Protein Science (1995), *4*:1217–1232. Cambridge University Press. Printed in the USA.

Building proteins from C_α coordinates using the dihedral probability grid Monte Carlo method

ALAN M. MATHIOWETZ[1] AND WILLIAM A. GODDARD III

Materials and Molecular Simulation Center, Beckman Institute (139-74),
Division of Chemistry and Chemical Engineering, California Institute of Technology,
Pasadena, California 91125

(RECEIVED December 29, 1994; ACCEPTED March 21, 1995)

Abstract

Dihedral probability grid Monte Carlo (DPG-MC) is a general-purpose method of conformational sampling that can be applied to many problems in peptide and protein modeling. Here we present the DPG-MC method and apply it to predicting complete protein structures from C_α coordinates. This is useful in such endeavors as homology modeling, protein structure prediction from lattice simulations, or fitting protein structures to X-ray crystallographic data. It also serves as an example of how DPG-MC can be applied to systems with geometric constraints. The conformational propensities for individual residues are used to guide conformational searches as the protein is built from the amino-terminus to the carboxyl-terminus. Results for a number of proteins show that both the backbone and side chain can be accurately modeled using DPG-MC. Backbone atoms are generally predicted with RMS errors of about 0.5 Å (compared to X-ray crystal structure coordinates) and all atoms are predicted to an RMS error of 1.7 Å or better.

Keywords: α carbons; Monte Carlo; protein modeling

The C_α coordinates of a protein provide a rough outline of its secondary and tertiary structure. Location of the C_α coordinates is an important early step in structural determination from X-ray crystallography (Jones et al., 1991), because these atomic positions can provide a framework for the rest of the structure. In addition, purely theoretical schemes to predict tertiary structure often use a simplified protein model containing only C_α coordinates (Friedrichs & Wolynes, 1989; Covell & Jernigan, 1990). Also, C_α coordinates can form a template for homology-based molecular modeling (Plaxco et al., 1989). However, the C_α coordinates do not provide sufficient information for understanding the most critical aspects of proteins such as binding and catalysis, which are determined by the chemical and steric properties of the protein backbone and side chains. Thus, it is necessary to provide a means for using the C_α coordinates of proteins to predict all other atomic coordinates.

Several methods for modeling complete protein structures from C_α coordinates have been published in recent years (Purisima & Scheraga, 1984; Reid & Thornton, 1989; Correa, 1990; Holm & Sander, 1991; Jones et al., 1991; Rey & Skolnick, 1992). The primary purpose for such methods is to speed and automate the process of building a protein model from crystallographic data (Jones et al., 1991), but several other uses have been suggested. Holm and Sander (1991) described how correct and incorrect protein folds can be evaluated by such methods, and Rey and Skolnick (1992) mentioned that their procedure may enable complete protein structures to be built from the C_α coordinates of a lattice representation. Our work was motivated by both of these factors: the desire to build full protein structures from lattice structures, and to provide a means for evaluating different lattice conformations. In addition, the "DPG Protein Builder" described here has been useful for homology modeling because it allowed us (Plaxco et al., 1989) to build a model of Hin recombinase from the C_α coordinates of λ Cro.

The process of building full protein conformations from C_α coordinates requires success in two areas: prediction of backbone conformations in the presence of explicit geometric constraints (the known C_α coordinates) and prediction of side-chain conformations constrained only by the conformation of the backbone and the presence of other side chains. Our method provides a consistent approach to solving both problems. Based primarily on Monte Carlo conformational searching, our technique differs significantly from previously published techniques, which range from the purely geometric (Purisima & Scheraga, 1984; Rey & Skolnick, 1992), to methods based primarily on database

Reprint requests to: William A. Goddard III, Materials and Molecular Simulation Center, Beckman Institute (139-74), Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125; e-mail: wag@wag.caltech.edu.

[1] Present address: Central Research Division, Pfizer, Inc., Groton, Connecticut 06340.

searches of several consecutive residues (Reid & Thornton, 1989; Holm & Sander, 1991; Jones et al., 1991), to molecular mechanics (Correa, 1990).

Our procedure for building protein structures from C_α coordinates uses the conformational probabilities of individual residues, rather than groups of residues. Thus, it does not depend upon the prior existence of fragments in the protein database that happen to have the same C_α geometries as those we are trying to fit. We use the dihedral probability grid Monte Carlo (DPG-MC) method to build first the backbone conformation (DPG-BACK) then the side chains (DPG-SIDE). The specific application of DPG-MC to the problem of modeling the complete structure of a protein from C_α (CA) coordinates is termed the DPG Protein Builder. The DPG-MC method modifies protein conformations one residue at a time, by choosing either new backbone (ϕ, ψ) or side-chain (χ) dihedral angles from probability matrices. In the DPG-BACK phase, the backbone is built one residue at a time. As the protein chain grows, the conformational space of the backbone is sampled using (ϕ, ψ) probability grids. The DREIDING force field (Mayo et al., 1990) is used to evaluate the energy of each structure, with additional harmonic constraint terms added between the template C_α coordinates and the C_α coordinates of the growing chain. After the entire backbone is built in this way, side-chain positions are optimized during a second DPG-MC simulation. The DPG-SIDE phase uses χ probability grids to modify one side-chain conformation at a time. Because DPG-MC uses random numbers both to determine whether new conformations are accepted or rejected and to choose new conformations, each run produces different results. Therefore, we generate numerous backbone conformations and select those with the best energy to use in the DPG-SIDE stage. Likewise, for each backbone conformation, several independent DPG-SIDE simulations are carried out and the structure (backbone and side chains) with the best overall energy is selected as the optimum model.

Results and discussion

Crambin

Our method for calculating complete protein structures from C_α coordinates is described in detail in the Methods section. The method was used to calculate several complete structures, ranging in size from crambin (46 amino acid residues) to myoglobin (153 residues), from the crystallographic C_α coordinates, and the results were compared to the full crystallographic structures. We used the "united atom" representation in which all heavy atoms and those hydrogens attached to heteroatoms are represented explicitly, whereas hydrogens attached to a carbon merely are represented implicitly as part of the carbon atom. The full structure of crambin was calculated using the C_α coordinates from the crystal structure (Hendrickson & Teeter, 1981). In the first phase, the DPG-BACK method was used to generate 20 different backbone conformations. Each conformation was generated using a different series of random numbers to control the selection of (ϕ, ψ) dihedrals as well as to determine which conformations would be accepted and which rejected. The conformational energies of the backbone, the RMS deviations (RMSDs) in backbone atoms, and (ϕ, ψ) dihedrals from each of these structures are listed in Table 1, ranked by energy. The average backbone RMSD for these 20 simulations was 0.527 Å.

Table 1. *Energy and RMSDs (atoms and dihedrals) for each of the 20 backbone conformations generated by DPG-BACK for crambin*

Energy (kcal/mol)	Atoms (Å)	Dihedrals[a] (deg)
335.3	0.494	22.05
338.4	0.430	19.43
363.3	0.543	25.75
363.8	0.495	26.00
366.4	0.515	28.69
376.9	0.576	29.40
377.6	0.545	29.88
393.2	0.582	32.96
465.5	0.668	42.27
577.1	0.483	28.94
597.6	0.481	31.15
652.7	0.572	33.77
796.9	0.588	33.13
797.1	0.430	21.49
822.7	0.498	27.47
850.3	0.505	27.74
872.4	0.595	33.38
1,445.3	0.589	32.08
5,266.2	0.447	27.67
5,700.5	0.513	34.44

[a] RMSD in (ϕ, ψ) dihedrals.

The average all-atom deviation was 1.696 Å. "All-atom" RMSDs refer to deviations in all the atoms represented explicitly in the united atom approach. It is apparent that there is only a small correlation between the backbone energy and the RMS fit to the crystal structure backbone. The backbone of the crystal structure itself has an energy of 759.8 kcal/mol, higher than 12 of the 20 model conformations. This is likely due to limitations of the force field, to effects of crystal packing and solvation, and to errors in the crystal structure. Nevertheless, in cases where the crystal structure is unknown, the backbone energy is the best criterion for selecting model structures. Other possible selection criteria, including C_α constraint energy and total energy including side-chain atoms, had even worse correlation with the deviation in the backbone coordinates (unpubl. data).

The five lowest-energy backbone conformations from DPG-BACK (Table 1) were used as a starting point for the DPG-SIDE phase. For each of the five backbone conformations, five DPG-SIDE simulations were carried out, using different random numbers. Each simulation involved 1,000 Monte Carlo steps using 10° probability grids and a simulation temperature of 300 K. The 25 conformations produced are listed in Table 2. Again, there is only a small correlation between energy and RMS fit to the crystal structure. Nevertheless, the fits are quite good, with an average RMSD from the crystal structure of 1.323 Å. All five backbone conformations were represented throughout the list of complete structures, so the backbone energy was not the determining factor in the overall energy.

The best energy conformation from the side-chain phase was chosen as the "model" conformation of crambin for detailed comparison to the "true" structure, the crystal structure (Hendrickson & Teeter, 1981). Table 3 gives a breakdown of the

Table 2. *Energy and RMSDs in atomic coordinates for each of the 25 crambin models produced by the DPG Protein Builder*

Energy	RMSD	Energy	RMS
668.1	1.386	1,039.0	1.337
669.2	1.367	1,074.0	1.519
688.2	1.132	1,111.8	1.153
691.6	1.259	1,304.6	1.332
706.6	1.313	1,696.1	1.468
757.3	1.170	2,225.6	1.272
767.8	1.449	2,576.8	1.393
793.9	1.430	3,023.2	1.486
801.3	1.278	3,077.1	1.391
823.0	1.243	3,105.8	1.487
860.7	1.297	3,334.5	1.221
947.9	1.111	3,383.6	1.484
971.7	1.102		

RMSD of the crambin model for different regions of the protein. Some of this information is shown graphically in Figure 1, where the backbone RMSD of each residue is shown. The largest deviations occur at the carboxyl-terminus, where residues 45 and 46 are very poorly modeled, especially considering that the C_α atoms, because of the constraining force, have a deviation from the crystal structure of less than 0.05 Å. Excluding these two residues, the backbone RMSD drops from 0.543 Å to 0.361 Å. The carboxyl-terminal residues are generally the worst modeled residues because there are fewer constraints on the structure. They usually lie on the surface of the protein, where there are fewer interresidue contacts and there is no $I + 1$ C_α to constrain the orientation of the terminal carboxyl group. In the crambin model, the Asn 46 side chain and the terminal carboxyl group have reversed positions, giving rise to a large error even though the chemical significance is small. The backbone RMSD is fairly consistent throughout the rest of the protein, with 34 of the 46 residues having deviations in the 0.1–0.4-Å range. The lowest backbone deviations are in the residues of the long α-helix, Helix 1, where the deviation in atomic coordinates is 0.209 Å, and the deviation in (ϕ, ψ) dihedrals is only 13.7°. The deviations are equally low (0.232 Å and 13.1°) for the first seven residues of Helix 2. However, the last residue in the helix starts a turn and is poorly modeled. In general, the turn regions before and after α-helices are the most poorly modeled residues other than those at the C-terminus. This is very apparent from both the graph in Figure 1 and the picture in Figure 3. These regions (particularly residues 5, 20, and 30) have nonstandard (ϕ, ψ) values that have very low probabilities in the (ϕ, ψ) probability grids. No (ϕ, ψ) probability grids were specifically developed for turn regions, but these might prove very valuable.

