Protein Science (1995), *4*:1203–1216. Cambridge University Press. Printed in the USA.

De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology

JOHN SPENCER EVANS,[1,3] ALAN M. MATHIOWETZ,[2,4] SUNNEY I. CHAN,[1] AND WILLIAM A. GODDARD III[2]

[1] Arthur Amos Noyes Laboratory for Chemical Physics (127-72), Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125
[2] Materials and Molecular Simulation Center, Beckman Institute (139-74), Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

(RECEIVED December 20, 1994; ACCEPTED March 21, 1995)

Abstract

We tested the dihedral probability grid Monte Carlo (DPG-MC) methodology to determine optimal conformations of polypeptides by applying it to predict the low energy ensemble for two peptides whose solution NMR structures are known: integrin receptor peptide (YGRGDSP, Type II β-turn) and S3 α-helical peptide (YMSEDEL KAAEAAFKRHGPT).

DPG-MC involves importance sampling, local random stepping in the vicinity of a current local minima, and Metropolis sampling criteria for acceptance or rejection of new structures. Internal coordinate values are based on side-chain-specific dihedral angle probability distributions (from analysis of high-resolution protein crystal structures). Important features of DPG-MC are: (1) Each DPG-MC step selects the torsion angles (ϕ, ψ, χ) from a discrete grid that are then applied directly to the structure. The torsion angle increments can be taken as $S = 60$, 30, 15, 10, or 5°, depending on the application. (2) DPG-MC utilizes a temperature-dependent probability function (P) in conjunction with Metropolis sampling to accept or reject new structures.

For each peptide, we found close agreement with the known structure for the low energy conformational ensemble located with DPG-MC. This suggests that DPG-MC will be useful for predicting conformations of other polypeptides.

Keywords: computational chemistry; importance sampling; Monte Carlo; peptide conformation; protein conformation; protein folding

A full understanding of protein function requires knowledge of the three-dimensional structure. Unfortunately, experimentally determined structures for most proteins are unavailable. Consequently, it is essential to develop approaches for predicting secondary and tertiary structures of proteins.

In the past decade, several approaches to protein structure prediction have evolved, including: (1) lattice search methods (Covell & Jernigan, 1990; Skolnick & Kolinski, 1990, 1991; Rey & Skolnick, 1991, 1993); (2) homology data search (Lupas et al., 1991; Rooman et al., 1992; Srinivasan et al., 1993; Geourjon & Deleage, 1994); (3) genetic algorithms (Judson et al., 1993; McGarrah & Judson, 1993); and (4) Monte Carlo methods (Paine & Scheraga, 1985; Lambert & Scheraga, 1989; Nayeem et al., 1991).

An exhaustive search of the conformational space for the protein is usually not computationally practical, because the size of the space grows exponentially with the size of the molecule (Ngo & Marks, 1992). Hence, many simplifications have evolved to the computational requirements. Examples include: (1) reduced numbers of atoms by the use of the C_α or virtual bond representation of the protein backbone (Purisima & Scheraga, 1984; Covell & Jernigan, 1990; Rey & Skolnick, 1991; Skolnick & Kolinski, 1990, 1991; Mathiowetz, 1992); (2) pairwise functions for side-chain–side-chain interactions, allowing rapid evaluation of structure energies (Miyazawa & Jernigan, 1985; Rey & Skolnick, 1991, 1993; Sippl et al., 1992; Kocher et al., 1994); and (3) the use of existing structures as "templates" for generating un-

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.

[3] Present address: Department of Chemistry, New York University, New York, New York 10010.

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

Abbreviations: DPG-MC, dihedral probability grid Monte Carlo; BP-MC, biased probability Monte Carlo; IRP, integrin receptor peptide; Q, partial atomic charge; vdW, van der Waals; PDB, Brookhaven Protein Data Bank; ϵ_0, dielectric constant; S, grid spacing; T_{MC}, Monte Carlo temperature; E_Q, Coulombic potential; Q_{net}, net molecular charge; r18, the non-Pro, non-Gly amino acids; RMSD, RMS deviation.

known protein structures (Rooman et al., 1992; Sippl et al., 1992; Madej & Mossing, 1993; Srinivasan et al., 1993). The use of existing protein structural data to guide the selection of structures is attractive for the following reasons. First, there exists a large data set of protein crystal structures (~800) encompassing a wide range of polypeptide backbone and side-chain arrangements. Hence, this data set contains implicit information regarding secondary and tertiary structure, sterics, and folding. Second, the data set provides a range of backbone and side-chain dihedral angles for each amino acid type. This information can be used to reduce computational effort by excluding regions of the energy map from the conformational search, thus permitting a more focused search in biologically relevant conformer regions.

As a means of integrating protein database knowledge with conformational search methodologies, we have developed an internal coordinate search algorithm that is guided by residue-specific dihedral angle probabilities as determined from known protein crystal structures (from the Brookhaven Protein Data Bank). This dihedral probability grid Monte Carlo method (Mathiowetz, 1992) involves the use of (1) importance sampling (Lambert & Scheraga, 1989), (2) local step procedures (i.e., random steps occur in the vicinity of a current local minima), and (3) Metropolis sampling for acceptance or rejection of new structures (Metropolis et al., 1953).

Some important features of DPG-MC are: (1) Each step of the simulation involves the selection of torsion angles (ϕ, ψ, χ) from the DPG, which are then applied directly to the structure. The torsion angle resolution can be taken as $S = 60$, 30, 15, 10, or 5°, depending on the application. (2) A temperature-dependent probability function, P, is used in conjunction with Metropolis sampling (to accept or reject new structures whose energies are *less* favorable than the starting structure). This ensures a Boltzmann distribution of conformations. The number of accepted structures increases with the Monte Carlo temperature. This creates a larger number of "bad" structures, but it also provides alternative "paths" that may eventually lead to lower energy minima because lower acceptance rates require that the simulation take a more "downhill" path through conformational space (Mathiowetz, 1992). (3) Cartesian coordinate minimization steps can follow the torsional perturbation. (4) Any force field can be utilized for energy evaluations. (5) The DPG-MC backbone and side-chain grid probabilities represents a diverse, highly refined subset of PDB protein crystal structures (64 structures, with resolution ≤2.0 Å and R-factor <20%) whose sequence overlap is less than 25%.

Herein we test the DPG-MC method using as benchmarks Met^5-enkephalin; the 7-residue integrin receptor peptide, YGRGDSP, which assumes a β-turn structure (Johnson et al., 1993); and the S3 peptide, YMSEDELKAAEAAFKRHGPT, which forms an α-helix structure (Lyu et al., 1993). The latter two peptides have known NMR solution structures. We find that DPG-MC gives excellent results for predicting structure and that the protein dihedral probability grids are effective for selecting minima.

Results

Using Met^5-enkephalin as the test case or benchmark, we first consider how efficiently DPG-MC finds conformational minima. We consider: (1) the "best" structure energy as a function of the Monte Carlo step size; (2) the conformer energy fluctuations over the course of the search; (3) the effect of S and T_{MC} on conformer selection; (4) the effect of dihedral probabilities on selection of minima; and (5) the use of Cartesian minimization following dihedral torque but prior to Metropolis evaluation.

Conformer minima selection as a function of grid spacing, S

We evaluated the effect of dihedral angle grid point size on minima selection at room temperature (300 K) (see Fig. 1). In the first third of each simulation (<60,000 steps), it is apparent that smaller grid spacings ($S = 5$, 10, 30°) lead to lower energy minima, with the $S = 5°$ grid providing the best approach to the minimum structure. This result is expected because small changes in the torsion angles offer a larger number of pathways on the energy map. The large, stepwise changes of the $S = 60°$ simulation may be useful for large atom assemblies, where an initial coarse grid search can locate rapidly the minima to be refined further by the use of denser grids. As the simulation proceeds (>100,000 steps), we observe energy convergence for all grids except 60°. As shown in Table 1, the acceptance rate for a given T_{MC} is not greatly affected by the choice of S. We conclude the following regarding grid spacing: (1) Small S grid values (<30°) are preferable for determining lowest energy structures, particularly in short MC simulations. (2) Larger gridpoint values (60°) may be useful for initial coarse searches of conformational space.

Monte Carlo temperature and best energy

Examination of the "best" conformer energies for a sampling of T_{MC} at the optimum S value (5° grids) (Fig. 2A) reveals rapid energy convergence in the first 10,000 steps of the simulation (12–15 kcal/mol for 0, 300 K; 8–10 kcal/mol for 1,000–10,000 K). As the simulation proceeds out to 200,000 steps, smaller conformer energy changes are observed. Low temperature simulations generally lead to lower energy conformations. The conformer structures generated at each step of the simulations (Fig. 2B) vary greatly in energy, depending on T_{MC}.

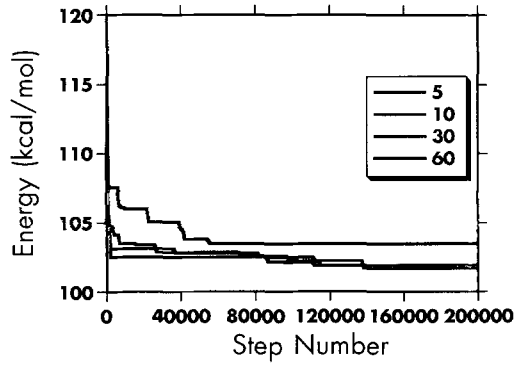


Fig. 1. Effect of grid spacing (S) on the final energy of Met^5-enkephalin. Energy of the "best" structure is given (every 10^3 steps) as a function of the Monte Carlo step, number using T_{MC} = 300 K. The DREIDING force field was used.

