PROTEINS: Structure, Function, and Genetics 14:213–223 (1992)

Fast and Simple Monte Carlo Algorithm for Side Chain Optimization in Proteins: Application to Model Building by Homology

Liisa Holm and Chris Sander
European Molecular Biology Laboratory, D-6900 Heidelberg, Federal Republic of Germany

ABSTRACT **An unknown protein structure can be predicted with fair accuracy once an evolutionary connection at the sequence level has been made to a protein of known 3-D structure. In model building by homology, one typically starts with a backbone framework, rebuilds new loop regions, and replaces nonconserved side chains. Here, we use an extremely efficient Monte Carlo algorithm in rotamer space with simulated annealing and simple potential energy functions to optimize the packing of side chains on given backbone models. Optimized models are generated within minutes on a workstation, with reasonable accuracy (average of 81% side chain χ_1 dihedral angles correct in the cores of proteins determined at better than 2.5 Å resolution). As expected, the quality of the models decreases with decreasing accuracy of backbone coordinates. If the backbone was taken from a homologous rather than the same protein, about 70% side chain χ_1 angles were modeled correctly in the core in a case of strong homology and about 60% in a case of medium homology. The algorithm can be used in automated, fast, and reproducible model building by homology.**

Key words: protein folding, protein structure, rotamers, simulated annealing

INTRODUCTION

The protein folding problem remains essentially unsolved in spite of 20 years of persistent efforts. The three principal difficulties lie in the delicate energetic balance between enthalpy and entropy, in the highly cooperative character of the folding process, and in the fact that protein sequences have evolved to satisfy diverse functional constraints and are not optimized to merely fold up correctly. Meanwhile, quietly, and as a result of much hard work by people who solve structures and determine sequences experimentally, a simple but powerful way to make predictions concerning the 3-D structure of proteins has been developed. This partial solution to the folding problem is model building by homology. It works only sometimes, but when it works, it works like a charm. For example, a correct approximate model of the HIV protease was built based on the identification of a catalytic triad characteristic of the acid proteases.[1]

Model building by homology exploits the evolutionary fact that in protein families the structure of the core is highly conserved in spite of divergence of sequences. In general, core backbones can be superposed with a root mean square deviation of C_α atom (rmsd) positions of less than 1 Å if sequence identity in the core is 50% or higher.[2] Loop conformations are more variable, necessarily so if there are insertions or deletions. Observations that similar substructures occur time and again in known protein structures have led to a widely used and convenient method to generate loop conformations by replacing the backbone by a fragment retrieved from the structural database.[3] More time-consuming methods include systematic search of backbone torsion angles,[4,5] molecular dynamics,[6] and distance geometry,[7] followed by final evaluation of the quality of loop conformations in the context of the rest of the protein.

The optimization of side chain packing is a complex combinatorial problem because of multiple minima in energy as a function of dihedral angles. The complexity can be reduced to a manageable level by two kinds of simplification. Several methods perform local energy minimization of very few residues in the field of otherwise fixed protein atoms.[8–10] An alternative is to restrict the number of conformational states that side chains can occupy.[11–13] These latter methods exploit the observation that the statistical distribution of side chain χ angles has sharp peaks[14] and restrict an energetic search to sampling only the most heavily populated subspace. In one such approach,[11] we developed a Monte Carlo algorithm for optimizing side chain conformations in the context of generating full protein coordinates from a mere C_α trace. In this pa-

Received August 15, 1991; accepted December 26, 1991.
Address reprint requests to either Liisa Holm or Chris Sander, European Molecular Biology Laboratory, Postfach 10-2209, Meyerhofstrasse 1, D-6900 Heidelberg, Federal Republic of Germany.

per, we show how the Monte Carlo algorithm can be optimized and used efficiently to explore the multiple minima of side chain conformations and demonstrate its potential in practical model building by homology.

METHODS

Protein Coordinates

The method for side chain construction was tested using only the backbone coordinates of known structures and letting a computer program fill in side chain coordinates. The predicted models were then compared to the original fully known coordinates. All protein coordinates were retrieved from the Protein Data Bank.[15] Backbone reconstruction from the C_α trace was carried out using MaxSprout.[11] Proteins which are homologous to the one being built were excluded from the fragment database.