The side-chain modeling is not as successful as the backbone modeling, with the average deviation in atomic coordinates being near 2.0 Å. This is not surprising because the backbone is more highly constrained than the side chains: each peptide unit in the polypeptide backbone is covalently constrained at both ends by the positions of two consecutive C_α's, whereas the side chains are usually constrained by only a single covalent attachment to a C_α. The constraints on the side-chain conformations are primarily steric in nature: side chains in the interior of a protein can have considerable steric overlap and their conformations must be correlated to allow for closest packing. The DPG-SIDE calculations are also much slower than the DPG-BACK and far fewer conformations are sampled per dihedral angle. Figure 2 shows the residue-by-residue side-chain RMS for the crambin model. Two side chains stand out: Arg 17, analyzed below in the discussion of Figure 5, and Asn 46, the C-terminal residue that in the model has the side chain and C-terminal carboxyls flipped, as mentioned above.

Another measure of the modeling accuracy of the side chains is the deviation in side-chain dihedral angles, χ, defined as the absolute value of the difference between the dihedral in the model and in the crystal structure. Of the 37 χ^1's in crambin, 24 have deviations less than 30°. Eleven of the χ^1's have deviations between 90° and 150°, indicating a rotation from one minimum to the next. Only two have deviations between 30° and 90°. It is important to note that 5 of the 13 side chains with side-chain deviations greater than 30° are cysteine residues involved in disulfide bridges in the crystal structure. The DPG Protein Builder does not currently account for the presence of disulfide

Table 3. *RMSDs for different regions of the crambin model*

Region	Residues	Backbone (Å)	(ϕ, ψ) (deg)	All atoms (Å)	Side chains (Å)
All	1-46	0.543	25.8	1.386	2.010
No C-term	1-44	0.361	23.0	1.248	1.841
Helix 1	7-19	0.209	13.7	1.658	2.347
Helix 2	23-30	0.394	22.3	1.026	1.454
Sheet 1	1-4	0.417	22.3	1.146	1.771
Sheet 2	32-35	0.315	19.2	1.070	1.530
Turn 1	41-44	0.571	32.9	1.853	1.853
N-terminus	1-2	0.559	31.1	1.184	1.688
C-terminus	45-46	1.872	67.9	3.175	4.682
Coil	Others	0.373	28.5	0.511	0.728

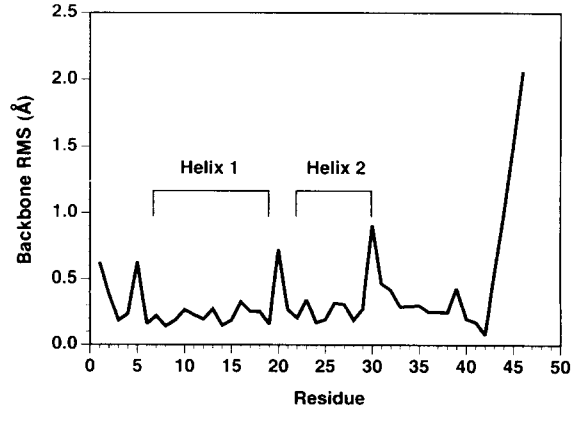

Fig. 1. Distribution by residue of backbone RMS errors for the crambin model relative to the crystal structure.

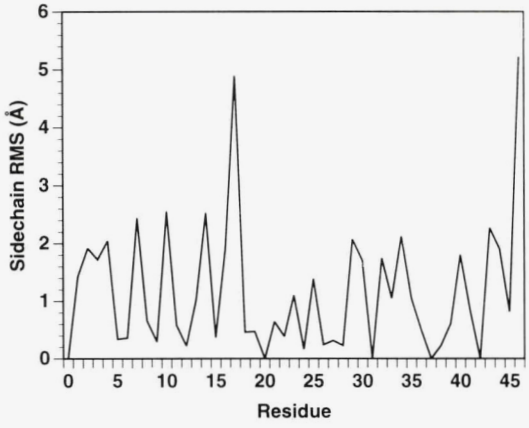

Fig. 2. Side-chain RMS errors for the crambin model relative to the crystal structure.

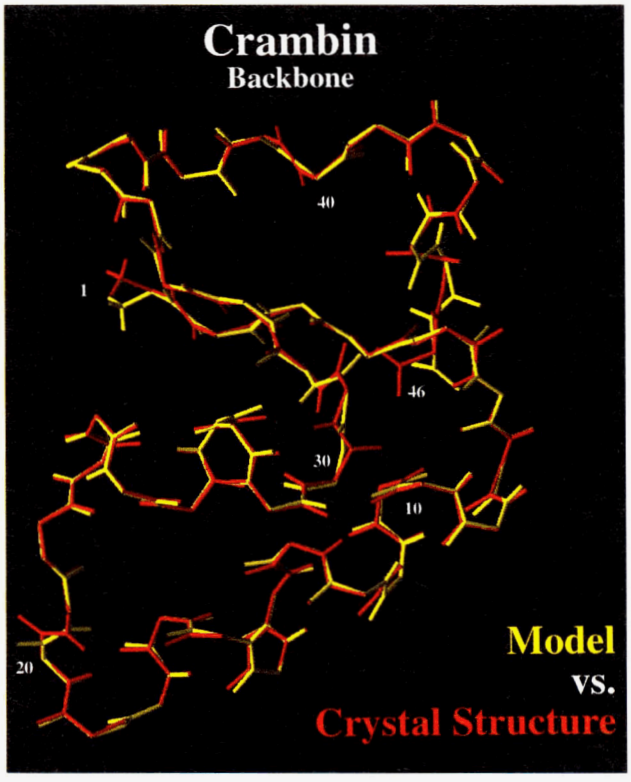

Fig. 3. Peptide backbone of the model and crystal structures of crambin. RMSD is 0.538 Å.

bridges. The disulfide bonds are not included in the Monte Carlo energy evaluations. Such a term could be included and would certainly improve the results for these residues. RMSDs for the different backbone and side-chain dihedrals are shown in Table 4. Although the side-chain dihedrals are not as well modeled as the backbone, the results are encouraging with respect to other methods. As discussed below, the DPG Protein Builder provides results for flavodoxin χ dihedrals as good or better than other methods, and these results for crambin are better still.

Differences between the crambin model and the crystal structure are shown in detail in Figures 3, 4, and 5. Figure 3 shows the model and crystal structure backbones for the entire protein. For most of the protein, it is difficult to distinguish between the two structures. Only in the turn regions after the two helices is the difference readily apparent. The two following figures show the complete structures of the two helices of crambin. Helix 2, shown in Figure 4, is very well modeled, with an all-atom RMSD of 1.03 Å. In terms of the all-atom deviation, it is the best modeled region of the protein (see Table 3). The picture shows this quite well, with both side-chain and backbone atoms showing little difference between the two structures, except for Thr 30 on the C-terminal (right) end of the helix. As explained above, this residue begins a turn in the backbone conformation and is poorly sampled during the DPG-BACK phase. The Helix 1 backbone, in contrast, is modeled quite well throughout its length, including Pro 19 at its C-terminal end. However, Helix 1, shown in Figure 5, has many large side chains that are difficult to model. Large errors can be seen in Asn 14 and Arg 17. The latter has a particularly large impact on the RMSD. Excluding Arg 17, the crambin model has an RMSD of 1.207 Å, rather than 1.386 Å. However, this incorrect conformation of Arg 17 may be energetically more favorable than other conformations more similar to the crystal structure. Of the next four lowest-energy conformations listed in Table 2, all five have more native-like conformations of Arg 17, but all are higher in energy.

Table 4. *RMSDs in various types of dihedrals for the crambin model and percentages of each type of dihedral with deviations less than 30° or more than 90°*

Dihedral	Number	RMSD (deg)	Deviation <30° (%)	Deviation >90° (%)
ϕ	45	22.3	86.7	0.0
ψ	45	28.8	75.6	2.2
ω	45	5.4	100.0	0.0
χ^1	37	69.6	62.2	29.7
χ^2	21	84.5	38.1	28.6
χ^3	8	75.1	25.0	37.5
χ^4	7	34.9	71.4	0.0
χ^5	2	9.8	100.0	0.0

The crambin model illustrates several general findings for simulations using the DPG Protein Builder. The lowest-energy structures from the DPG-BACK and DPG-SIDE phases are usually among the best models built, but are not necessarily the very best. Regardless, the backbone models from DPG-BACK are consistently good, and almost any one of them provides an acceptable model of the true backbone. The model backbones are especially good in regions of regular secondary structure such as helices and sheets, but rather poor in turn regions. These results are obtained consistently in different simulations. There is a much larger variation among the results from DPG-SIDE. This may be due to the constraints of time; the number of 1,000 Monte Carlo steps was selected largely in order to keep the simulation time below 10 min, so that large numbers of different

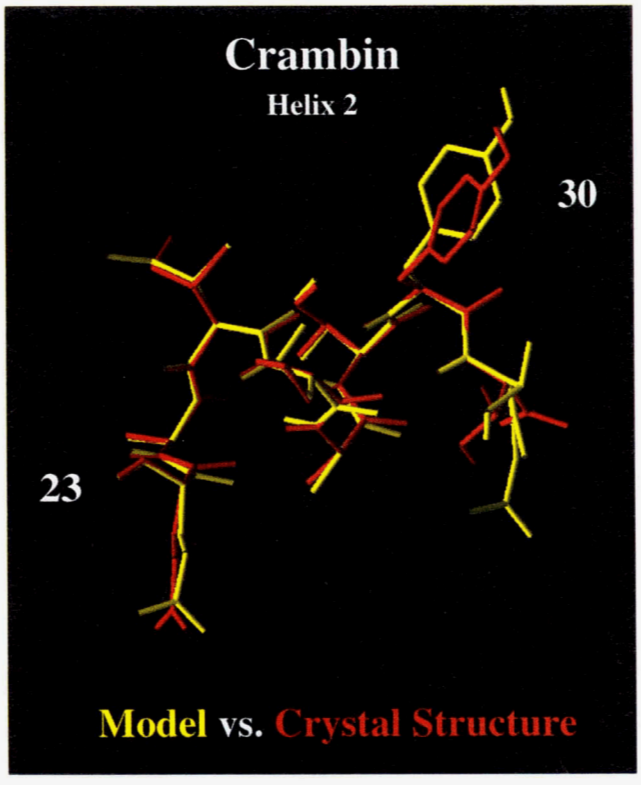

Fig. 4. Comparison of helix 2 (residues 23–30) in the model and crystal structures. RMSD is 1.026 Å for all atoms and 0.394 Å for backbone atoms.

conformations could be evaluated. Better and more consistent results might be obtained by substantially longer calculations. Nevertheless, between 40% and 60% of χ^1 dihedrals are modeled correctly.