Table 1. *Comparative analysis of Met[5]-enkephalin DPG-MC simulations*[a]

S (deg)	T_{MC} (K)	Minimization cycles[b]	Starting structure	Accepted[c] (%)	Best energy[d] (kcal/mol)
5	0	0	Ext	0.03	100.6 (88.7)
	300	0	Ext	13.5	101.9 (82.9)
	1,000	0	Ext	46.2	104.2 (89.5)
	5,000	0	Ext	76.3	107.9 (91.0)
	10,000	0	Ext	83.6	108.5 (75.3)
10	0	0	Ext	0.02	103.2 (78.7)
	300	0	Ext	13.8	101.7 (93.8)
	1,000	0	Ext	46.1	105.8 (88.4)
	5,000	0	Ext	77.4	108.3 (90.1)
	10,000	0	Ext	83.5	108.3 (91.9)
30	0	0	Ext	0.01	103.1 (70.6)
	300	0	Ext	15.2	101.9 (86.2)
	1,000	0	Ext	44.3	104.3 (79.2)
	5,000	0	Ext	76.7	107.9 (77.0)
	10,000	0	Ext	83.0	107.9 (91.0)
60	0	0	Ext	0.01	107.7 (99.0)
	300	0	Ext	18.9	103.5 (87.5)
	1,000	0	Ext	44.5	103.6 (84.8)
	5,000	0	Ext	73.0	107.7 (89.9)
	10,000	0	Ext	79.4	107.9 (100.6)
5	300	1	Ext	13.2	100.9 (78.8)
	300	10	Ext	13.6	97.0 (82.8)
	500	0	β	–	107.5 (84.7)
		0	α	–	108.1 (81.2)
30[e]	0	0	Ext	0.01	110.9 (87.9)
	300	0	Ext	2.0	109.0 (94.7)
	1,000	0	Ext	20.1	109.0 (94.1)
	5,000	0	Ext	50.5	109.1 (94.2)
	10,000	0	Ext	607	109.1 (94.2)

[a] All DPG-MC simulations utilized 200,000 total steps. Unless otherwise noted, five runs were performed for each category.

[b] The number of minimization steps prior to Monte Carlo.

[c] "Accepted" refers to the percentage of 200,000 structures leading to low energy minima. Values represent the average for five parallel runs.

[d] Energy of the global minima, determined as the "best" structure generated from five parallel runs under the given conditions. For reference, the extended starting structure of Met[5]-enkephalin has an energy of 118.73 kcal/mol. The value in parentheses is the energy of the fully minimized best structure, determined as described in the text.

[e] Simulations utilized $T_{MC} = 0$, 300, 1,000, and 5,000 K, for $S = 5°$. Best energies represent the average energy of the best structures generated by each of the four runs.

Higher temperature simulations lead to greater fluctuations in energy and a wide variety of conformations, many of which represent "moves" away from the minima. The random acceptance of unfavorable minima during high temperature DPG-MC simulations sometimes led to lower energy final states (see Table 1 and Fig. 2A). Thus, at earlier stages of the search (20,000 steps) the 10,000 K run has reached a lower energy state compared to the 5,000 K run.

Database probabilities lead to lower energy minima structures

To determine how well the PDB H64 database probabilities lead to selecting lower energy conformers, we ran parallel DPG-MC simulations in which all grid points of a given spacing S (e.g., 30°) were assigned equal probabilities (Fig. 3). This situation corresponds to a "standard" Monte Carlo Metropolis sampling algorithm, where torsion angles are given random, discrete values. Compared to simulations using equal probabilities, the use of DPG clearly leads to selecting lower energy minima structures (3–7 kcal/mol difference). The acceptance rate for equal probability simulations is lower than that obtained for DPG simulations (Table 1). A comparison of structural energy changes during the course of each simulation reveals the following: (1) Compared to the DPG probability simulations, little if any improvement in structural energy occurs after the start of the equal probability simulation (Fig. 3). (2) The final minimized structures obtained for equal probability simulations are higher in energy than for DPG, leading to an ensemble whose energies are clustered near a single value (Table 1).

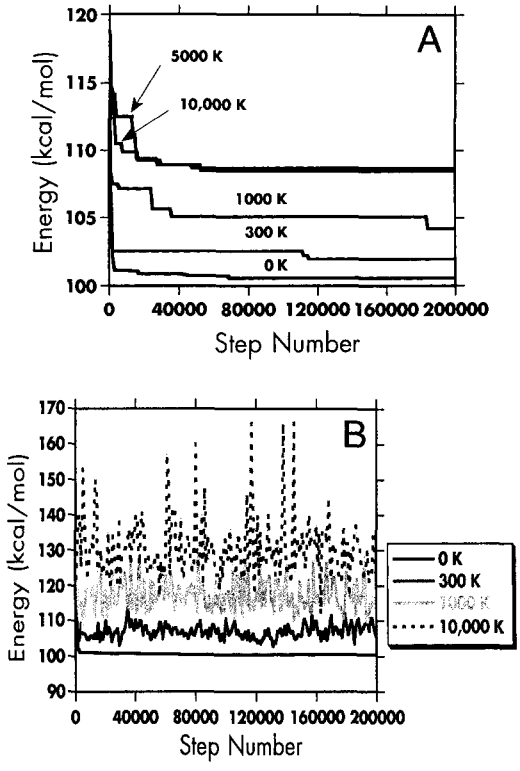


Fig. 2. Energy profile for DPG-MC Met[5]-enkephalin simulations as a function of the T_{MC} for $S = 5°$. **A:** "Best" structure energy/every 10^3 steps as a function of the Monte Carlo step number. **B:** Energy of each conformer generated (every 10^3 steps) of the Monte Carlo. The DREIDING force field was used.

These findings indicate that equal probability grid Monte Carlo simulations are more prone to minima "trapping" in local minima regions of the potential energy map and are less effective in exploring conformational space.

To ascertain if the DPG-MC program exhibits any bias toward starting structure conformation, we ran parallel simulations in which the starting structure of Met[5]-enkephalin was set to either all α-helix or all β-strand conformations. As shown in Table 1, the best structures obtained for the all-extended, α-helix, or β-strand starting structures show negligible energy differences, indicating that the starting conformation has little effect on the outcome of the conformational search.

DPG-MC with Cartesian minimization

To determine the effectiveness of the Monte Carlo with minimization approach, we performed Cartesian minimization (1 or 10 steps of conjugate-gradient minimization) for each "new" structure prior to Metropolis sampling. At a given T_{MC} and S value, it is evident from Figure 4 that a minimization step in the DPG-MC simulation improves minima selection. This was also noted in earlier Monte Carlo polypeptide studies (Nayeem et al., 1991). The use of a single minimization step leads to a 0.45-kcal/mol difference in "best" structure energy, whereas the 10-step minimization resulted in a 4-kcal/mol difference.

We did not pursue DPG-MC runs using a larger number of minimization steps because we found that the increased computational cost for DPG-MC with minimization was not justified. For a typical Met[5]-enkephalin 200,000-step DPG-MC run (with no cutoff), a single step of minimization increases the run time by a factor of 0.9, whereas a 10-step minimization increases it by a factor of 8.8. Table 1 compares the results of DPG-MC runs featuring 0, 1, and 10 steps of minimization in terms of acceptance rate, the number of best structures, and, most importantly, the structural energy obtained after the best conformer of each run was minimized to convergence in Cartesian space. We observed no significant difference in terms of acceptance rate. However, comparing the final minimized structures obtained from all DPG-MC runs, we found that the lowest energy conformers were generated by DPG-MC runs that featured 0 steps of minimization (Table 1: 5, 30° simulations). Inclusion of 1 or 10 minimization steps within the DPG-MC program does lead to an overall lower energy state for the simulation. However, such minimizations do not improve the probability of finding the global minima and require 2–10 times the computational effort.

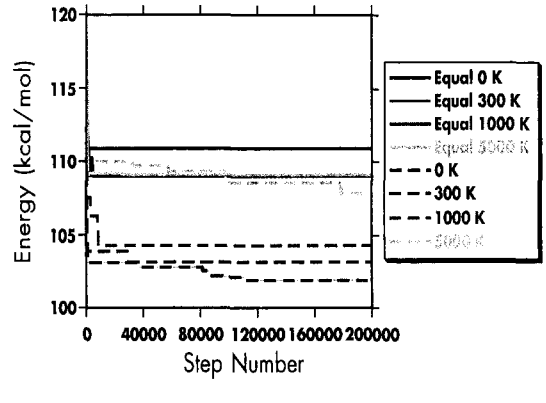


Fig. 3. Energy profile for DPG-MC Met[5]-enkephalin simulations utilizing biased or equal probability grid values. DPG-MC simulations were run utilizing T_{MC} values of 0, 300, 1,000, and 5,000 K. $S = 30°$ for all runs. The plot represents the "best" structures obtained every 10^3 steps. The term "Equal" denotes the equal probability runs. The DREIDING force field was used.