Rotamer Library

The statistical distribution of sidechain χ angles is relatively sharp.[14] Unless stated otherwise, results reported here are generated using the rotamer library of Tuffery et al.[13] The number of rotamers is 1 for proline, 3 for cysteine, serine, threonine, valine, and aspartic acid, 4 for asparagine, 5 for isoleucine, 6 for leucine and histidine, 7 for tryptophan, 10 for glutamine, glutamic acid, methionine, and arginine and 16 for lysine.

Computational Details

3-D coordinates

The coordinates of backbone atoms (N, C_α, C, O, C_β) and, optionally, fixed side chains are taken directly from the input file and form the set of *static atoms*. If C_β coordinates are not present in the input file (e.g., in modeling a Gly→non-Gly mutation), then the position of C_β is calculated from N, C_α, and C coordinates using standard bond lengths and angles. Prosthetic groups, ligands, crystallographic water molecules, and hydrogen atoms are ignored.

Atoms which belong to the same side chain can only move between the discrete rotamer conformations. Cartesian coordinates for all possible side chain conformations are generated by mounting the appropriate prototype rotamer on the backbone using a least squares fit of the N, C_α, C, and C_β atoms and adjusting the dihedral χ angles as specified in the rotamer library.

Energy lookup tables

The total energy is calculated by summing pairwise atomic interaction energies using the truncated 6–9-potentials in Table I and a cutoff radius of 6 Å. Covalent, 1–3 and 1–4 neighbors are excluded. A 3-D grid is used in the search for neighbors so that the computation time for this part grows roughly linearly with the number of atoms in the system.

TABLE I. Parameters for the A/r^6-B/r^9 Potential*

Atom types	r(min) [Å]	r(trunc) [Å]	A	B
S–S	2.0	1.25	131.1	24.6
N–O	3.0	1.75	5039	280
All others	4.0	2.25	67000	1573

*Depending on atom type, the potential energy minimum is at r(min).
For $r < r$(trunc), the potential energy is constant at $E[r(\text{trunc})]$. Energy units are arbitrary.

Calculating the energy of the system after small adjustments in atomic positions is the most time-consuming step in conventional energy minimization and molecular dynamics because it requires in principle a double loop over all atoms. In our discrete model all possible pair interactions can be precalculated and stored in tables. The array *estatic* contains an entry for each rotamer conformation that a given side chain can adopt, and holds its interaction energy with all static atoms (backbone and fixed sidechains). Rotamer conformations with bad clashes with static atoms are rejected after the calculation of *estatic*. A clash is considered bad if the energy is above 10 energy units (Table I). If no rotamer has an energy below 10, then the cutoff is set at 10 energy units above the best rotamer for the residue. Between one-fourth and one-third of the rotamers are rejected by this filter. Rotamer–rotamer interactions are calculated only for the remaining rotamers and stored in the array *epair*. To save memory, only repulsive/attractive rotamer–rotamer interactions with an absolute value larger than 0.1 energy units are kept.

The configuration of the protein model is represented as a 1-D list of the currently selected rotamer for each residue in the protein. The total energy is the sum of *estatic* and *epair* over all currently selected rotamers. In this way, the calculation of the energy difference between two successive configurations of the model is reduced to a few additions and subtractions (Fig. 1).

Monte Carlo simulated annealing optimization of rotamer choice

The optimization problem is combinatorial: find the set of rotamers, one per residue, which minimizes the total energy. The key idea of Monte Carlo optimization is iterative improvement in which controlled uphill steps can occasionally be incorporated in the search of minima in complex energy landscapes.[16] The higher the temperature, the more uphill steps are accepted. In simulated annealing, the temperature of the system is gradually lowered until it settles in a minimum.[17]

The initial rotamer configuration is random. An iteration step consists of trying a change of the rota-