Larger proteins

Although the variables discussed in the Methods section could be tuned to specific classes of proteins, the same values were used for six different proteins listed in Table 5. These proteins have widely different structures, as indicated by the percentages of their secondary structures that are α-helical and β-sheet. Four of the six proteins are included in the subset of 64 crystal structures used to develop the Monte Carlo probability grids. Of the other two, the flavodoxin structure is merely a different form (oxidized) than the one used in the data set (semiquinone), and the plastocyanin is homologous, but not identical, to a structure used in the data set. It is unlikely that this has any significant effect on the results because individual ϕ, ψ, and χ values from any one structure have only a small influence on the probabilities used in the conformational sampling.

For each of these six proteins, the C_α coordinates from the listed crystal structure were used to rebuild the backbone conformation 20 times, as described in the preceding sections for crambin. In each case, all prosthetic groups, such as the myoglobin heme, were removed from the crystal structure, as were any cofactors or solvent molecules. Each of the 20 backbone conformations was compared to the crystal structure and the results were analyzed. Table 6 lists the average RMSD as well as the standard deviation (σ) for the 20 structures. Also listed are the RMSDs for the lowest energy conformation and the conformation with the best fit. Again, it is seen that the lowest energy conformation is never the one with the best fit to the crystal structure. However, the lowest energy conformation was better than average for five of the six proteins. We were not able to identify any systematic differences between the low energy structures and the best fit structures. During homology modeling or crystallographic model building, it would be best to try

Table 5. *Proteins modeled using the DPG Protein Builder*[a]

Protein	PDB	Reference	Size	% Helix	% Sheet
Crambin	1CRN	[1]	46	45.7	17.4
BPTI	5PTI	[2]	58	27.6	25.9
Plastocyanin	7PCY	[3]	98	7.1	58.2
Ribonuclease A	7RSA	[4]	124	26.7	46.8
Flavodoxin	3FXN	[5]	138	37.7	26.8
Myoglobin	1MBD	[6]	153	79.1	0.0

[a] The reference crystal structure is given along with the number of residues in the protein and the percentage of these that are in α-helices and β-sheets. References: [1], Hendrickson and Teeter (1981); [2], Wlodawer et al. (1984); [3], Collyer et al. (1985); [4], Wlodawer et al. (1988); [5], Smith et al. (1977); [6], Phillips (1980).

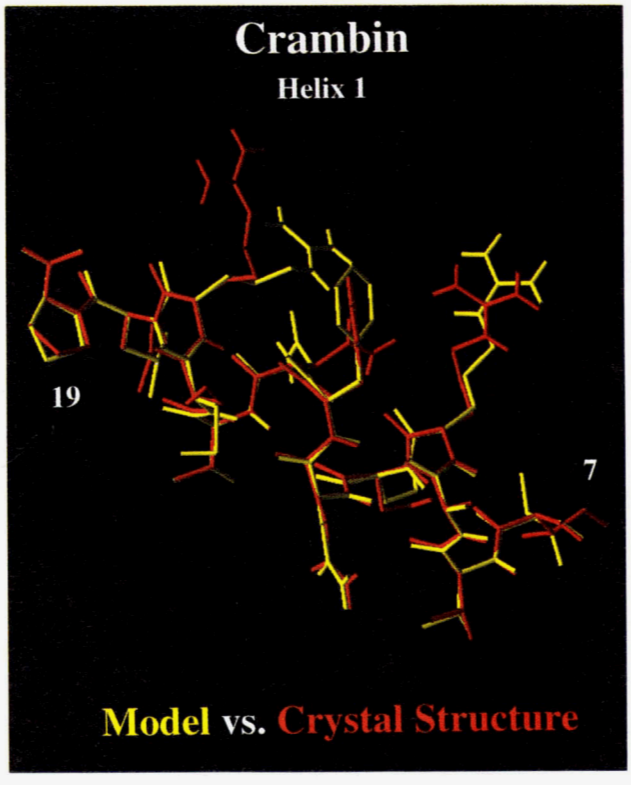

Fig. 5. Helix 1 (residues 7–19) in the model and crystal structures. RMSD is 1.658 Å for all atoms and 0.209 Å for backbone atoms.

Table 6. *Results from DPG-BACK constructions of the backbone conformations for several proteins*

Crystal structure	Backbone RMSD			
	Average	σ	Best E	Best fit
1CRN	0.527	0.062	0.494	0.430
5PTI	0.610	0.065	0.582	0.506
7PCY	0.550	0.048	0.602	0.470
7RSA	0.601	0.052	0.551	0.530
3FXN	0.593	0.050	0.577	0.509
1MBD	0.453	0.033	0.451	0.366

several of the backbone models, rather than merely the lowest energy one, in order to provide several templates for side-chain modeling or refinement.

Comparing Tables 5 and 6, it is clear that the size of the protein has little effect on the accuracy of DPG-BACK. In fact, the largest protein, myoglobin, is consistently modeled most accurately. This is not surprising considering the crambin results, where the average backbone deviations were approximately 0.2 Å for helical residues. The protein myoglobin, with almost 80% of its residues in α-helices, is greatly benefited by the accuracy with which the method models helices. Plastocyanin is also modeled relatively well, even though it is a β-sheet protein, with little helical content. The large β-sheet content is probably also a favorable factor, as these conformations are also very well represented by the probability grids. It is proteins such as bovine pancreatic trypsin inhibitor (BPTI), with only about 50% α-helix and β-sheet content, which are relatively poorly modeled. Even in this case, most of the protein is modeled quite accurately and the overall RMSD is greatly increased by the poor modeling of the C-terminal residues. The average RMSD for residues 1–54 is 0.501 Å.

DPG-SIDE simulations were carried out on flavodoxin and plastocyanin, building five complete structures from each of the top five backbone conformations from DPG-BACK. The same parameters were used for these simulations as were used for DPG-SIDE simulations of crambin. The energy and all-atom RMSD for each of the 25 conformations were evaluated and the results were analyzed. Table 7 lists the results for these two proteins, along with those for crambin. Unlike DPG-BACK, the results for DPG-SIDE are highly dependent on the size of the protein, with the average deviation increasing substantially for larger proteins. In DPG-BACK simulations, each residue was sampled the same number of times, regardless of the size of the protein. In the DPG-SIDE simulations, however, each simulation involved a total of 1,000 Monte Carlo steps. For crambin, this meant that the average residue was varied 27 times during the simulation (alanine and glycine residues are not affected). For plastocyanin, the 73 relevant dihedrals were sampled an average of 14 times; for flavodoxin, the average was 8.5. Clearly, the side chains of flavodoxin are not being adequately sampled. Unfortunately, the cpu time required for the simulations also grows substantially as the size of the protein grows. Although the 1,000 Monte Carlo steps take 7 min for crambin, they require nearly 20 min for plastocyanin (on one processor of a Silicon Graphics 4D/380 workstation) and more than 40 min for flavodoxin. Therefore, it is computationally expensive to increase the number of steps for flavodoxin. Nevertheless, the results for flavodoxin are comparable to or better than published results using other methods.

Table 7. *Results from DPG-SIDE constructions of the side chains of crambin, plastocyanin, and flavodoxin*

Crystal structure	All-atom RMSD		
	Average	Best E	Best fit
1CRN	1.323	1.386	1.102
7PCY	1.483	1.398	1.299
3FXN	1.796	1.663	1.607

The lowest energy conformation of flavodoxin was chosen for comparison with other methods. This protein has become a standard test case for published methods of building complete structures from C_α coordinates. This includes both methods based on molecular mechanics (Correa, 1990) and those using database searches to determine conformations for multiple-residue peptide fragments from the protein (Reid & Thornton, 1989; Holm & Sander, 1991). Table 8 lists several measures of the accuracy of these models. "Peptide flips" refer to the number of peptide units (the planar backbone unit between the C_α coordinates) that are rotated by more than 90° from the crystal structure. This occurs seven times in our model, compared to only five and four times in the fragment-matching methods (Reid & Thornton, 1989; Holm & Sander, 1991). This is the only measurement by which the DPG Protein Builder appears deficient, using flavodoxin as the case study. The other proteins we studied did not have such a large number of peptide flips. The lowest energy structures of these proteins had between zero (crambin) and five (plastocyanin and ribonuclease A) peptide flips. In most of the other measures, the DPG Protein Builder is comparable to, or better than, the other published methods. It is currently not quite as accurate as the method of Holm and Sander (1991) but is comparable in most respects, even though it is based on a more general approach to protein modeling: dihedral probability grid Monte Carlo. The DPG-MC method is applicable to

Table 8. *Comparison of the results for flavodoxin versus other methods*

Atoms	Reference			DPG model
	[RT]	[C]	[HS]	
RMSD, all atoms (Å)	1.73	1.64	1.57	1.66
RMSD, main chain (Å)	0.57	0.49	0.48	0.57
RMSD, side chain (Å)	2.41	–	2.19	2.31
Peptide flips	5	–	4	7
Correct χ^1 (%)	40	–	44	41
Correct χ^1, χ^2 (%)	17	–	25	24

[a] "Correct" refers to dihedrals predicted to within 20° of their crystal structure values. [RT], Reid and Thornton (1989); [C], Correa (1990); [HS], Holm and Sander (1991).

unconstrained systems as well as those constrained by a priori knowledge of the C_α coordinates.