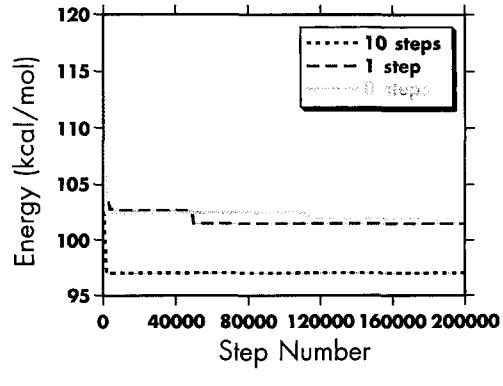


Fig. 4. Effect of Cartesian minimization (conjugate gradients) on the energy profile for the DPG-MC Met[5]-enkephalin simulations. DPG-MC simulations were conducted using $S = 5°$, at $T = 500$ K. Shown are the "best" structures obtained every 10^3 steps. The DREIDING force field was used.

Minimum energy structures

The most important test of DPG-MC simulations is the ability to accurately predict three-dimensional structure of polypeptides. Met^5-enkephalin is a suitable polypeptide for energy determinations, but we found that it was unacceptable as a structural benchmark for the following reasons: (1) Although Met^5-enkephalin has become a popular benchmark for Monte Carlo studies (Paine & Scheraga, 1985; Nayeem et al., 1991; Shin & Jhon, 1991), the use of different force fields in these studies makes inappropriate the comparison of efficacy on the basis of energetics. (2) It is now known that small peptides, such as Met^5-enkephalin, can populate a wide range of conformational states in solution (Dyson et al., 1988; Wright et al., 1990; Merutka et al., 1993). This distribution of low-lying conformational states make determination of the global minima difficult and perhaps meaningless.

For these reasons, we focused our attention on two peptides exhibiting strong conformational preferences. IRP or YGRGDSP has seven residues and exhibits a preference for a Type II β-turn structure in solution (Johnson et al., 1993). Peptide S3 or YM SEDELKAAEAAFKRHGPT is a 20-residue polypeptide whose solution structure is primarily α-helix in the first 15 residues of the sequence (50% helicity as determined by CD spectrometry and NMR) (Lyu et al., 1993).

The use of two different secondary structure "benchmarks" (helix and turn) provides a more rigorous test of global minima determination based both on energy and on the distribution of backbone and side-chain dihedrals. In each case, we know the preferred conformer state in solution and can unambiguously compare theoretical results with known experimental structures. In addition, we can also examine the range of conformational preferences or isoenergetic states for each peptide as determined by DPG-MC.

Integrin receptor peptide

For IRP, the dihedral distributions and structures for the four best minima structures (out of 2.4 million total structures) are presented in Figures 5 and 6. For the four minima structures, we obtained a mean total DREIDING energy of 65.5 ± 1.2 kcal/mol. As shown in Figure 6, the lowest energy IRP conformers form a turn structure exhibiting excellent overlap, with an overall RMS deviation (RMSD) among conformers of 1.36 Å for backbone and side-chain atoms (omitting H atoms). The greatest structural divergence is found at the termini (Fig. 6), where structural stability tends to be weakest (Wright et al., 1990). We did not observe β-turn-specific G_2 carbonyl-D_5 amide hydrogen bonding in any of the conformers, as was reported for YGRGDSP in solution (Johnson et al., 1993). Figure 5 shows that the dihedral preferences for IRP are primarily centered in three regions of the Ramachandran map: (1) near $\psi = -70°$, $\phi = -65°$; (2) near $\psi = 90°$, $\phi = 5°$; and (3) in a region bounded from $\phi = 80$ to 180° and $\psi = -50$ to $-120°$.

For β-turn structures (I–VIII), typical ϕ and ψ values range from +90° to −120° and +30° to −120°, respectively (Wilmot & Thornton, 1990). The majority of the IRP dihedrals (60%, $\psi = -70°$, $\phi = -65°$) exhibit good agreement with the ϕ,ψ distribution for the Type I β-turn and Type I β-turn distortions (Wilmot & Thornton, 1990). A smaller percentage (30%) of the IRP dihedrals ($\phi = 80$ to 180° and $\psi = -50$ to $-120°$; $\psi = 90°$, $\phi = 5°$) fall into the Type II β-turn category. Hence, DPG-MC predicts the lowest energy ensemble for IRP to be a β-turn with both Type I (predominant) and Type II β-turn conformers. This is in qualitative agreement with the results from NMR spectroscopy (see above).

S3 peptide

For S3, the four lowest energy conformational states (of 8 million total structures) (Figs. 5, 7) show two major secondary structure groupings: α-helix (representing 55–60% of the total number of dihedrals) and β-strand (representing 40–50%). For the four minimum energy structures, we obtained a mean total energy of 158.4 ± 7.3 kcal/mol (DREIDING), indicating a greater variation in conformer energy for this ensemble than obtained for IRP. This is further confirmed in Figure 7, where the four minima structures exhibit good overlap, with an average RMSD of 4.38 Å (backbone atoms only) and 5.91 Å (backbone and side-chain atoms). We attribute this wider range of conformers to the following: (1) The presence of several long side chains (Lys, Glu, Arg) in S3 allows for adopting of a variety of χ torsions, leading to a larger number of distinct low energy conformations. (2) These same side chains are charged, leading to

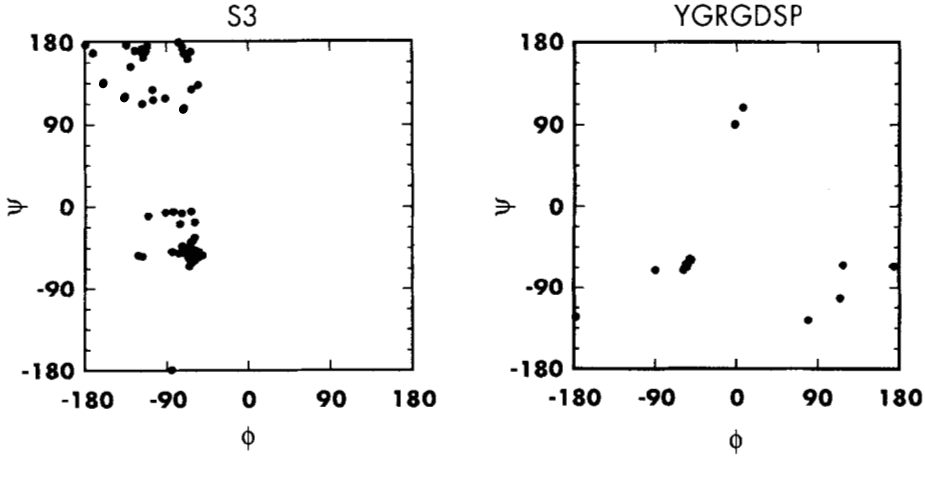


Fig. 5. Ramachandran (ϕ, ψ) plots for the low energy ensembles of IRP and S3 polypeptides. Scatter plots represent the four lowest energy conformers for each polypeptide as determined by DPG-MC simulations.

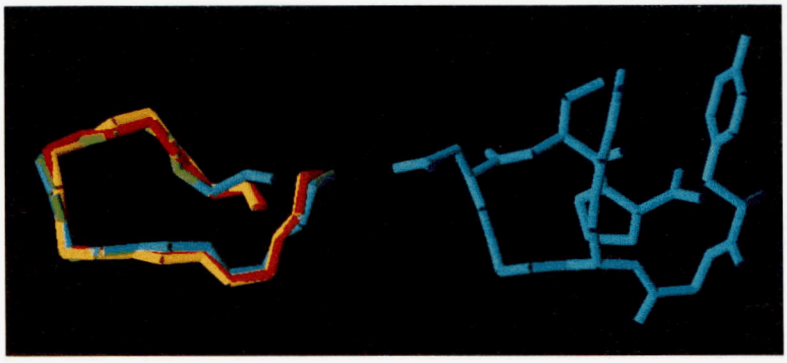


Fig. 6. All-atom representation of the ITP optimum structure. Views of the four superimposed lowest energy conformer states shown in Figure 5. All-atom backbone-only structures are shown on the left, with the global minimum structure (blue) represented in its entirety on the right. Hydrogen atoms are removed for clarity. Note the turn-like structure.

attractive–repulsive interactions that may lead to several distinct low energy conformations.

The minimum energy structures adopt a right-handed helical conformation over 55–60% of the overall length, beginning at the N-terminus and progressing to the midpoint of the peptide sequence. The lowest energy conformation (denoted in red in Fig. 7) features α-helical secondary structure from residues 1 through 14. The remaining structure (residues 15–20) is predominantly β-strand, with only 1–2 residues exhibiting any helical dihedral preferences. Experimentally, NMR NOESY studies have revealed that the S3 peptide has α-helical structure from residues 2–13 and 15–17, with the remaining portion of the peptide being undefined (Lyu et al., 1993). Our theoretical structure quantitatively mimics these features. Globally, the peptide conformation appears to adopt a "bent" or hairpin fold (Fig. 7).