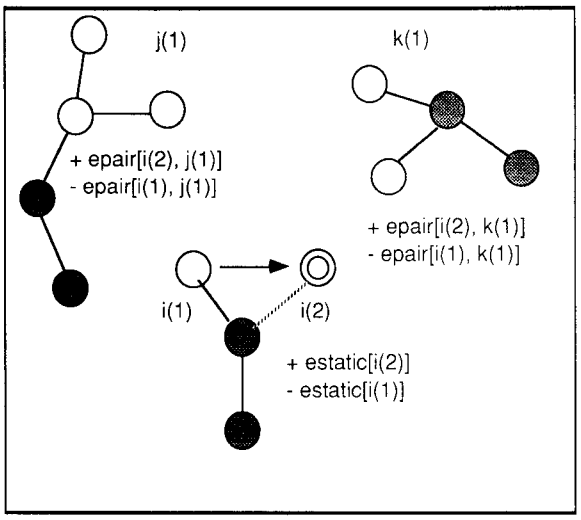

Fig. 1. Precalculation of pair energies speeds up the calculation of total energy after a change in rotamer configuration. This simplified representation of a protein environment shows residue *i* is in contact with residues *j* and *k*. Hatched circles are static atoms with fixed positions throughout the optimization (shown here are C_α and C_β). Open circles are atoms at γ, δ, ϵ etc. positions along the side chain; the position of these atoms depends on the conformational state of the residue, indicated by the rotamer number in parentheses. The open concentric circle marks a new atomic position for the side chain of residue *i* after one Monte Carlo move. The energy difference between states {*i*(1) *j*(1) *k*(1)} and {*i*(2) *j*(1) *k*(1)} is computed from precalculated terms *estatic* and *epair* as indicated in the figure.

mer state of one randomly chosen residue. The probability p of accepting the change is $p = 1/[1 + \exp(\alpha \cdot \Delta E)]$, where α corresponds to the inverse temperature and ΔE is the energy difference between the new and old configuration. The formula used here is modified from the original Metropolis criterion [$p = 1/\exp(\Delta E/kT)$] to make the initial probability (small α), of accepting uphill steps one-half and independent of the energy difference. The decision of acceptance or rejection is made by comparing p to a random number in the interval (0,1). The optimization is started from $\alpha = 0.0$, but at each step α is increased by an amount δ. Iteration is terminated when the maximum number of steps has been reached ($\alpha = \alpha_{max}$) or there is no improvement in energy for a specified number of steps. The predicted conformation is the rotamer configuration which gave the lowest total energy during the entire simulation.

Model building by homology

3-D models of three pepsins (4APE, 3APR, and 1PSG) were built with 3APP as template. Suitable candidates were identified by a fast C_α–C_α distance search[3] between five anchor residues on both sides of the loop, and sorted according to C_α rmsd (root mean square positional deviation) fit in the anchor regions. Loops which clashed with the framework were rejected, after which the loop with the best C_α rmsd fit was selected. The fragment database consisted of the four pepsins mentioned above. In many, though not all, cases the native loop was selected and attached to the core of 3APP (the purpose here was to simulate an ideal loop backbone conformation predictor). For technical reasons, the insert of three helices (residues 8–49) in 1PSG relative to 3APP was not included in the model. After completing the backbone, side chains were optimized using the Monte Carlo algorithm.

Assessment of Models

The quality of side chain construction was assessed by rms positional deviation of side chain atoms, including C_β atoms, although these are generally taken from the input file. The rmsd error is strongly influenced by large sidechains. Another measure, side chain dihedral χ angle deviation from the X-ray structure, gives equal weight to small and large side chains. Sidechain χ_1 angles were classified as correct if the deviation was less or equal to 30°. Glycines and alanines were excluded from the calculation of the percentage of correctly predicted residues.

The structural core is defined as residues with less than 20% solvent accessibility in the X-ray structure, as defined by the solvent accessible surface area (program DSSP[18]) compared to extended peptides.[20] Residues not in the core are exposed residues.

As a measure of randomness of the rotamer distribution in optimized models, we calculate an entropy term at a given residue from $\sum[-f_i \cdot \ln(f_i)]$, where f_i is the frequency of a rotamer state i, the sum is over all rotamers for the amino acid type and rotamers with $f_i = 0$ do not contribute to the sum. The extreme values are zero for unique rotamer choice and $\ln(n)$ for n equally frequent rotamers, i.e., 0.7 if $n = 2$, 1.1 if $n = 3$, 1.4 if $n = 4$, and so on.