Conclusions

The DPG Protein Builder is a new method for building complete protein structures from C_α coordinates using DPG-MC. Most of the previous methods (Reid & Thornton, 1989; Holm & Sander, 1991; Jones et al., 1991) use database searches to find conformations for several consecutive residues that match the configuration of the C_α coordinates being used as a template. DPG-MC, in contrast, uses probabilities for individual residues to guide Monte Carlo searches. The DPG Protein Builder produces results as good as or better than previously published methods for the protein flavodoxin. In general, backbone conformations are modeled accurately to within 0.6 Å RMSD from the crystal structure. Most of the error comes at the C-terminal ends and in turns, whereas the extended secondary structures (α-helices and β-sheets) are modeled much better, with a typical RMSD of 0.3 Å or better. Side-chain conformations are not modeled as accurately. Side-chain RMSDs greater than 2.0 Å can be expected for large proteins, where the computational cost of optimizing all side chains concurrently is very large. The side-chain deviation for the small protein crambin was much better, averaging 1.87 Å for 25 models. Overall RMSDs are typically better than 2.0 Å, and depend primarily upon the amount of time spent optimizing the side-chain conformations. The calculations performed here were not optimized for accuracy alone but for speed as well. In real-world cases where the best possible model is desired, it would be possible to significantly increase the number of conformations sampled in both the DPG-BACK and DPG-SIDE stages, thereby improving the accuracy of both the backbone and side chains.

Methods

Dihedral probability grid Monte Carlo

DPG-MC is a method developed for predicting the conformations of peptides and proteins by searching their torsional degrees of freedom. The DPG-MC method combines two of the best features from other torsion-space conformational search methods developed to study peptide conformations: Monte Carlo importance sampling and grid searching. Like the importance sampling method of Lambert and Scheraga (1989) and biased probability Monte Carlo (Abagyan & Totrov, 1994), the method described here assigns probabilities to different (ϕ, ψ) combinations, and conformations are generated according to those probabilities, rather than completely at random or through an exhaustive search of all possibilities. However, unlike either of these methods, our probabilities are designed to work within the framework of a grid search method, i.e., only discrete values are chosen for the dihedral angles. There are three primary advantages to using discrete values for dihedral angles, rather than sampling from a continuum: (1) the conformational space is reduced to a finite number of possible conformations per dihedral angle, (2) the probabilities can be generated to reflect known (ϕ, ψ) distributions more accurately because they are not forced to fit a functional form, and (3) the method is easily extended to side chain (χ) dihedrals. Because no functional form is necessary to specify the probabilities, grids can be developed for any necessary dimensionality. They range from one-dimensional grids for small side chains to five-dimensional grids for arginine.

Grid searches have been employed in many conformational studies, such as those designed to predict protein loop structures (Bruccoleri & Karplus, 1987) and those employed in the study of organic molecules (Lipton & Still, 1988). The conformational space in a grid method is still large, as each dihedral can assume $360/S$ conformations, where S is the grid spacing. Therefore, these methods usually employ sophisticated schemes for eliminating combinations that cause steric overlap. In contrast, the DPG-MC method implicitly includes a great deal of steric information through the use of probability grids: probabilities are assigned to different protein backbone (ϕ, ψ) and side chain (χ) dihedrals according to their distributions in known protein structures. The sampling is biased toward the sterically allowed amino acid conformations seen in nature, so the simulation focuses on optimization of long-range interactions.

In the DPG-MC method, conformations of a peptide or protein are generated by rotating backbone (ϕ, ψ) and/or side chain (χ) dihedral angles of individual amino acids. The conformations are not chosen randomly, but are selected from probability grids calculated from a selected subset of proteins from the Brookhaven Protein Data Bank (PDB). Each grid is an N_d-dimensional matrix, where N_d is the number of dihedrals involved. For instance, backbone sampling involves two-dimensional grids, and each point on the grid is the probability of choosing a particular (ϕ, ψ) pair. The grids have $S°$ spacing, where $S = 5$, 10, 15, 30, or 60. Therefore, (ϕ, ψ) grids have N_S points, where $N_S = (360/S) \times (360/S)$. The probabilities were derived from a set of high-resolution protein crystal structures by partitioning every (ϕ, ψ) pair into S-degree bins. The probabilities, $P(\phi, \psi)$, are normalized so that

$$\sum_{i=1}^{360/S} \sum_{j=1}^{360/S} P(\phi_i, \psi_j) = 1. \quad (1)$$

Side-chain probability grids have varying dimensionality, depending upon the number of dihedrals needed to specify the conformation. This ranges from $N_d = 1$ for small side chains like serine and threonine, to $N_d = 5$ for arginine. For alanine and glycine, $N_d = 0$.

The PDB subset

Dihedral probabilities integral to DPG-MC must be based on a judicious choice of structural data that are both diverse and accurate. The PDB now contains more than 2,000 protein crystal structures; however, many proteins are represented numerous times or are highly homologous to other proteins in the PDB data set. Including identical or nearly identical structures would distort the probability distribution in favor of geometries found in those particular proteins. In order to eliminate highly redundant structures, we carried out pairwise sequence comparisons among 503 proteins in our initial PDB data set using the "Align" program from W.R. Pearson's FASTA sequence analysis package (Pearson & Lipman, 1988). Any protein with greater than 25% sequence identity with another protein of higher resolution was eliminated. This homology elimination process reduced our data set from 503 proteins to 121. We further reduced the data

Table 9. *Crystal structures used in the H64 data set*

PDB code	Resolution (Å)	R	PDB code	Resolution (Å)	R	PDB code	Resolution (Å)	R
1AMT	1.5	0.155	1UBQ	1.8	0.176	3BLM	2.0	0.163
1BP2	1.7	0.171	1UTG	1.34	0.23	3CLA	1.75	0.157
1CRN	1.5	0.114	1XY1	1.04	0.088	3DFR	1.7	0.152
1CSC	1.7	0.188	256B	1.4	0.164	3GRS	1.54	0.186
1CSE	1.2	0.178	2AZA	1.8	0.157	3RNT	1.8	0.137
1CTF	1.7	0.174	2CA2	1.9	0.176	451C	1.6	0.187
1ECA	1.4	0.183	2CCY	1.67	0.188	4CPV	1.5	0.215
1FB4	1.9	0.189	2CDV	1.8	0.176	4FD1	1.9	0.192
1GD1	1.8	0.177	2CPP	1.63	0.19	4FXN	1.8	0.200
1GMA	0.86	0.071	2CYP	1.7	0.202	4INS	1.5	0.153
1GP1	2.0	0.171	2ER7	1.6	0.142	4PTP	1.34	0.171
1HOE	2.0	0.199	2GBP	1.9	0.146	5CPA	1.54	0.190
1I1B	2.0	0.189	2LTN	1.7	0.177	5CYT	1.5	0.171
1L19	1.5	0.153	2MHR	1.7	0.158	5PTI	1.0	0.200
1LZ1	1.5	0.177	2MLT	2.0	0.198	5RXN	1.20	0.115
1MBA	1.6	0.193	2OVO	1.5	0.199	5TNC	2.0	0.155
1MBD	1.4	0.188	2RSP	2.0	0.144	6TMN	1.6	0.171
1NXB	1.38	0.24	2SGA	1.5	0.126	7RSA	1.26	0.15
1PAZ	1.55	0.18	2SNS	1.5	N.A.	9PAP	1.65	0.161
1PCY	1.6	0.17	2WRP	1.65	0.180	9WGA	1.8	0.175
1PPT	1.37	0.279	3B5C	1.5	0.16			
1THB	1.5	0.196	3BCL	1.9	0.189			

set to 64 high-quality crystal structures that had 1.5 Å resolution data or better or had better than 2.0 Å resolution and *R*-factors below 20%. This data set, which we call H64, was used to create our probability grids in this work. The 64 crystal structures comprising this data set are listed in Table 9.

Backbone (ϕ, ψ) probability grids

The backbone probability grids were determined by partitioning every (ϕ, ψ) pair in the proteins comprising the H64 data set into bins of size $S° \times S°$ and normalizing. We have determined separate probability grids for each amino acid, but it is sufficient to use individual grids for the three major residue types: glycine, which has no side chain; proline, whose side chain forms a closed loop with the backbone; and the other 18 "standard" L-amino acids. The (ϕ, ψ) probabilities are significantly different for these three residue types, as can be seen in Figure 6. The shape of the grid depends not only on the data, but on the grid spacing, *S*. A narrower spacing allows for much greater conformational flexibility, which is especially important in simulations of constrained systems. It is clear from Figure 6 that no simple functional form would accurately represent the (ϕ, ψ) probabilities seen in protein crystal structures.

We have also used the secondary structure designators in the protein database (HELIX, SHEET, and TURN) to obtain separate probability grids for the α-helix, β-sheet, and random coil structural classes. Coil residues were defined as those not marked as belonging to HELIX, SHEET, or TURN regions. We decided not to create grids for β-turn residues because the four residues involved in a turn usually have completely different (ϕ, ψ) conformations and it would be counterproductive to treat them identically. Eight-dimensional probability grids generated for sequences of four consecutive (ϕ, ψ) pairs would have peaks for particular turn conformations as well, but the total number of turns in our set of crystal structures is tiny compared to the immense number of grid points on an eight-dimensional grid. Such grids would have little advantage over a method that simply tries all known turn configurations. Six proteins in the H64 database had no HELIX, SHEET, or TURN designators, and we excluded these from secondary structure analyses. The remaining 58 proteins with secondary structure designators comprise the SS58 data set, which we used to create the probability grids shown in Figure 7. The coil grid in Figure 7 contains significant probabilities for both α-helix and β-sheet conformations, but the probabilities are much lower than those in the "all-structures" grid. Presumably, residues in the coil regions are not participating in the extended hydrogen bonding networks or involved in the large-scale dipole–dipole interactions of α-helices and β-sheets. Therefore, the coil probability grids are more indicative of the inherent conformational energies of individual residues and, therefore, are the grids that most closely resemble classic Ramachandran plots (Ramachandran et al., 1963) and (ϕ, ψ) potential energy maps (Brant et al., 1967). These secondary structure-specific grids are useful only when the secondary structure is known beforehand. This is not the case for an ab initio prediction of protein conformation, but is for simulations used in conjunction with C_α coordinates, homology modeling, or secondary structure prediction algorithms.