Discussion

The results of this study suggest that the DPG-MC de novo protein conformational search algorithm provides improved capabilities for determining the low energy conformational states of polypeptides. In DPG-MC the trial conformations are selected on the basis of most probable internal coordinate values. This directs the local step moves toward energy minima, which are more biologically relevant (i.e., sterics, side-chain folding), increasing the efficiency of the search (e.g., see Fig. 4). With an appropriate selection of T_{MC}'s (high in the beginning and decreasing with time), the Metropolis sampling provides alternative paths that are not restricted to "downhill" searches, but periodically can "accept" unfavorable structures serving as alternative starting points. This avoids getting trapped into "local" minima (Fig. 4; Table 1).

With DPG-MC we are not limited to any particular force field for energy evaluation. Thus, depending on the situation, different energy potentials can be used to evaluate conformer "acceptance," e.g., AMBER or CHARMM for all-atom simulations of proteins, DREIDING (Mayo et al., 1990) for organics or modified proteins, and the universal force field, UFF (Rappè et al., 1992), for simulations involving tethered metal atoms.

In addition, one can include various approximations for including solvent electrostatics, e.g., POLARIS (Lee et al., 1993) for structure evaluation based on solvent electrostatics. Collectively, these features should allow DPG-MC to be a useful tool for de novo protein structure prediction.

Optimal DPG-MC parameters

The choice of S and T_{MC} can influence the outcome of the conformational search.

Grid coarseness, S

We considered several choices for S, the grid spacing for ψ, ϕ, and χ. From the Met^5-enkephalin simulations, it is clear that smaller S values ($\leq 30°$) lead to improved convergence. The 60° grid spacing is useful for larger structures, allowing a better conformation sampling at this "coarse" level that can then be refined using the smaller S grid spacings.

Monte Carlo temperature

A second consideration is the sequence of T_{MC}'s. The initial stages required high T_{MC} to better sample conformational space. This generates a larger percentage of "bad" structures (Fig. 2B; Table 1) but also causes the search to "jump" to new

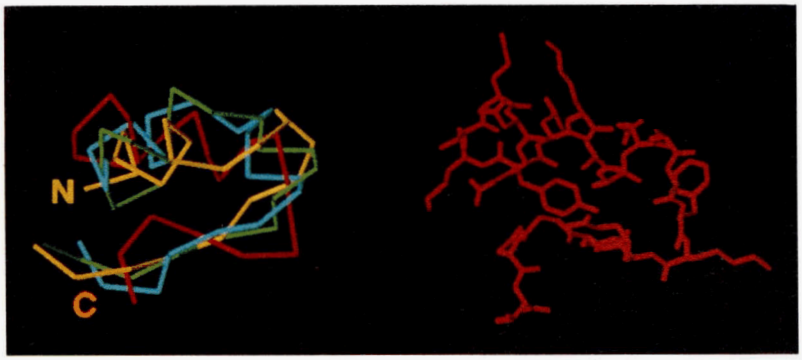


Fig. 7. C_α and all-atom representation of optimum S3 peptide. Views of the four superimposed lowest energy conformer states (C_α traces) shown in Figure 5. The global minimum structure is shown in red and represented in its entirety on the right. Note the helical backbone structure in the N-terminus of half of the polypeptide molecule. The C-terminal half of the polypeptide adopts a β-strand-like structure. Hydrogen atoms have been removed for clarity.

energy valleys (Fig. 2A, 5,000 K simulation; Table 1, note lower energy states obtained at $T > 300$ K). This is followed by incremental decreases in the T_{MC} as the conformations settle successively into better energy valleys. This "annealing" approach (Kawai et al., 1989; Nayeem et al., 1991; Shin & Jhon, 1991) initially generates a larger variety of structures. As T_{MC} drops, the best S should also drop to keep the acceptance rate high. This focuses the conformational search to the better energy regions.

Energy evaluation and search efficiency

As discussed previously, 1–10 cycles of Cartesian minimization prior to applying the Metropolis criterion lead to lower energy minima (Fig. 3). To reduce the computational expense, this can be done with short nonbond cutoffs during the initial stages (high T_{MC}, large S) with increased cutoffs for latter parts of the search. The accuracy can be increased by applying a smoothing procedure such as the optimum spline switch (Ding et al., 1992). This approach considerably reduces the runtime (J.S. Evans, S.I. Chan, & W.A. Goddard III, manuscript in prep.). In the latter stages, larger nonbond cutoffs lead to increased structural accuracy (Schreiber & Steinhauser, 1992).

PDB database: Limitations and expectations

Limitations inherent within any database method are: (1) the size and comprehensiveness of the database; (2) the population and variety of the secondary structures; and (3) biases in the protein folding patterns within this sampling.

We have opted for a small, highly refined, minimally overlapping protein crystal structure subset as the basis for the dihedral probability grids. As shown in the dihedral distributions (Fig. 8; Table 2), the (ϕ, ψ) distribution from the r18 data set exhibits the following biases: (1) globular protein structures predominate; (2) certain polypeptide secondary structures (α-helix, β-strand) are highly probable; and (3) smaller S values (5°, 10°) have non-zero gridpoints covering a smaller percentage of the possible conformations.

With regard to secondary structure, the r18 data set is poorly represented in quadrants III and IV of the Ramachandran map. Therefore, the database may accurately generate helical and β-strand segments de novo but may fail to generate with the same accuracy all β-turn structures (Wilmot & Thornton, 1990), as well as other "nonstandard" secondary structures (Matsushima et al., 1990; Gerstein & Chothia, 1991). This bias will be reduced as additional highly refined structures are added to the database.

These studies of low energy conformational ensembles for the IRP and S3 peptides (Figs. 5, 6, 7) reveal that DPG-MC can accurately predict the predominant secondary structural preference (i.e., global minima) for polypeptides. Thus, this algorithm accounts for the residue-specific helical and nonhelical portions of the S3 peptide, as determined from NMR data (Lyu et al., 1993). More importantly, DPG-MC provides a means of determining the conformational ensemble of low energy structures having internal energies close to the global minima. Solution-state NMR spectroscopy has shown that some peptides are conformationally flexible in solution, with significant populations of a number of different conformer states (Dyson et al., 1988; Wright et al., 1990; Merutka et al., 1993). The existence of the conformational ensemble is significant, particularly in terms of antibody–peptide antigen structure–function relationships (Dyson et al., 1988) and protein folding processes (Wright et al., 1990; Merutka et al., 1993). One of the most important factors

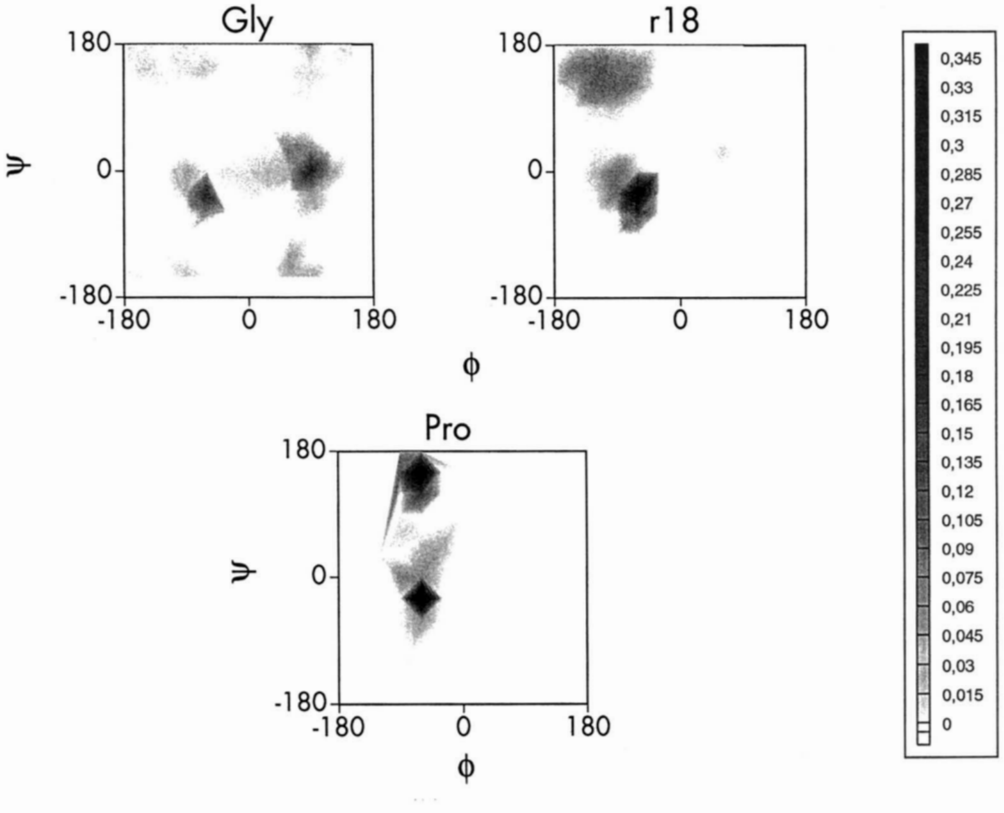


Fig. 8. The $S = 30°$ probability grids, (ϕ, ψ), for the H64 PDB data set. Higher probability regions denoted by increasing gray scale values; unpopulated regions are indicated in white. Probabilities are given in the legend.