RESULTS

Toward an Optimal Protocol

Convergence

The key technical difficulty in conformational searching is locating the global minimum in a complex energy landscape. The flatter the potential function and the fewer the degrees of freedom, the easier it is to find the global optimum. As shown below, the present setup with simple 6–9-potentials and 1–16 rotamer states per residue makes the optimization problem non-trivial, yet not prohibitively complicated.

The dependence of the efficiency of the Monte Carlo algorithm on annealing parameters (δ, α_{max}) is characterized in Figures 2–4 by way of energy, rotamer entropy, and acceptance ratio. The convergence of energy and of rotamer choice between in-

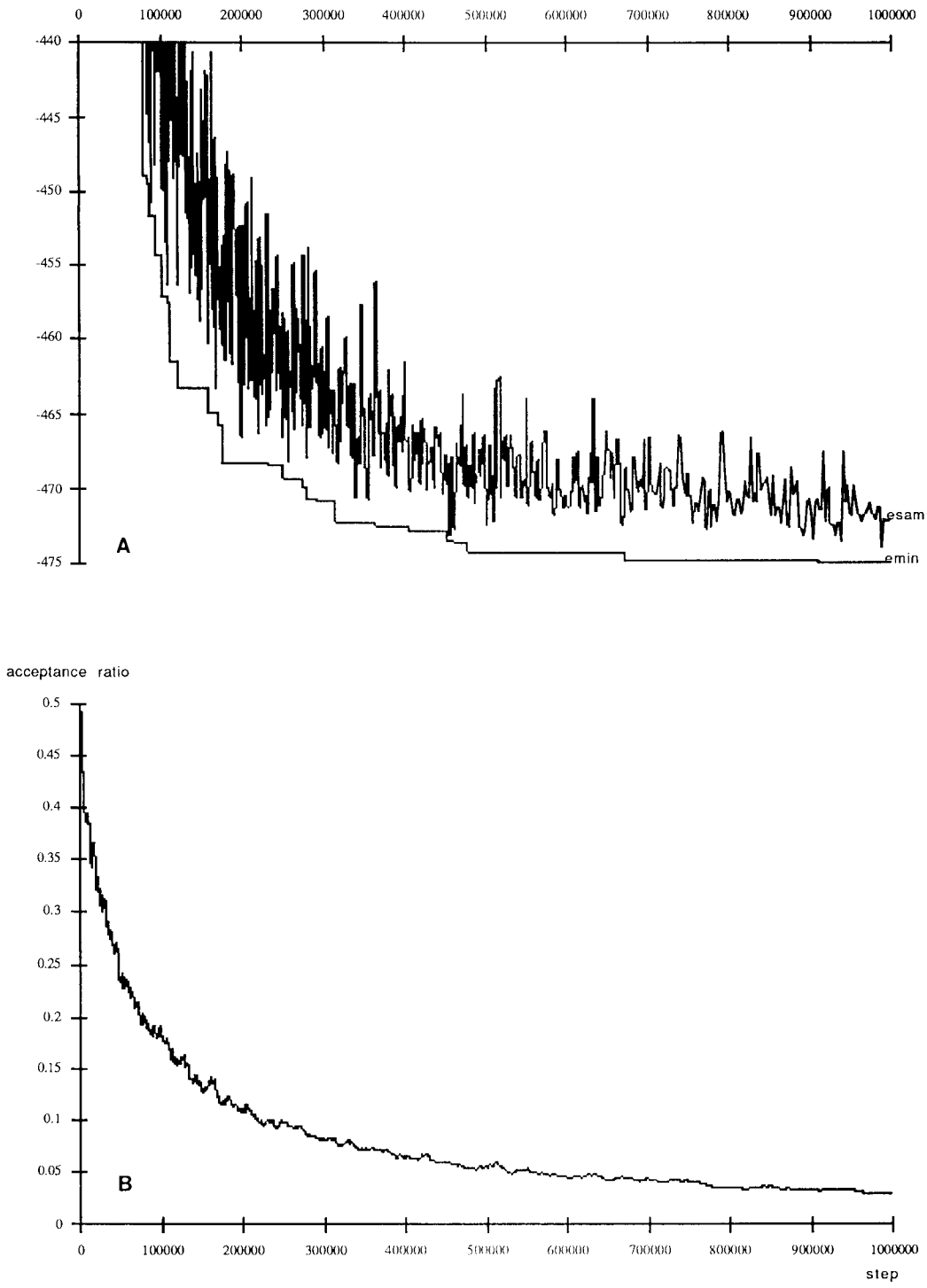

Fig. 2. Optimization of energy during a single run. Time is given in terms of the number of Monte Carlo steps, energy in arbitrary units. This figure is for flavodoxin (3FXN) using a cooling rate of $\delta = 0.00001$, so the entire run covers the temperature range $\alpha = 0.0$ to $\alpha = 10.0$. **(A)** Energy as a function of time. The energy of the initial random configuration is of the order of +2,500 energy units. With time, fluctuations decrease: the probability of large uphill steps in energy is diminished as the system is cooled down. The lower curve (emin) is the best energy found so far, and the upper curve (esam) the energy at the time of sampling (every 100 accepted steps). **(B)** Acceptance ratio averaged over 500 accepted steps as a function of time. Late in the run, the acceptance ratio is small and the rate of change in energy becomes very slow.