Side-chain (χ) probability grids

Although every amino acid backbone can be specified by the same three dihedral angles, ϕ, ψ, and ω, there is a far greater diversity among side-chain dihedrals, χ. At the extremes are gly-

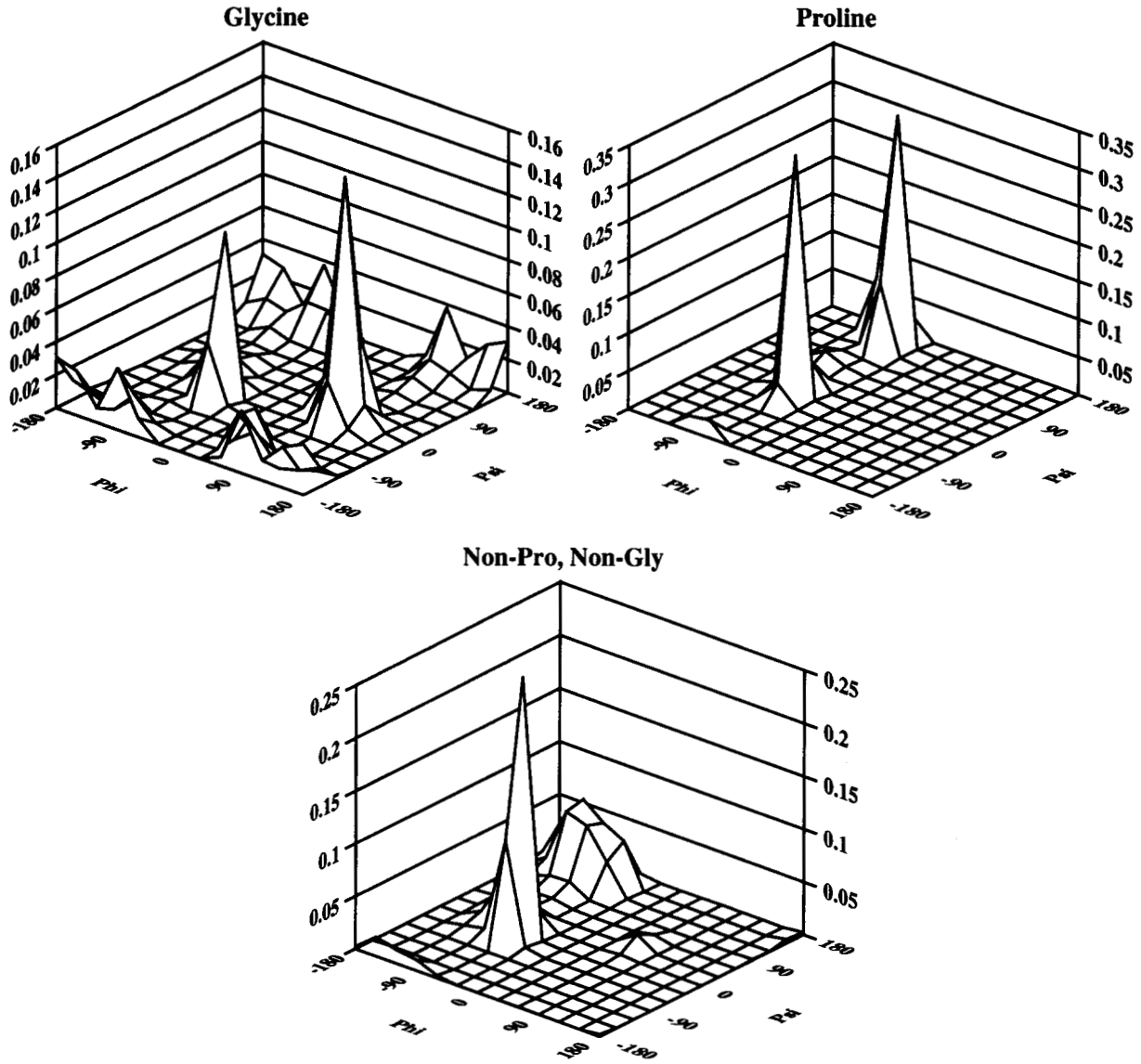

Fig. 6. The 30° (ϕ, ψ) grids for the three major residue types: glycine, proline, and standard (non-Pro, non-Gly). The height of the plot (vertical axis) at a particular (ϕ, ψ) is the normalized probability $P(\phi, \psi)$.

cine, which has no side chain, and tryptophan, which has 12 χ dihedral angles if you include those in the indole ring. Our simulations do not modify dihedral angles that affect only hydrogen positions (i.e., rotation of methyl groups), or those involved in rings, so the number of dihedrals is significantly reduced. Both alanine and glycine have zero DPG-MC side-chain dihedrals ($N_\chi = 0$), whereas tryptophan, tyrosine, phenylalanine, and histidine have only two, despite being very large side chains. The values of N_χ for the common amino acids, excluding alanine and glycine, are given in Table 10. Although proline is a ring, we allow χ^1 to vary while holding the C_δ atom fixed. This enables reasonable conformations of χ^1 to χ^4 to be sampled by modifying only a single dihedral, χ^1.

Table 10 also lists the number of occurrences of each amino acid in the H64 data set as well as the number of populated (nonzero) grid points and the maximum possible grid points at each spacing level. Many of the probability grids are sparse, with only a small fraction of the grid points populated. In most instances, this implies that a random search would sample many conformations never seen in nature. In some cases, however, it is clear that the number of populated grids is limited by the sample size rather than conformational propensities of the side chains. The multidimensional grid points ($N_d \geq 3$) at the finer spacings have nearly as many occupied grid points as the sample size (almost every conformation occupies a different grid point). This extreme variability is due primarily to the enormous number of possible conformations available for these structures (Table 10), rather than unusual flexibility in the individual torsions of these side chains. The (χ^1, χ^2) distributions of these residues makes this more clear (Mathiowetz, 1992): only lysine has an unusually large number of populated conformations (98 of 144 at 30°) when only χ^1 and χ^2 are considered. Arginine (60), methionine

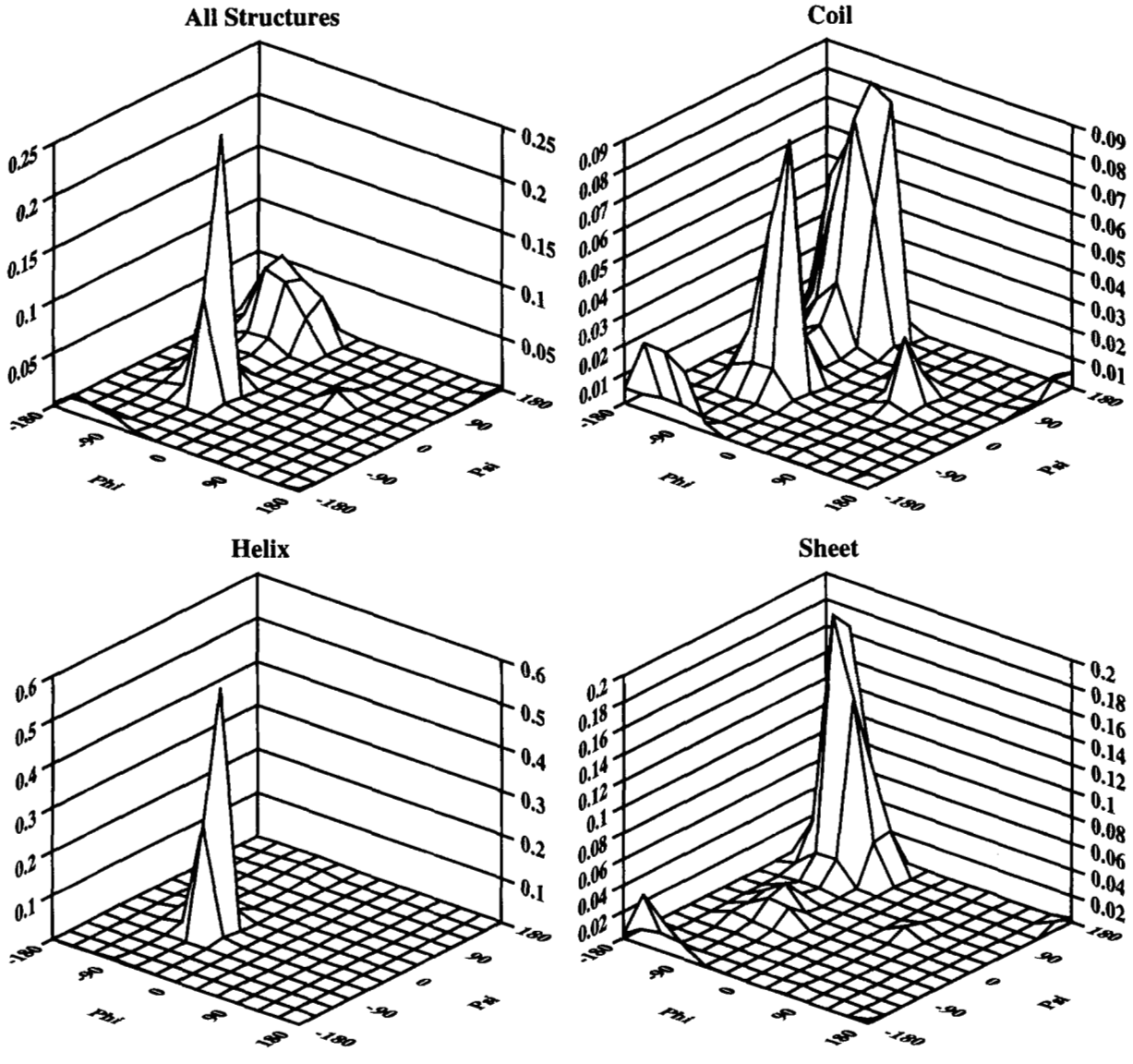

Fig. 7. The 30° (ϕ, ψ) grids of different structural classes for standard residues (non-Pro, non-Gly) in the SS58 data set. The vertical axis is $P(\phi, \psi)$.

(48), glutamine (67), and glutamic acid (86) have values typical of the smaller amino acids.