Table 2. *Statistical analysis of (ϕ, ψ) grid points for the H64 database*

Grid spacing, S (deg)	Number of grid points, N_S	Measure[a]	Gly	Pro	r18
60	36	Non-zero	30 (83.3%)	10 (27.8%)	33 (91.7%)
		$P > P_{avg}$	12 (33.3%)	4 (11.1%)	7 (19.4%)
30	144	Non-zero	83 (57.6%)	24 (27.8%)	110 (76.4%)
		$P > P_{avg}$	33 (22.2%)	11 (7.6%)	23 (16.0%)
15	576	Non-zero	198 (34.4%)	48 (8.3%)	253 (43.9%)
		$P > P_{avg}$	127 (22.0%)	48 (8.3%)	78 (13.5%)
10	1,296	Non-zero	312 (24.1%)	76 (5.9%)	429 (33.1%)
		$P > P_{avg}$	312 (24.1%)	76 (5.9%)	172 (13.3%)
5	5,184	Non-zero	593 (11.4%)	164 (3.2%)	1114 (21.5%)
		$P > P_{avg}$	593 (11.4%)	164 (3.2%)	630 (12.2%)

	Boundaries				
Quadrant[b]	ϕ	ψ	Gly	Pro	r18
I	Negative	Positive	14.8%	54.4%	47.8%
II	Negative	Negative	29.9%	45.6%	49.4%
III	Postive	Postive	29.1%	0.0%	2.4%
IV	Positive	Negative	26.1%	0.0%	0.4%

[a] Non-zero, number of non-zero grid points (% of total number of grid points); $P > P_{avg}$, number of grid points in high-probability regions (i.e., $P > \langle P \rangle$) (% of total grid points).
[b] Quadrant of Ramachandran map.

in determining the ensemble is the inclusion of solvation effects (implicit or explicit) within the prediction algorithm. However, we opted for a simple screened, distance-dependent Coulombic potential (4) with $\epsilon_0 = 80$, to approximate the dielectric properties of water. This simple approximation leads to low energy peptide structures for S3 and IRP that agree with experimental observations (Johnson et al., 1993; Lyu et al., 1993). However, there are alternatives for implicit modeling of solvation effects in DPG-MC. One method includes the use of explicit counterions (e.g., Na^+, Cl^-) localized near charged side-chain moieties (e.g., Glu, Asp, His, Lys, Arg). Using the branching algorithm, each counterion can be considered the endpoint of each branch and therefore will travel with its assigned side chain during a dihedral displacement step. In fact, this approach has been applied to the study of lowest energy conformational states for polyelectrolyte peptides (i.e., Poly-L-$(Asp)_{20}$, Poly-L-$(PSer)_{20}$) and will be presented elsewhere (Evans, 1992; J.S. Evans, S.I. Chan, & W.A. Goddard III, manuscript in prep.).

Although DPG-MC exhibits an ability to locate global minima structures for small peptides (Figs. 1, 2, 3, 4), we have not established that DPG-MC will avoid trapping into secondary minima for larger structures such as proteins. However, for larger systems, DPG-MC can be used as part of an overall hierarchical protein structure refinement scheme (Mathiowetz, 1992): (1) initially virtual bond C_α structures are generated by a lattice search algorithm; (2) the all-atom structure is reconstructed from the virtual bond representation, and the conformation is refined by the DPG-MC program; (3) finally, the all-atom structure is subjected to molecular dynamics refinement and/or minimization (Mathiowetz, 1992). This approach is currently being used for a number of systems.

Biased probability Monte Carlo (BP-MC)

While preparing this manuscript, the biased probability Monte Carlo (BP-MC) method appeared in the literature (Abagyan & Totrov, 1994). This method is similar to DPG-MC. It utilizes a dihedral Brookhaven protein database, nonlocal step procedures, and an optimal probability distribution function (Abagyan & Totrov, 1994). The BP-MC method randomly selects the subspace first, then makes a step to a new random position independent of the previous position but according to the predefined continuous probability distribution. (Compared with DPG-MC, the BP-MC utilizes a larger and somewhat lower resolution data set for determining the (ϕ, ψ) distribution.) It uses 191 total structures (rather than 64), allows a resolution up to 2.4 Å (rather than 2.0 Å), and includes 35% sequence overlap (rather than 25%). For side-chain angles, it includes 161 total structures with resolution ≤2 Å and 50% sequence overlap.

The BP-MC method was applied to the de novo prediction of all-helical and nonhelical peptides, but was not tested for its ability to accurately predict β-strand or β-turn structures, nor for sequences that feature a mixture of secondary structures.

Computational details

Selection of the protein database and creation of the dihedral probability grids

In order to generate a set of dihedral probabilities representing a diverse sampling of accurately known protein structures, we considered the 503 proteins in the PDB with resolution ≤1.5 Å, or with resolution ≤2.0 Å and R-factors <20%.

Using the "align" program (Pearson, 1990), we conducted pairwise sequence comparisons and eliminated any protein having greater than 25% sequence homology with another protein of higher resolution. This resulted in a data set consisting of 64 proteins (termed H64) (Table 3).

We next analyzed the ϕ, ψ, and χ dihedral angles for each of the 20 amino acids in each protein structure of the H64 data set. For each amino acid, a grid is constructed as an N-dimensional matrix, where N is the number of dihedrals involved. For example, backbone sampling involves two-dimensional grids, and each point on the grid represents the probability of choosing a particular (ϕ, ψ) pair. The grids were constructed with spacings of $S = 5$, 10, 15, 30, or 60°. The probabilities for each point on the grid were derived by partitioning every (ϕ, ψ) pair and χ from the H64 set into S-degree bins. The probabilities, P, were normalized such that:

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

where $\phi_i = iS$ and $\psi_j = jS$. We constructed separate backbone (ϕ, ψ) probability grids for each amino acid. However, we found that it is sufficient to use individual grids for the three major residue types: (1) glycine (G), no side chain; (2) proline (P), whose side chain forms a closed loop with the backbone; and (3) r18, or the remaining 18 "standard" amino acids.

As an example, the (ϕ, ψ) probability distributions for the three amino acid types, are given in Figure 8 for $S = 30°$. The boundary regions of this grid are less smooth than for the finer $S = 5°$ and 10° grid spacings (Mathiowetz, 1992); however, we found that $S = 30°$ leads to broader spanning of conformational space than the finer grids. The (ϕ, ψ) grids are substantially different for Pro, Gly, and r18 (Fig. 8; Table 2). Gly has high-probability conformations in all four quadrants of the Ramachandran plot because there is no R group to sterically restrict conformations (Table 2). Pro, in contrast, has few high-probability (ϕ, ψ) grid points and is centered in two quadrants of the Ramachandran map due to the geometrical constraints of the imine ring (Table 2; Fig. 8); note the ϕ angle restriction near −60° in Figure 8. The r18 data points lead to a very narrow distribution in the high-probability α-helical region and a broad, low-probability distribution in the β-sheet region (Fig. 8). However, the β-sheet quadrant, I, has nearly the same *overall* probability as the α-helix quadrant, II (Table 2).

Side-chain probability grids, χ, were constructed for each amino acid side chain. We did not consider dihedral angles that affect only hydrogen positions or those involved in rings (except Pro). The number of χ dihedrals (N_χ) under consideration is small: $N_\chi = 0$ for Ala and Gly; $N_\chi = 2$ for His, Tyr, Trp, and Phe (Table 4). Although Pro is a ring structure, we allow χ_1 to vary while holding the C_δ atom fixed. This enables reasonable conformations of χ_1-χ_4 to be sampled by modifying only the single dihedral, χ_1. We determined the occurrence of each amino acid and the total number of grid points at each S level, along with the χ distribution and the number of populated grid points for amino acids that possess significant χ_1 and χ_2 dihedrals (Table 4; Fig. 9, for $S = 30°$). It is difficult to display the higher dimensional grids for Arg, Met, Gln, Glu, and Lys in their entirety. It is evident that conformational variability (i.e., the number of possible conformations) increases as a function of N_χ (Table 4; note populated grid points for Lys, Arg, Glu, and Gln). The χ distributions, particularly for χ_1 and χ_2, exhibit significant variations from amino acid to amino acid (Fig. 9), and many (χ_1, χ_2) conformations for a given side chain have similar probabilities.

The Monte Carlo method

A diagrammatic representation of the DPG-MC algorithm is given in Figure 10. Briefly, the conformations of a polypeptide are generated by (1) randomly selecting a residue, (2) randomly choosing which dihedral to alter ((ϕ, ψ) pair, or χ), and then (3) obtaining the corresponding dihedral value from the amino acid-specific probability grid.

In DPG-MC we do not currently alter the value of ω (this could be included). Because Pro forms a closed loop with the backbone, the imine ϕ remains fixed during the DPG-MC simulation. Next, (4) the energy of the new structure is evaluated and compared to the previous structure. If the energy of the new structure is lower, it is saved and used as the starting point for the next torsional motion. If the new structure is *not* lower in energy, then it is accepted with a probability of

$$P = \exp\left[-\frac{\Delta E}{k_B T_{MC}}\right], \tag{2}$$

where k_B is the Boltzmann constant and T_{MC} is the simulation temperature (Metropolis et al., 1953). A high T_{MC} creates a large number of "bad" structures but also provides alternative "paths" that may eventually lead to lower energy minima. (5) Minimization in Cartesian coordinate space (e.g., steepest descents, conjugate-gradient, Fletcher–Powell) may be performed following the conformational selection but prior to Metropolis sampling. We used the DREIDING force field (Mayo et al., 1990), which yields excellent results for structural minimization; however, other force fields, i.e., AMBER (Weiner et al., 1986), CHARMM (Brooks et al., 1983), MM3 (Lii & Allinger, 1991), and UFF (Rappè et al., 1992), can be utilized as well.