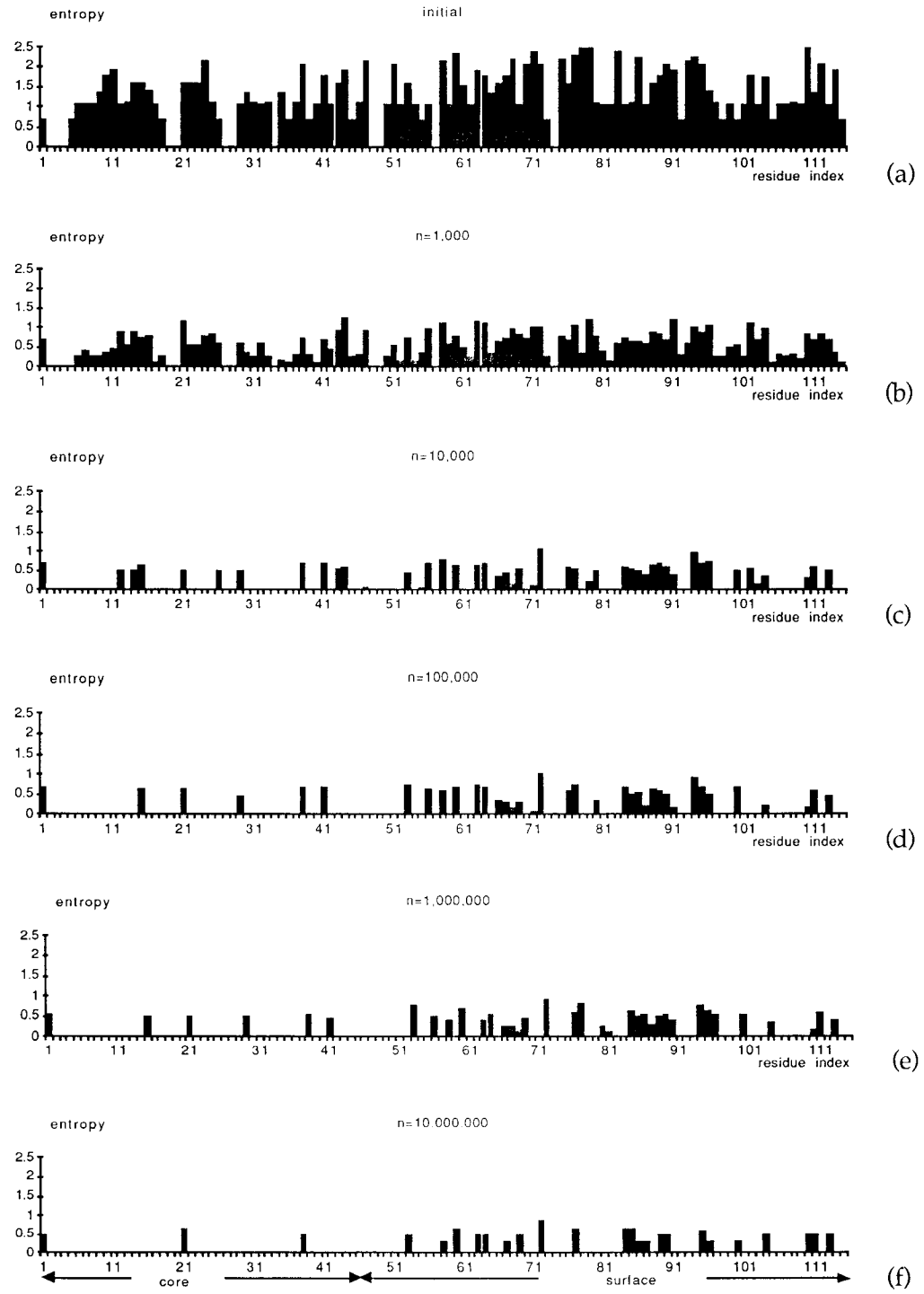

Fig. 3. Global optimum? Rotamer choices converge toward a unique arrangement of side chains as run lengths are increased and the rate of cooling is decreased. Each of the panels **(a)** to **(f)** contains rotamer entropies for flavodoxin (3FXN) residues from 100 independent Monte Carlo optimization runs. The runs start from different initial rotamer configurations and use different random number generator seeds. Residues are sorted by relative solvent accessibility (Gly, Ala, Pro excluded). Residues with index 1–46 are in the core, 47–115 are on the surface. A unique arrangement of side chains corresponds to zero entropy for all residues. **(a)** Initial configurations, after removal of rotamers which clash with the backbone. **(b)–(f)** Optimized configurations obtained with **(b)** $\delta = 0.01$, **(c)** $\delta = 0.001$, **(d)** $\delta = 0.0001$, **(e)** $\delta = 0.00001$, **(f)** $\delta = 0.000001$. Only 10 runs instead of 100 were run for (f). The computational effort is increased 10-fold between panels, from 1,000 steps per run in (b) to 10,000,000 in (f). The total entropies of the system from (a) to (f) are 148, 62, 25, 20, 18, and 13, respectively. Eleven residues have a single rotamer (i.e., zero entropy) at stage (a): C128, E59, F66, F69, F85, F99, F131, L43, Y88, Y106, W95. In **(f)** the global optimum has almost been reached for core residues: the three core residues which have not yet converged (nonzero entropy) are C53, L49, and E62.