Several two-dimensional χ probability grids are shown in Figure 8. One-dimensional grids are simple probability versus dihedral graphs (Mathiowetz, 1992), whereas higher dimensional grids cannot easily be visualized. There is a great deal of variety even among residues with the same number of significant χ dihedrals. We should point out that some of the χ's actually have a periodicity of 180°, rather than the 360° as shown. This arises when two branches are the same (as in Asp, where the two carboxylate oxygens are chemically identical but only the one labeled O_δ is used to specify χ^2). This labeling in the PDB is not always done the same way, hence there are separate peaks at 150° and −30°. This does not affect our Monte Carlo simulations, because the orientations will be simulated identically, with a total probability equal to the sum of the individual probabilities. The great variety in side-chain conformations can also be seen in Table 11, which lists the highest probability side-chain conformation for each amino acid.

It is interesting to compare the values in Table 11 to the side-chain rotamers of Ponder and Richards (1987). The methods are not equivalent in that Ponder and Richards divide side-chain conformations into a small number of rotamers and then find the average χ values for each rotamer. Ponder and Richards also use a different set of proteins than are used here. Nevertheless, the results are very similar. If the Ponder and Richards rotamer χ values are rounded off to the nearest 30°, the most probable conformations for 14 of the 18 amino acids are the same as in our work (Table 11). The four that do not match are threonine, proline, glutamic acid, and histidine. The results for threonine are nearly the same because it has two conformations with nearly equal probabilities and the two methods simply reverse their

Table 10. *Number of populated (non-zero) grid points for various side chains at different grid spacings as tabulated from the H64 set of crystal structures*

Amino acid	N_χ	Sample	120°	60°	30°	15°	10°	5°
Cys	1	283	3	4	10	15	21	34
Pro	1	568	2	3	5	9	13	22
Ser	1	925	3	6	12	24	35	70
Thr	1	791	3	6	12	23	32	54
Val	1	991	3	6	12	23	30	51
Maximum	1		3	6	12	24	36	72
Asn	2	634	9	28	82	198	282	465
Asp	2	728	9	31	84	200	296	485
His	2	317	9	27	66	125	170	253
Ile	2	603	8	23	56	89	134	238
Leu	2	1,025	9	27	67	135	191	343
Phe	2	491	8	23	51	119	175	318
Trp	2	179	8	19	39	73	98	141
Tyr	2	453	9	22	52	107	172	294
Maximum	2		9	36	144	576	1,296	5,184
Glu	3	699	26	116	200	528	528	*688*
Gln	3	409	24	83	312	331	331	*404*
Met	3	241	20	54	120	185	218	*240*
Maximum	3		27	216	1,728	$>10^4$	$>10^4$	$>10^5$
Lys	4	858	67	288	580	775	*834*	*858*
Maximum	4		81	1,296	$>10^4$	$>10^5$	$>10^6$	$>10^7$
Arg	5	438	116	195	322	*421*	*429*	*436*
Maximum	5		243	7,776	$>10^5$	$>10^6$	$>10^7$	$>10^9$

[a] N_χ is the number of DPG dihedrals for each amino acid ($N_\chi = 0$ for alanine and glycine). "Sample" is the number of occurrences of each amino acid in the H64 set. Numbers in italics indicate cases where the number of occupied grid points is at least 95% of the sample size. "Maximum" refers to the maximum possible number of grid points for a given N_χ and grid spacing.

order. The other three cases differ primarily because Ponder and Richards calculate averages rather than strictly binning all conformations. Differences in the methodologies also lead to differences in the computed probabilities. In general, the probability of a particular gridpoint is lower than for the equivalent rotamer because conformational space has been divided into much smaller bins.

DPG protein builder: Backbone phase (DPG-BACK)

During the first stage of the Protein Builder, the backbone-modeling stage termed DPG-BACK, the protein is built one residue at a time until the entire protein has been built. As each residue l is added, its geometry is initially built from the standard peptide geometries in the BIOGRAF peptide library (Molecular Simulations, Inc., 1992), then the backbone (ϕ, ψ) and side-chain (χ) dihedrals are rotated to their most probable conformations according to the relevant probability grids. A Monte Carlo simulation using (ϕ, ψ) probability grids is then used to search the conformational space of a "pulse" of residues: the last p residues of the current chain (residues $(l-p+1)$ through l). The residues preceding the pulse are held fixed and are not included in the energy calculations. Simulations in which these early residues are held fixed, but included in the energy calculation, are considerably slower and give worse results. The side chains are also ignored during the chain-building phase; they are added in the second stage after the backbone conformation has been built. Long-range interactions are, therefore, not explicitly included in the backbone-building stage, despite their great importance in protein folding and packing interactions. Rather, they are implicitly included in that the C_α coordinates themselves represent the global fold. The backbone-building phase attempts to generate a polypeptide backbone that has an optimized local geometry *and* fits a particular global arrangement of C_α's. The energy used during the Monte Carlo simulations is essentially the DREIDING energy of the backbone atoms of the pulse, plus harmonic terms constraining the pulse C_α coordinates to the true coordinates. The best conformation sampled during the Monte Carlo simulation is saved and then optimized by conjugate gradients minimization. This process proceeds sequentially, with each new residue being involved in several optimization cycles before finally being held in its final position as the pulse moves beyond it.

The DPG-BACK simulations are aided by predetermination of the secondary structure where possible. There is a high correlation between the (ϕ, ψ) dihedrals of a protein and its C_α coordinates, so knowledge of the C_α coordinates can limit the possible (ϕ, ψ) values. The most common secondary structural elements, α-helices and β-sheets, have very specific C_α configurations, as described by the virtual angle ζ (defined by $C_\alpha(i-1)$, $C_\alpha(i)$, and $C_\alpha(i+1)$) and virtual dihedral γ (defined by $C_\alpha(i-1)$ through $C_\alpha(i+2)$). Analysis of the (ζ, γ) distributions of the proteins in the H64 data set showed that α-helix and β-sheet residues almost always have ζ and γ values

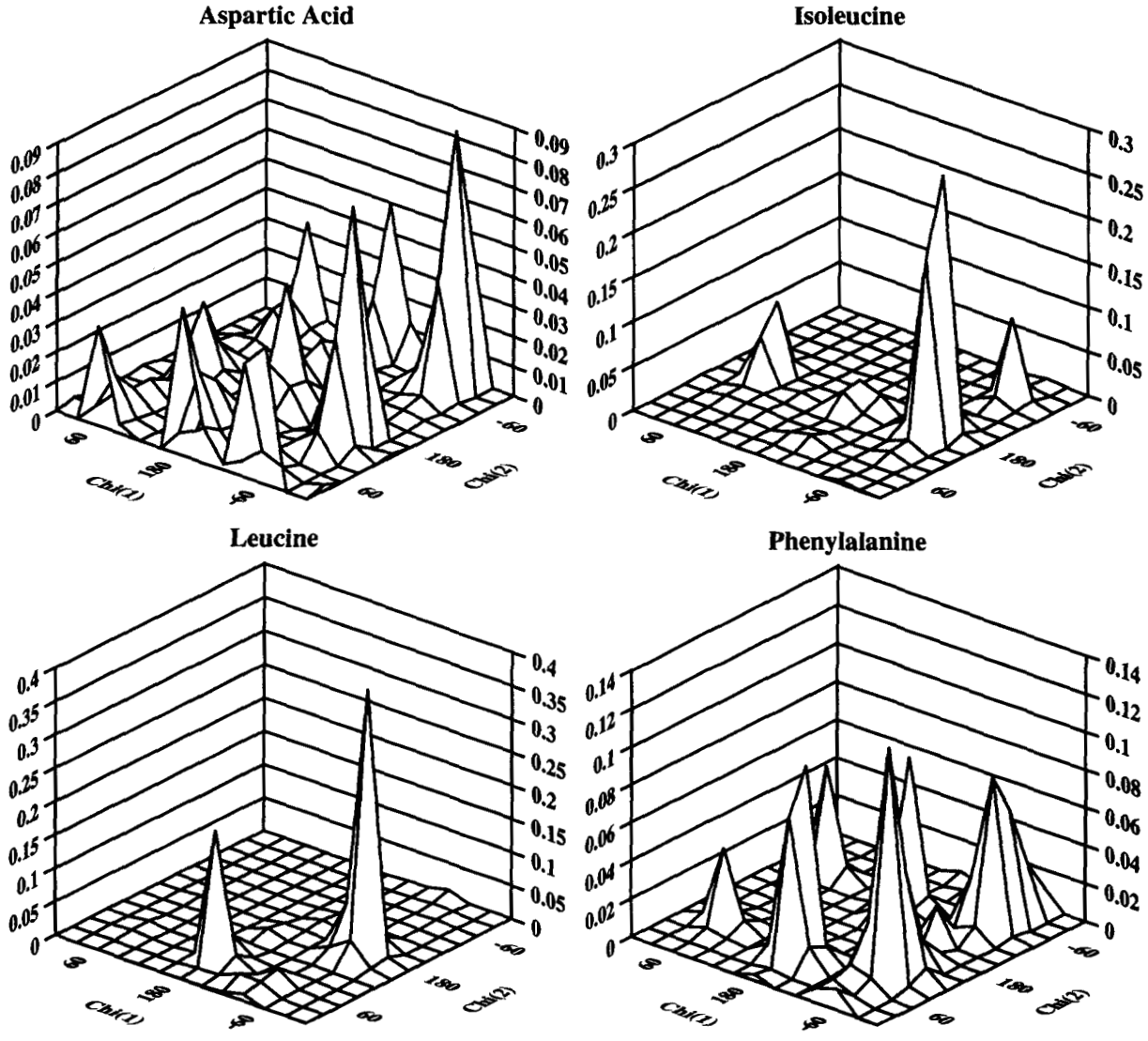

Fig. 8. The 30° χ grids for several different amino acids with two significant dihedrals. The vertical axis is $P(\chi^1, \chi^2)$.

within the ranges specified in Table 12. Residues with (ζ, γ) distributions in one of these two regions are assumed to have (ϕ, ψ) values common to that secondary structure class; when their ϕ and ψ conformations are sampled during the chain-building process, the (ϕ, ψ) grids determined for α-helix or β-sheet residues are used. Residues with (ζ, γ) values falling outside this region are sampled using the generic (ϕ, ψ) probability grids. Eighty-five percent of the residues in the H64 data set having ϕ and ψ values within the high-probability β-sheet region listed in Table 12 also have (ζ, γ) values within the specified region. The correlation is even higher for α-helices, where 88% of the residues with α-helix (ϕ, ψ) values have (ζ, γ) values within the corresponding range. If there were no variation in bond lengths and angles in the protein backbone, the (ζ, γ) angles would provide almost completely sufficient information to determine the (ϕ, ψ) angles, according to the method developed by Purisima and Scheraga (1984). Unfortunately, the variability in real conformations is too high for this exact method to work, and (ϕ, ψ) angles must be derived from simulation methods such as the one presented here. Nevertheless, the correlation between (ζ, γ) and (ϕ, ψ) angles is sufficient to determine which residues should be sampled using the α-helix and β-sheet (ϕ, ψ) grids. The use of these grids for the appropriate residues improves our results significantly.