The DPG-MC program relies on a "branching" algorithm (Abe et al., 1984) to define torsion angles of side chains in terms of a set of connected atoms that "branches" off of the backbone. All torques occur at branch points. An additional benefit of the branch algorithm is the ability to model counterions or other atoms/molecules at side-chain sites, where the algorithm counts these counterions as part of the side-chain "branch." Hence, the counterions move with the side chain as the dihedrals are changed (it will not be "orphaned").

Procedure for the determination of polypeptide conformational minima

All calculations presented here were performed using a modified version of BIOGRAF (Molecular Simulations, Inc., 1992) running on Silicon Graphics workstations (models 4D/380, 4D/35, 4D/25, and Indigo XZ were used). Most simulations used the DREIDING force field (Mayo et al., 1990), but the AMBER force field (Weiner et al., 1986) was used for some comparisons.

To determine the performance parameters of the Monte Carlo program, we utilized the five-residue peptide Met^5-enkephalin, which has been used as a benchmark molecule for theoretical simulations of polypeptide conformation (Lambert & Scheraga,

Table 3. *H64 set of Brookhaven protein crystal structures*[a]

PDB	Protein	Resolution (Å)	*R*
1AMT	Alamethicin (*Trichoderma*)	1.5	0.155
1BP2	Phospholipase A2 (bovine)	1.7	0.171
1CRN	Crambin (Abyssinian cabbage)	1.5	N/A
1CSC	Citrate synthetase (chicken)	1.7	0.168
1CSE	Subtilisin (*Bacillus subtilis*)	1.7	0.188
1CTF	L12 50S ribosomal (*Escherichia coli*)	1.2	0.178
1ECA	Hb erythocruori (*Chironomus*)	1.4	0.136
1FB4	IgG FAB (human)	1.9	0.189
1GD1	D-Glyceraldehyde dehydrogenase	1.8	0.177
1GMA	Gramicidin A	0.86	0.071
1GP1	Glutathione peroxidase (bovine)	2.0	0.171
1HOE	α-Amylase inhibitor (*Streptomyces*)	2.0	0.199
1I1B	Interleukin-1b (human)	2.0	0.189
1L19	Lysozyme (bacteriophage T)	1.7	0.153
1LZ1	Lysozyme (human)	1.5	0.177
1MBA	Metmyoglobin (sea hare)	1.6	0.193
1MBD	Deoxymyoglobin (sperm whale)	1.4	N/A
1NXB	Neurotoxin (sea snake)	1.38	N/A
1PAZ	Pseudoazurin$_{ox}$ (*Alcaligenes*)	1.55	0.180
1PCY	Plastocyanin Cu(II) (*P. populus*)	1.6	0.170
1PPT	Avian pancreatic polypeptide	1.37	N/A
1THB	Hb, T-state (human)	1.5	0.196
1UBQ	Ubiquitin (human)	1.8	0.176
1UTG	Uteroglobin$_{ox}$ (rabbit)	1.34	0.230
1XY1	β-Mercaptoproprionate	1.04	0.088
256B	Cyt$_{ox}$ b_{562} (*E. coli*)	1.4	0.164
2AZA	Azurin$_{ox}$ (*Alcaligenes*)	1.8	0.157
2CA2	Carbonic anhydrase (human)	1.9	0.176
2CCY	Cyt *c'* (*Rhodospirillum*)	1.67	0.188
2CDV	Cyt c_3 (*Desulfovibrio*)	1.8	0.176
2CCP	Cyt P_{450} (*Pseudomonas*)	1.63	0.190
2CYP	Cyt *c* peroxidase (yeast)	1.7	0.202
2ER7	E aspartic protein (chestnut blight)	1.6	0.142
2GBP	D-Gal/D-Glu binding (*E. coli*)	1.9	0.146
2LTN	Pea lectin (garden pea)	1.7	0.177
2MHR	Myohemerythrin (sipunculan worm)	1.7	0.158
2MLT	Mellitin (honey bee)	2.0	0.198
2OVO	Ovomucoid 3rd domain (pheasant)	1.5	0.199
2RSP	Rous sarcoma virus protease	2.0	0.144
2SGA	Proteinase A component (*Streptomyces*)	1.5	0.126
2SNS	Staphylococcal nuclease (*Staphylococcus*)	1.5	N/A
2WRP	Trp repressor (*E. coli*)	1.65	0.180
3B5C	Cyt$_{ox}$ b_5 (bovine)	1.5	0.160
3BCL	Bacteriochlorophyll A protein	1.9	0.189
3BLM	β-Lactamase (*Staphylococcus*)	2.0	0.164
3CLA	Type III chloramphenicol binding protein (*E. coli*)	1.75	0.157
3DFR	Dihydrofolate reductase (*Lactobacillus*)	1.7	0.152
3GRS	Glutathione reductase (human)	1.9	0.186
3RNT	Lys 25 ribonuclease T1 (*Aspergillus*)	1.8	0.137
451C	Cyt$_{red}$ C_{551} (*Pseudomonas*)	1.6	0.187
4CPV	Parvalbumin (Ca^{2+}) (carp)	1.5	0.215
4FD1	Ferredoxin (*Azotobacter*)	1.9	0.192
4FXN	Flavodoxin (*Clostridium*)	1.8	0.200
4INS	Insulin (porcine)	1.5	0.153
4PTP	β-Trypsin (diisopropyl-) (bovine)	1.34	0.171
5CPA	Carboxypeptidase-a (bovine)	1.54	N/A
5CYT	Cyt c_{red} (tuna)	1.5	0.171
5PTI	Trypsin inhibitor (bovine)	1.0	0.200
5RXN	Rubredoxin$_{ox}$ (Clostridium)	1.20	0.115
5TNC	Troponin C (avian)	2.0	0.155
6TMN	Thermolysin (*Bacillus thermophilus*)	1.6	0.171
7RSA	Ribonuclease A (bovine)	1.26	0.15
9PAP	Papain (papaya)	1.65	0.161
9WGA	Agglutinin (wheat germ)	1.8	0.175

[a] Standard PDB abbreviations are used. *R*, *R*-factor, expressed as decimal; N/A, not available; ox, oxidized form; red, reduced form.

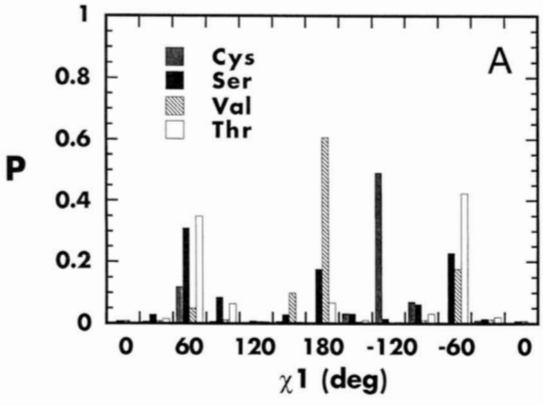
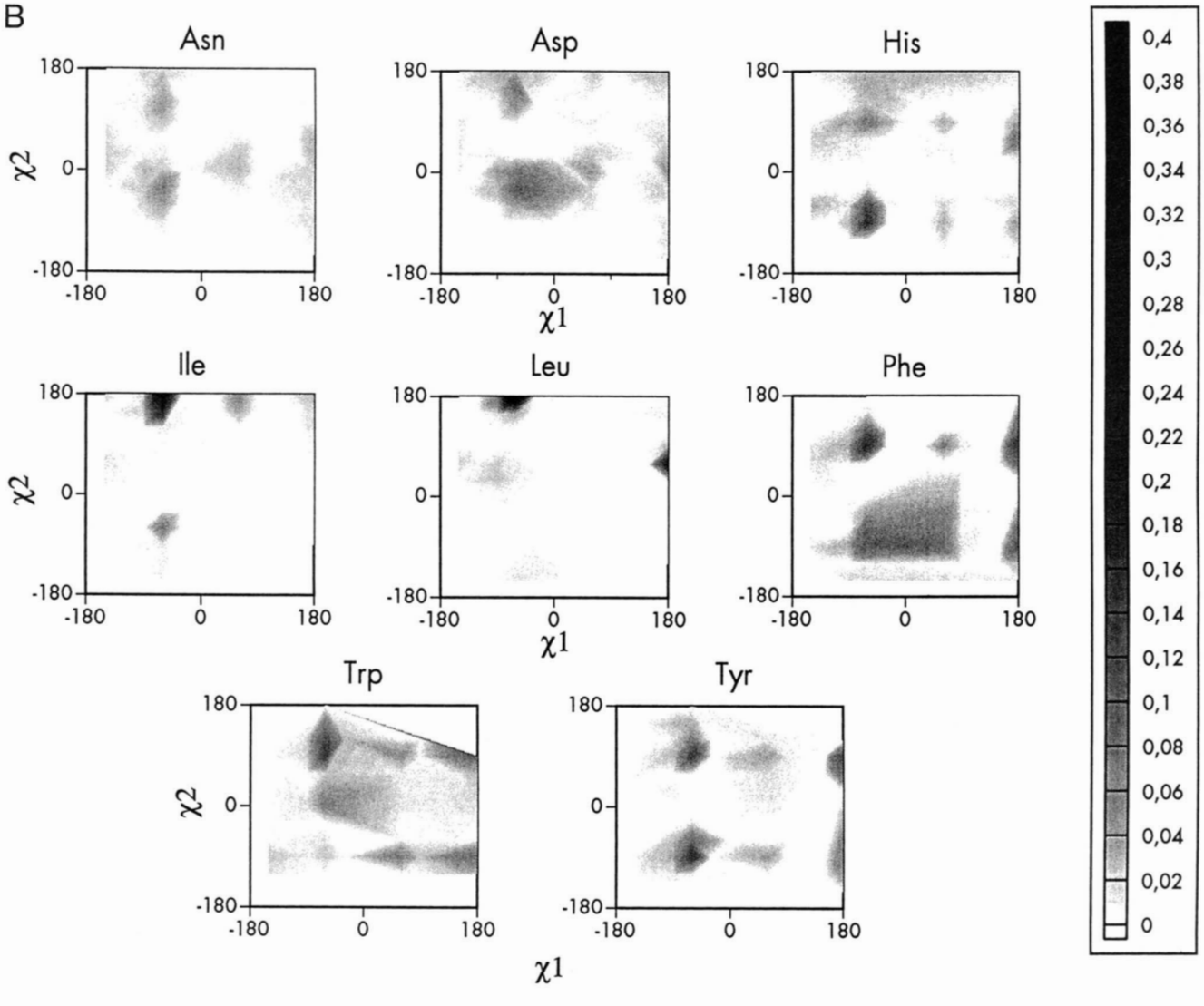