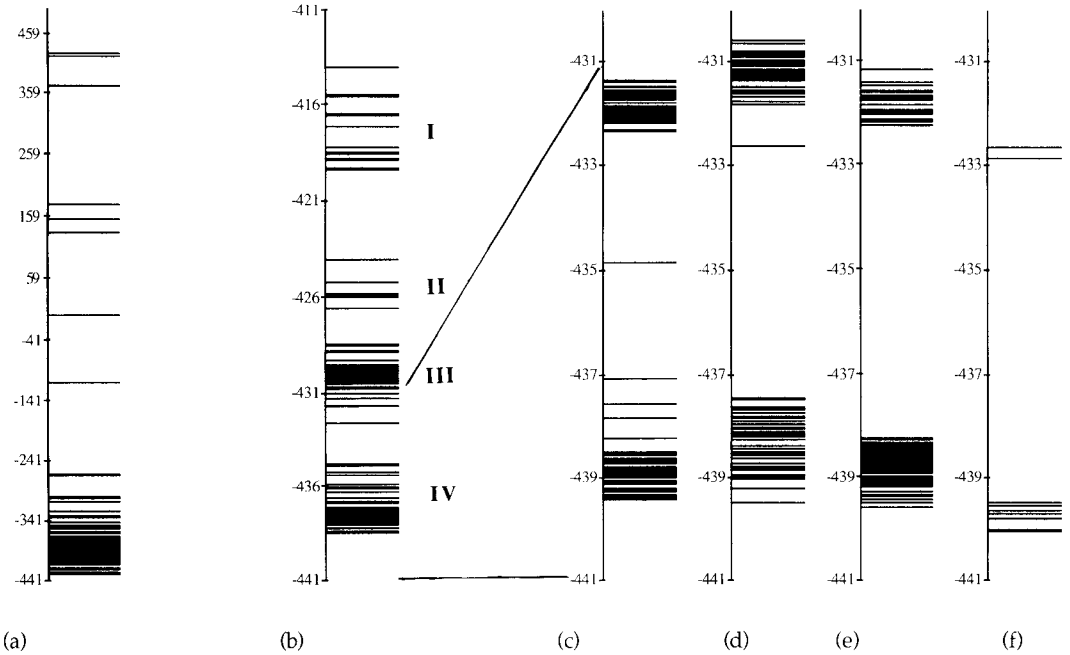

Fig. 4. Approach to the optimum: distribution of final energies in many runs. Each of the figures **(a)–(f)** represents the distribution of final energies in a set of 100 runs with a particular cooling rate δ. Between sets of runs, the cooling rate is changed: **(a)** $\delta = 0.01$; **(b)** $\delta = 0.001$; **(c)** $\delta = 0.001$ using learned frequencies collected during 100 previous runs; **(d)** $\delta = 0.0001$; **(e)** $\delta = 0.00001$; **(f)** 10 runs with $\delta = 0.000001$. Note that the energy scales are different in (a), (b), (c)–(f). With slower annealing, the range of final energies narrows and occurs at lower energies. A number of "bands" (clusters of final energies) are apparent in all sets of runs, indicative of "difficult" (for the algorithm) transitions. For example, half the runs in (d) did not reach an energy lower than about −432; all of the runs with a lower energy achieved a final energy of about −438. In (b) the transition from band I to band III requires a cooperative change in the conformation of the triplet I77, V82, C108, band II is mixed, and the transition from band III to band IV is a cooperative conformational change of the triplet C53, M56, E62. The latter two bands remain in (c)–(f).

dependent optimization runs give an indication of the probability of finding the global minimum. Provided that cooling is not too fast, the algorithm converges to configurations of similar low energy and with similar packing in the core.

Shape of the energy surface

The decrease in entropy as the energy of the system decreases shows that only a small fraction of possible side chain arrangements leads to acceptable packing (Fig. 2). Unique rotamers are chosen for the majority of core residues. Aromatic residues in particular are highly constrained by backbone clashes.