Monte Carlo simulations depend on random numbers and produce a different backbone conformation each time the calculation is run. However, an exhaustive search is much more computationally intensive, even if only a few conformations were allowed for each residue. A complete sampling of just the top 20 (ϕ, ψ) conformations for each residue in a three-residue pulse would require evaluation of 8,000 different conformations. In contrast, we are able to obtain excellent results from only 200 Monte Carlo steps as each amino acid is added. The Metropolis criterion (Metropolis et al., 1953) rejects conformations producing very bad energies, allowing the conformational sampling to focus on low-energy conformations. It is therefore possible

Table 11. *Highest probability grid point for each amino acid for S = 30°* [a]

	P_{max} (%)	χ^1	χ^2	χ^3	χ^4	χ^5
Cys	49.1	−60				
Pro	38.2	0				
Ser	31.0	60				
Thr	42.4	−60				
Val	60.6	180				
Asn	9.5	−60	−30			
Asp	9.3	−60	−30			
His	13.2	−60	−90			
Ile	31.5	−60	180			
Leu	38.8	−60	180			
Phe	12.6	−60	90			
Trp	12.3	−60	90			
Tyr	13.9	−60	−90			
Glu	3.8	−60	180	150		
Gln	2.9	−60	180	−30		
Met	5.8	−60	−60	−60		
Lys	4.7	−60	180	180	180	
Arg	2.1	−60	180	180	−150	0

[a] P_{max} is the probability of this particular conformation.

Table 12. *Secondary structure correlations*[a]

	γ_i	ζ_i
α-Helix	$25° < \gamma_i < 75°$	$80° < \zeta_i < 110°$
β-Sheet	$160° < \gamma_i < -75°$	$100° < \zeta_i < 145°$
	ϕ_i	ψ_i
α-Helix	$-90° < \phi < -30°$	$-60° < \psi < 0°$
β-Sheet	$-165° < \phi < -45°$	$100° < \psi < 180°$

[a] The C_α angles, (ζ, γ), corresponding to α-helix or β-sheet conformation. Greater than 85% fall within the corresponding (ϕ, ψ) region listed in the lower table. ζ_i and γ_i are defined in the text.

to quickly build backbone conformations. A typical simulation takes approximately 15 s per residue on one processor of a Silicon Graphics 4D/380 workstation, or less than 12 min for the 46-residue protein crambin. Speed is crucial for simulations where different C_α conformations are being evaluated, for instance, when numerous conformations are generated by a lattice-based protein structure prediction method (Covell & Jernigan, 1990). In cases where a single set of C_α coordinates is being used, it may not be necessary to limit the calculations to a matter of minutes. In these cases, several simulations can be run, using different random numbers for the Monte Carlo calculation. Each will produce a slightly different backbone conformation. From these, the lowest energy conformations are selected for the second stage of the calculation.

DPG Protein Builder: Side-chain phase (DPG-SIDE)

The best-energy conformations generated in the DPG-BACK phase were evaluated *without* regard to side-chain positions. During the chain-building process, energies were determined for only a small pulse of residues; all previous residues were ignored. However, after the chain is built, the energy of the entire backbone is evaluated and this value is used to determine which backbone conformations are used in stage 2, DPG-SIDE. The side-chain conformations are optimized by a DPG-MC simulation using χ probability grids. In this DPG-SIDE stage, the backbone atoms are held fixed but are included in the energy calculation. Because the backbone is held fixed, constraints to the C_α coordinates are removed. In these calculations, at every Monte Carlo step, one side chain is selected at random and a new side-chain conformation is chosen for it according to the residue-specific χ probability grid. The energy of the new conformation is calculated, and the Metropolis criterion is used to accept or reject this structure. Because the Metropolis acceptance probability (Metropolis et al., 1953) is dependent upon ΔE, the change in energy, only the energy of the side chain being modified needs to be evaluated; all other interactions can be ignored. This results in a huge speed increase over calculations that re-evaluate the entire energy of the protein at every step. Using this method, the second stage can be quite rapid. For the small protein crambin, which has 46 residues and 396 atoms in the DREIDING calculations, 1,000 Monte Carlo steps require 7 min of cpu time, whereas plastocyanin, with 98 residues and 857 atoms, requires 22 min for 1,000 steps. Like the backbone-building process, the side-chain-modeling process is a stochastic simulation, dependent upon random numbers. Therefore, it is useful to run the simulation several times, using different random number seeds, and to use the lowest-energy structures for further studies.

Variables

There is a considerable number of variables that affect the efficiency of the DPG Protein Builder. The most important of these are:

1. *Spacing* (S). The DPG-MC grid spacing, as described above.
2. *Temperature*. The constant controlling the Monte Carlo acceptance rate.
3. *Steps*. The number of conformations sampled by the DPG-MC calculation. In the DPG-BACK phase, this refers to the number of backbone conformations sampled each time a residue is added to the growing chain. In the DPG-SIDE phase, it refers to the total number of conformations sampled.

There are two additional variables affecting the DPG-BACK phase:

1. *Pulse*. The number of residues used in the conformational sampling as each new residue is added.
2. *Harmonic constraint* (K_c). Force constant of the harmonic constraint between the C_α of the protein being built and the target C_α coordinate.

In order to determine which combination of parameters is most effective, we ran numerous simulations using crambin (Hendrickson & Teeter, 1981) as a model. This protein was cho-

sen because of its small size (allowing rapid calculations) and because it contained α-helix, β-sheet, and β-turn regions. Backbone phase parameters were evaluated by running 20 simulations for each set of parameters, building the complete crambin backbone from its C_α coordinates. The efficacy of the parameters was determined by averaging, over the 20 runs, the RMSDs from the crystal structure for the backbone atoms of the models produced. This average correlated very well with a second measure of the accuracy of the backbone model: the RMSDs in the (ϕ, ψ) dihedrals. Not every variable had a large impact on the results. In particular, the simulation temperature and the grid spacing had smaller effects than did the pulse size, the harmonic constraint, or the number of Monte Carlo steps.

The average RMSDs from 20 backbone phase simulations are shown in Figure 9 for several temperatures and pulse sizes. These simulations were run using 200 Monte Carlo steps for each pulse, a grid spacing of 10°, and a C_α constraint of 1,000 (kcal/mol)/Å^2. There are no consistent trends with respect to temperature. For pulse lengths of three or four, the best results are obtained at a temperature of 1,000 K. However, for longer pulses, higher temperatures are more favorable. The pulse length itself has a much bigger impact on the results. There is a consistent trend favoring shorter pulse lengths at all temperatures except 5,000 K, where a pulse of six is better than a pulse of five. It was clear from numerous other simulations that a pulse length of three gave the best results, with four residues being slightly worse and larger numbers significantly worse. The number of possible (ϕ, ψ) conformations grows exponentially with the number of residues in the pulse, so smaller pulse lengths are clearly favored in that a larger percentage of their conformational space can be searched during the Monte Carlo calculation. This makes up for the fact that important hydrogen bonding interactions occur between residues i and $i+4$ in α-helices, a fact that would favor a pulse length of at least four. In addition, the time of the simulation is roughly proportional to p, so a pulse length of three is preferable from the standpoint of speed, as well.

Another important variable in these simulations is the force constant of the harmonic constraint between the C_α's of the

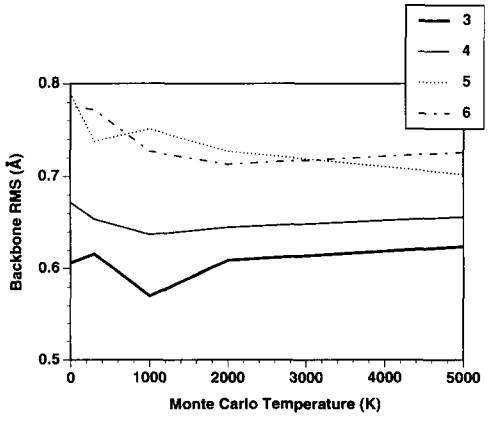

Fig. 9. Average RMSDs for the crambin backbone built using DPGBACK. Each calculation started with the C_α coordinates from the crystal structure and determined 20 independent structures (based on 200 MC steps) using particular temperatures and pulse sizes. The 10° (ϕ, ψ) grid was used with $K_{\text{constraint}}$ = 1,000 kcal/mol. The pulse size is 3, 4, 5, or 6 as indicated at the upper right.

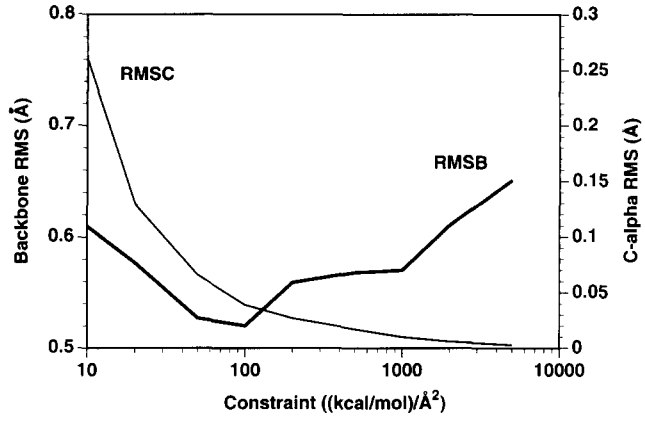

Fig. 10. Dependence of the RMS error in the backbone atoms (RMSB) and in the C_α coordinates (RMSC) built using different C_α constraint force constants. Based on 20 independent structures using T = 1,000 K and a pulse size of 3.

protein chain being built and the input C_α coordinates. The energy of each constraint is given by the expression

$$E_c = \tfrac{1}{2} K_c (r_i)^2, \tag{2}$$

where K_c is the force constant and r_i is the distance between the C_α coordinate of residue i in the model and in the template. There is a constraint of this type for each residue in the pulse. There is an additional constraint, with a weak force constant of $K_c/10$ and an offset of 2.0 Å, between the carbonyl carbon of the most recently added residue, l, and the template C_α of residue $l+1$. This helps to orient the final residue of the growing chain. Figure 10 shows the effect of the constraint on the average RMS errors in the backbone atoms (RMSB) and the C_α coordinates (RMSC). These simulations were run at a temperature of 1,000 K, using a grid spacing of 10° and a pulse length of three. As should be expected, deviations for the C_α coordinates decrease exponentially as the force constant increases. However, the fit of the entire backbone has a minimum of 0.520 Å when K_c = 100 (kcal/mol)/Å^2. This is substantially less than a typical DREIDING force constant of 700 (kcal/mol)/Å^2 or more for bond stretches. Therefore, the C_α constraints do not cause distortions in the geometries during the conjugate gradients minimization stage that follows the Monte Carlo.