Fig. 9. A: χ_1 Probability histogram for the H64 DPB data set using $S = 30°$. **B:** (χ_1, χ_2) probability contour plot for the H64 PDB data set using $S = 30°$. Higher probability regions are denoted by increasing gray scale values; regions that are unpopulated are indicated in white (see legend).



1989; Nayeem et al., 1991; Dorofeyev & Mazur, 1993). The small size of Met[5]-enkephalin (22 dihedral angles, 75 total atoms) permits rapid energy evaluation, and thus we can track the DPG-MC simulation over a large number of steps and numerous runs without requiring excessive time to complete each run. Single-temperature Monte Carlo simulations were conducted at 0, 300, 1,000, 3,000, 5,000, and 10,000 K, using 2×10^5 steps/run. Grid spacings of 5, 10, 30, and 60° were utilized in parallel runs; we found little difference between the 10° and 15° grid simulations. In order to determine the effect of Monte Carlo

Table 4. *Populated grid points for side-chain dihedrals, χ*

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

[a] $N_\psi = 0$ for Gly and Ala.
[b] Number of occurrences of each residue in the H64 data set. For each S value, the number of grid points with a population non-zero grid points is given.
[c] Number of ψ dihedrals.

with minimization, some runs featured N steps of conjugate-gradient minimization (where $N = 1$ or 10) after each dihedral torque but prior to Metropolis sampling. The Met[5]-enkephalin polypeptide was constructed in the α-NH_3^+/α-COO^- form using the peptide builder function within BIOGRAF. The initial starting structure was set in the all-extended conformation ($(\phi, \psi) = -180°$; $\chi = 60°$). The net charge of the peptide (Q_{net}) was set to 0.000 and charge equilibration was performed using Charge Equilibration (QEq) (Rappè & Goddard, 1991). The structures were then minimized to convergence using 10 steps of steepest descents followed by conjugate-gradient minimization (RMS force < 0.1, at 300 K) (Kini & Evans, 1991). These minimized structures were then input into the DPG-MC simulations. At the end of each DPG-MC simulation, the lowest energy minima structure was subjected to minimization in Cartesian coordinate space using 10 steps of steepest-descents minimization, followed by conjugate-gradient minimization to convergence (RMS force < 0.1, 300 K). For both DPG-MC and minimization runs, a dielectric (ϵ_0) of 1 was utilized.

For the IRP and S3 peptides, the procedure was identical to that described for Met[5]-enkephalin, except that a larger number of Monte Carlo steps was utilized per simulation run (3×10^5 steps/run, eight total runs for YGRGDSP; 1×10^6 steps/run, eight total runs for S3), a grid spacing of 5° was used exclusively, and one step of conjugate-gradient minimization was performed at each step of the DPG-MC run. Asp and Glu

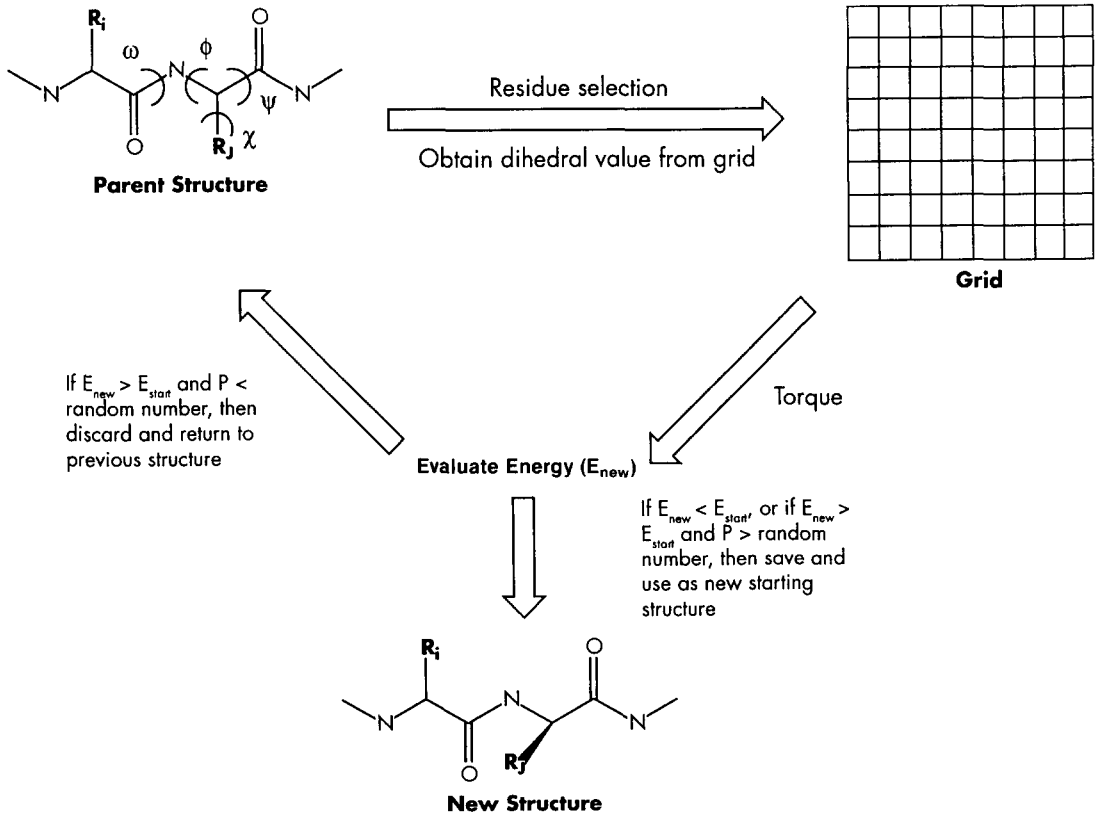


Fig. 10. Schematic representation of the DPG-MC algorithm. A local move in conformational space begins (upper left) by starting from a random point on the polypeptide, where an internal coordinate ((ϕ, ψ) pair, or χ_n) is identified. The amino acid at this position is identified and a dihedral value is obtained from the corresponding H64 data set grid (upper right). The new dihedral value is applied to the structure, and the internal energy is computed (center). This energy is compared to the starting conformer energy (bottom). Metropolis sampling is utilized to save or reject the structure. This continues to the upper left.

residues were represented as deprotonated, negatively charged species; Lys and Arg as protonated, positively charged species; and His as a neutral species (assuming pH > 7.0). Both termini were represented as charged species as per Met^5-enkephalin. Q_{net} for IRP and S3 were equilibrated to 0.000 and −1.000, respectively. Because the structures of IRP and S3 were determined in solution using NMR spectroscopy (Johnson et al., 1993; Lyu et al., 1993), we included the electrostatic screening effects in the Monte Carlo simulation. The use of explicit solvent molecules leads to more accurate simulations of protein and peptide structures (Steinbach et al., 1991; Smith & Pettitt, 1992; Collura et al., 1994). However, the computational cost for such simulations is prohibitive. Implicit approaches to modeling the electrostatic effects of solvent on charged polypeptides use distance-dependent Coulombic potentials in conjunction with an optimal value for the dielectric constant, ϵ (Fersht & Sternberg, 1989; Naylor & Goddard, 1989; Mehler & Solmajer, 1991; Moult, 1992; Arnold & Ornstein, 1994; Collura et al., 1994).

