *J. Mol. Biol.* **195**, 555-579.
77. Smith, W.W., Burnett, R.M., Darling, G.D. & Ludwig, M.L. (1977). Structure of the semiquinone form of flavodoxin from clostridium MP. Extension of 1.8 Å resolution and some comparisons with the oxidized state. *J. Mol. Biol.* **17**, 195-225.
78. Thanki, N., *et al.*, & Wlodawer, A. (1992). Crystal structure of a complex of HIV-1 protease with a dihydroxyethylene-containing inhibitor: comparisons with molecular modeling. *Protein Sci.* **1**, 1061-1072.
79. Artymiuk, P.J. & Blake, C.C.F. (1981). Refinements of human lysozyme at 1.5 Å resolution. Analysis of non-bonded and hydrogen-bond interactions. *J. Mol. Biol.* **152**, 737-762.
80. James, M.N.G. & Sielecki, A.R. (1983). Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.* **163**, 299-361.
81. Borkakoti, N., Moss, D.S., Stanford, M.J. & Palmer, R.A. (1984). The refined structure of ribonuclease-A at 1.45 Å resolution. *J. Crystallogr. Spectr. Res.* **14**, 467-494.
82. Holmes, M.A. & Matthews, B.W. (1982). Structure of thermolysin refined at 1.6Å resolution. *J. Mol. Biol.* **160**, 623-639.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from <http://biomednet.com/cbiology/fad> — for further information, see the explanation on the contents pages.

Erratum

Protein sidechain conformer prediction: a test of the energy function

Robert J Petrella, Themis Lazaridis and Martin Karplus

Folding & Design 30 September 1998, 3:353–377

On page 366, first paragraph, the fourth sentence should read as follows:

The results indicate that the conformers around $(\chi_1, \chi_2) = (\chi_1, 40^\circ \text{ or } 45^\circ)$ and $(\chi_1 + 45^\circ, 195^\circ \text{ or } 200^\circ)$ are the most nearly symmetric or pseudosymmetric pairs.

On page 356, the bottom row of Table 3 should appear as follows:

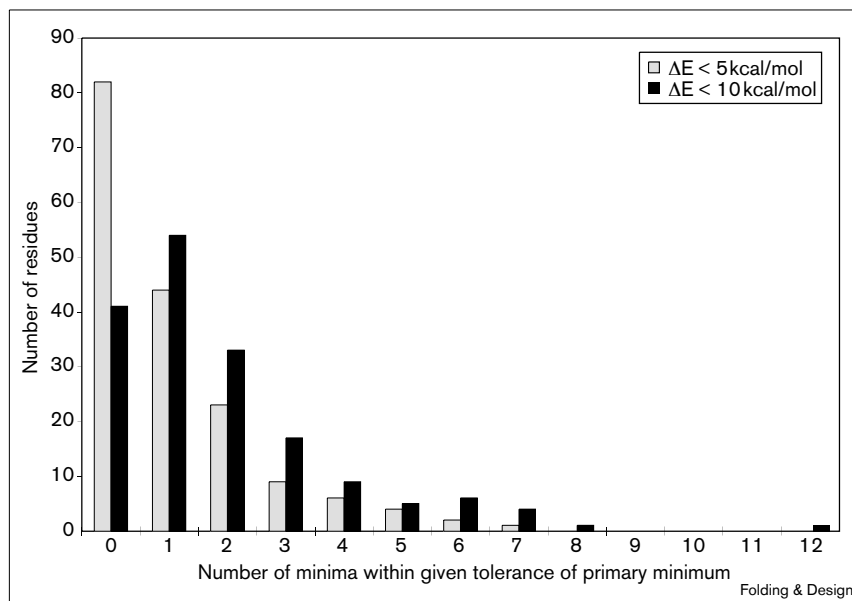
Table 3

Number of correctly predicted residues for c_1, c_2 rotations by residue type.

Amino acid type	All residues					Core residues				
	Number of residues	χ_1	%	$\chi_2 \chi_1$	% [†]	Number of residues	χ_1	%	$\chi_2 \chi_1$	%
All	1142	991		564		574	545		347	

On page 362, Figure 3, there should be visible a single sidechain with 12 alternative minima within 10 kcal/mol as follows:

Figure 3



Histogram of the distribution of energies in the relative minima occurring in sidechain dihedral energy maps of thermolysin. The plot counts minima of rank two and lower for sidechains having at least two heavy-atom dihedral angles. Degenerate χ_2 conformers in phenylalanine and aspartate are excluded. 52% and 76% of all residues have at least one local minimum within 5 kcal/mol and 10 kcal/mol, respectively, of their absolute minima.