As each new residue is added, the pulse of residues is optimized first by the Monte Carlo conformational search, then by 100 steps of conjugate gradients minimization. Both stages are important. The minimization process is necessary to provide flexibility in the bond lengths and angles of the protein model, in order to match closely the specific C_α geometry of the protein being built. Although the minimization process makes only small adjustments in the conformation of the pulse residues, it makes a substantial difference in the results. With no minimization, the errors in the backbone model built up very quickly. Using the same parameters that produced an average backbone deviation of 0.52 Å when minimization was included, the DPG Protein Builder produces crambin backbone models with an average RMSD of 1.32 Å when no minimization is involved. Our parameters were optimized for simulations including minimiza-

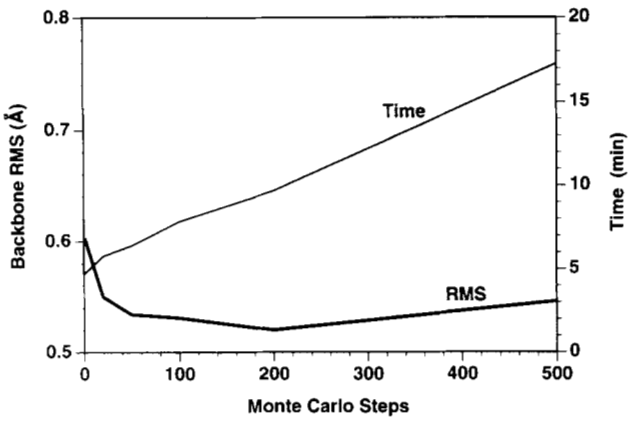

Fig. 11. Dependence in the backbone RMS error on the number of Monte Carlo steps. Based on 20 independent structures using $T = 300$ K and a pulse size of 3. The conclusion is that 50 steps is sufficient. We used 200 for additional calculations.

tion and probably are not optimal for simulations without minimization. Nevertheless, it is clearly preferable to include the minimization process. It is also important to include the Monte Carlo conformational search. The results using different numbers of Monte Carlo steps are shown in Figure 11. Simulations with one step correspond to simply using the highest probability conformation from the (ϕ, ψ) grids for each residue; no other conformations are sampled. Although the results for this case are good (0.60 Å RMS), the results are clearly improved by the use of even a small number of Monte Carlo steps, and get better as the number of steps increases. The standard error in these averages is typically 0.01 Å, so there is little statistical significance to the improvements above 50 steps. Nevertheless, in order to increase the number of conformations sampled while keeping the simulation time to 10 min per crambin backbone conformation, we chose to use a value of 200 Monte Carlo steps for most simulations.

The choice of grid spacing was based upon simulations of the pentapeptide Met^5-enkephalin (Mathiowetz, 1992), which showed that the best results are obtained using a grid spacing of 10°. The 10° dihedral spacing appears to provide the best balance between conflicting trends that arise as the grid spacing becomes smaller: there are far more possible conformations, so the protein can assume more low-energy conformations, but the fraction of the total conformational space that can be sampled with a given number of Monte Carlo steps decreases.

After DPG-BACK has determined the backbone models, the DPG-SIDE phase optimizes the side-chain χ's. In these calculations, the backbone is held fixed while the side chains are modified by randomly choosing new conformations according to the χ probability grids. The most important variables for these simulations are the grid spacing, the temperature, and the number of Monte Carlo steps. A grid spacing of 10° was selected for these calculations in order to be consistent with the grid spacing chosen for the DPG-BACK phase. Results improved consistently as the number of Monte Carlo steps was increased, but improvement slowed after about 500 steps; therefore, a value of 1,000 was used for the calculations reported. As discussed, this number may be insufficient for large proteins, but for crambin it represents more than 25 conformations per residue for the 37 non-alanine, non-glycine optimized during these simulations.

Table 13. *Values used for production runs of the DPG Protein Builder*

Variable	Phase 1	Phase 2
Spacing	10°	10°
Temperature	1,000 K	300 K
Steps	200	1,000
Constraint	100 (kcal/mol)/Å^2	–
Pulse	3	–

In order to determine the best simulation temperature for the DPG-SIDE phase, 10 DPG-SIDE calculations were run at 0 K, 300 K, 1,000 K, and 5,000 K. The starting structure for these calculations was the crambin crystal structure, with its side chains rotated to their most probable conformations according to the 10° χ probability grids. This structure had an RMSD from the crystal structure of 1.52 Å; the deviation for side-chain atoms alone was 2.34 Å. For each simulation, 1,000 Monte Carlo calculations were run, after which the lowest energy conformation was saved and its overall RMSD from the crystal structure was recorded. The averages for the 10 simulations at each temperature were 1.06 Å (0 K), 0.99 Å (300 K), 1.12 Å (1,000 K), and 1.08 Å (5,000 K). As was found for the DPG-BACK simulations (see Fig. 9), there is not a large variation with respect to temperature. This is the case despite the fact that the acceptance rate for new structures rises from 7.7% at 0 K to 46.8% at 5,000 K. Apparently, the much greater acceptance rate of new structures does translate directly into the creation of more low-energy conformations. Simulations at 300 K were more consistently accurate, so this temperature was used in the simulations reported below. Table 13 lists the values used for DPG-BACK and DPG-SIDE simulations reported in the Results.

Computations

DPG-MC and the DPG Protein Builder were developed as an extension of the BIOGRAF program from Molecular Simulations, Inc. (1992). All calculations reported here were run on Silicon Graphics Power Series and Indigo workstations; all timing numbers were obtained from simulations run on a single processor of an SGI 4D/380.

Supplementary material in the Electronic Appendix

Dihedral probability tables have been included in the Electronic Appendix (subdirectory Mathiwtz.SUP of the SUPLEMNT directory). The (ϕ, ψ) probabilities are in files named phipsi.*s.res.class*, where "s" is the grid spacing, "res" is the residue type (standard, glycine, or proline), and "class" is the structural class (all, coil, helix, sheet). The (χ) probabilities are found in files named chi.*s.res*, where "s" is the grid spacing and "res" is the amino acid type (aspartic acid, asparagine, etc.). Values are listed from the most probable conformation to the least probable.

Acknowledgments

A.M.M. acknowledges a National Research Service Award/NIH Predoctoral Biotechnology Traineeship. The research was funded by DOE-

AICD. The facilities of the MSC are also supported by grants from the NSF (ASC-9217368 and CHE91-100289), Allied Signal, Asahi Chemical, Asahi Glass, BP America, Chevron, B.F. Goodrich, Teijin Ltd., Vestar, Xerox, Hughes Research Laboratories, and Beckman Institute. Some of these calculations made use of the JPL Cray and the NSF Pittsburgh Supercomputer Center.

References

Abagyan R, Totrov M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol 235*:983–1002.

Brant DA, Miller WG, Flory PJ. 1967. Conformational energy estimates for statistically coiling polypeptide chains. *J Mol Biol 23*:47–65.

Bruccoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers 26*:137–168.

Collyer CA, Guss JM, Sugimura Y, Yoshizaki F, Freeman HC. 1985. Structure of oxidized poplar plastocyanin at 1.6 Å resolution. *J Mol Biol 211*: 617–632.

Correa PE. 1990. The building of protein structures from α-carbon coordinates. *Proteins Struct Funct Genet 7*:366–377.

Covell DG, Jernigan RL. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry 29*:3287–3294.

Friedrichs MS, Wolynes PG. 1989. Toward protein tertiary structure recognition by means of associative memory Hamiltonians. *Science 246*:371–373.

Hendrickson WA, Teeter MM. 1981. Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature 290*:107–113.

Holm L, Sander C. 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from a C_α trace. Application to model building and detection of co-ordinate errors. *J Mol Biol 218*:183–194.

Jones TA, Zou JY, Cowan SW, Kjeldgaard M. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A 47*:110–119.

Lambert MH, Scheraga HA. 1989. Pattern recognition in the prediction of protein structure. II. Chain conformation from a probability-directed search procedure. *J Comput Chem 10*:817–831.

Lipton M, Still WC. 1988. The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J Comput Chem 9*:343–355.

Mathiowetz AM. 1992. Dynamic and stochastic protein simulations: From peptides to viruses [thesis]. Pasadena: California Institute of Technology.

Mayo SL, Olafson BD, Goddard WA III. 1990. DREIDING: A generic force field for molecular simulations. *J Phys Chem 94*:8897–8909.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*: 1087–1092.

Molecular Simulations, Inc. 1992. *BIOGRAF/POLYGRAF*. Burlington, Massachusetts: Molecular Simulations, Inc.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444–2448.

Phillips SEV. 1980. Structure and refinement of oxymyoglobin at 1.6 Å resolution. *J Mol Biol 142*:531–554.

Plaxco KW, Mathiowetz AM, Goddard WA III. 1989. Predictions of structural elements for the binding of Hin recombinase with the hix site of DNA. *Proc Natl Acad Sci USA 86*:9841–9845.

Ponder JW, Richards FM. 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.

Purisima EO, Scheraga HA. 1984. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers 23*:1207–1224.

Ramachandran GN, Ramakrishnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol 7*:95–99.

Reid LS, Thornton JM. 1989. Rebuilding flavodoxin from Cα coordinates: A test study. *Proteins Struct Funct Genet 5*:170–182.

Rey A, Skolnick J. 1992. Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J Comput Chem 13*:443–456.

Smith WW, Burnett RM, Darling GD, Ludwig ML. 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 117*:195–225.

Wlodawer A, Svensson LA, Sjolin L, Gilliland GL. 1988. Structure of phosphate-free ribonuclease A refined at 1.26 Å. *Biochemistry 27*:2705–2717.

Wlodawer A, Walter J, Huber R, Sjolin 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.