To implicitly simulate the electrostatic effects of solvent environment on peptide conformation, we utilized the screened distance-dependent Coulombic potential, E_Q (Mayo et al., 1990):

$$E_Q = \frac{(322.0637)Q_iQ_j}{\epsilon R_{ij}}, \quad (3)$$

where

$$\epsilon = \epsilon_0 R_{ij}, \quad (4)$$

Q_i and Q_j are the charges (in electron units) for atoms i and j, R_{ij} is the distance in angstroms, ϵ is the dielectric constant, and the constant 332.0637 converts units so that the energy is in kcal/mol. This should yield a good first approximation for short-range side-chain–side-chain electrostatic interactions (i.e., 9.0 Å or less) (Fersht & Sternberg, 1989; Mehler & Solmajer, 1991) while permitting long-range electrostatic interactions. To mimic solvation effects on the optimum geometry of IRP and S3, we utilized $\epsilon_0 = 80$, which approximates the dielectric of water (Arnold & Ornstein, 1994; Collura et al., 1994). Electrostatic (E_Q) and van der Waals (vdW) nonbonding potentials were evaluated (without cutoffs) at each step of the DPG-MC simulation and during the Cartesian minimizations.

Acknowledgments

We thank Dr. Siddharth Dasgupta for helpful advice during this study. J.S.E. acknowledges a Postdoctoral National Research Service Award from the NIH (NIDR 1F32-DE-05445) and a fellowship award from AMGEN Pharmaceuticals. A.M.M. acknowledges a National Research Service Award/NIH Predoctoral Biotechnology Traineeship. These studies were supported by a grant from DOE-AICD. The facilities of the MSC are also supported by grants from the NSF (CHE91-00284 and ASC-9219368), 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 were carried out on the NSF Pittsburgh Supercomputer and on the JPL Cray. This is Contribution 8949 from the Division of Chemistry and Chemical Engineering, California Institute of Technology.

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.

Abe H, Braun W, Noguti T, Gō N. 1984. Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins. General recurrent equations. *Comput Chem 8*:239–247.

Arnold GE, Ornstein RL. 1994. An evaluation of implicit and explicit solvent model systems for the molecular dynamics simulation of bacteriophage T4 lysozyme. *Proteins Struct Funct Genet 18*:19–33.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem 4*:187–217.

Collura VP, Greaney PJ, Robson B. 1994. A method for rapidly assessing and refining simple solvent treatments in molecular modeling. Example studies on the antigen-combining loop H2 from FAB fragment McPC603. *Protein Eng 7*:221–235.

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

Ding HQ, Karasawa N, Goddard WA III. 1992. Optimal spline cutoffs for Coulomb and van der Waals interactions. *Chem Phys Lett 193*:197–201.

Dorofeyev VE, Mazur AK. 1993. Investigation of conformational equilibrium of polypeptides by internal coordinate stochastic dynamics. Met^5-enkephalin. *J Biomol Struct Dyn 10*:143–167.

Dyson HJ, Rance M, Houghten RA, Lerner RA, Wright PE. 1988. Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. *J Mol Biol 201*:161–200.

Evans JS. 1992. NMR and computational studies of conformational folding in the biomineralization template, phosphophoryn [thesis]. Pasadena: California Institute of Technology.

Fersht AR, Sternberg MJE. 1989. Can a simple function for the dielectric response model electrostatic effects in globular proteins? *Protein Eng 2*:527–530.

Geourjon C, Deleage G. 1994. SPM: A self-optimized method for protein secondary structure prediction. *Protein Eng 7*:157–165.

Gerstein M, Chothia C. 1991. Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. *J Mol Biol 220*:133–149.

Johnson WC, Pagano TG, Basson CT, Madri JA, Gooley P, Armitage IA. 1993. Biologically active Arg-Gly-Asp oligopeptides assume a type II β-turn in solution. *Biochemistry 32*:268–273.

Judson RS, Jaeger EP, Treasurywala AM, Peterson ML. 1993. Conformational searching methods for small molecules. II. Genetic algorithm approach. *J Comput Chem 14*:1407–1414.

Kawai H, Kikuchi T, Okamoto Y. 1989. A prediction of tertiary structures of peptide by Monte Carlo simulated annealing method. *Protein Eng 3*:85–94.

Kini RM, Evans HJ. 1991. Molecular modeling of proteins: A strategy for energy minimization by molecular mechanics in the AMBER force field. *J Biomol Struct Dyn 9*:475–488.

Kocher JPA, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol 235*:1598–1614.

Lambert MH, Scheraga HA. 1989. Pattern recognition in the prediction of protein structure. III. An importance-sampling minimization procedure. *J Comput Chem 10*:817–831.

Lee FS, Chu ZT, Warshel A. 1993. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. *J Comput Chem 14*:161–185.

Lii JH, Allinger NL. 1991. The MM3 forcefield for amides, polypeptides and proteins. *J Comput Chem 12*:186–199.

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. *Science 251*:1162–1164

Lyu P, Wemmer DE, Zhou HX, Pinker RJ, Kallenbach NR. 1993. Capping interactions in isolated α-helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. *Biochemistry 32*:421–425.

Madej T, Mossing MC. 1993. Hamiltonians for protein tertiary structure prediction based on three-dimensional environment principles. *J Mol Biol 234*:480–487.

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

Matsushima N, Creutz CE, Kretsinger RH. 1990. Polyproline, β-turn helices. Novel secondary structures proposed for the tandem repeats within rhodopsin, synaptophysin, synexin, gliadin, RNA polymerase II, hordein, and gluten. *Proteins Struct Funct Genet 7*:125–155.

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

McGarrah DB, Judson RS. 1993. Analysis of the genetic algorithm method of molecular conformation determination. *J Comput Chem 14*:1385–1395.

Mehler EL, Solmajer T. 1991. Electrostatic effects in proteins: Comparison of dielectric and charge models. *Protein Eng 4*:903–910.

Merutka G, Morikis D, Bruschweiler R, Wright PE. 1993. NMR evidence

for multiple conformations in a highly helical model peptide. *Biochemistry 32*:13089–13097.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*:1087–1092.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.
Molecular Simulations, Inc. 1992. *BIOGRAF/POLYGRAF*. Burlington, Massachusetts: Molecular Simulations, Inc.
Moult J. 1992. Electrostatics. *Curr Opin Struct Biol 2*:223–229.
Nayeem A, Vila J, Scheraga HA. 1991. A comparative study of the simulated-annealing and Monte-Carlo-with-minimization approaches to the minimum energy structures of polypeptides: (Met)-enkephalin. *J Comput Chem 12*:594–605.
Naylor A, Goddard WA III. 1989. Application of simulation and theory to biocatalysis and biomimetics. In: Burrington JD, Clark DS, eds. *Biocatalysis and biomimetics (ACS Symposium Series 392)*. Washington, D.C.: American Chemical Society. pp 65–87.
Ngo JT, Marks J. 1992. Computational complexity of a problem in molecular structure prediction. *Protein Eng 5*:313–321.
Paine GH, Scheraga HA. 1985. Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. I. Backbone structure of enkephalin. *Biopolymers 24*:1391–1436.
Pearson WR. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol 183*:63–98.
Purisima EO, Scheraga HA. 1984. Conversion from a virtual bond chain to a complete polypeptide backbone chain. *Biopolymers 23*:1207–1224.
Rappè AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM. 1992. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J Am Chem Soc 114*:10024–10035.
Rappè AK, Goddard WA. 1991. Charge equilibration for molecular dynamics simulations. *J Phys Chem 95*:3358–3363.
Rey A, Skolnick J. 1991. Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. *Chem Phys 158*:199–219.
Rey A, Skolnick J. 1993. Computer modeling and folding of four-helix bundles. *Proteins Struct Funct Genet 16*:8–28.
Rooman MJ, Kocher JP, Wodak SJ. 1992. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry 31*:10226–10238.
Schreiber H, Steinhauser O. 1992. Cutoff size does strongly influence molecular dynamics results on solvated polypeptides. *Biochemistry 31*:5856–5860.
Shin JK, Jhon MS. 1991. High directional Monte Carlo procedure coupled with the temperature heating and annealing as a method to obtain the global energy minimum structure of polypeptides and proteins. *Biopolymers 31*:177–185.
Sippl MJ, Hendlich M, Lackner P. 1992. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin β_4. *Protein Sci 1*:625–640.
Skolnick J, Kolinski A. 1990. Simulations of the folding of a globular protein. *Science 250*:1121–1124.
Skolnick J, Kolinski A. 1991. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure, and dynamics. *J Mol Biol 221*:499–531.
Smith PE, Pettitt BM. 1992. Amino acid side-chain populations in aqueous and saline solution: Bis-penicillamine enkephalin. *Biopolymers 32*:1623–1629.
Srinivasan S, March CJ, Sudarsanam S. 1993. An automated method for modeling proteins on known templates using distance geometry. *Protein Sci 2*:277–289.
Steinbach PJ, Loncharich RJ, Brooks BR. 1991. The effects of environment and hydration on protein dynamics: A simulation study of myoglobin. *Chem Phys 158*:383–394.
Weiner SJ, Kollman PA, Nguyen DT, Case DA. 1986. An all atom force field for simulations of proteins and nucleic acids. *J Comput Chem 7*:230–252.
Wilmot CM, Thornton JM. 1990. β-Turns and their distortions: A proposed new nomenclature. *Protein Eng 3*:479–493.
Wright PE, Dyson HJ, Feher VA, Tennant LL, Waltho P, Lerner RA, Case DA. 1990. Folding of peptide fragments of proteins in aqueous solution. In: Wright PE, Case DA, eds. *Frontiers of NMR in molecular biology*. New York: Alan R. Liss, Inc. pp 1–13.