There are several configurations near the lowest energy configuration which differ by the choice of a few rotamers. Figure 4 shows samples obtained from 100 runs on flavodoxin at different cooling rates. Two energy bands are seen even at very slow annealing rates. The fine structure of the bands is obviously due to surface sidechains, which have higher rotational freedom (Fig. 3). The two major bands correspond to alternative packing arrangements of the triplet C53, E62, and M56 (one letter amino acid code, followed by residue number, Fig. 4f). All nearest neighbors of this triplet are modeled by unique rotamers. The energy difference between the bands corresponds to about 20 pairwise atom–atom interactions at optimal distance or to one clash. This is not much considering the rigid rotamers and steep potentials. The transition to the lower energy state requires a rare chain event: torsion of C53 in the core is only possible if E62 first moves outward where it has to push away M56, which is located at the surface. The longer the simulation, the larger fraction of the runs finds a way to the lower energy state. Note that the higher energy state is closer to the X-ray structure in this particular example.

Type of elementary step

A trial move in our standard protocol consists of changing the conformation of one side chain. In an attempt to incorporate cooperative transitions, the protocol was modified so that the conformation of two contacting side chains was changed at the same time. The same energy minimum was found, but with a lower acceptance ratio.

"Learned" rotamer weights

Statistics on rotamer choices in the optimized configurations from a large number of short Monte Carlo runs were collected and the "learned" rotamer frequencies were then used to bias the choice of rotamers. The optimized energies are lower compared to the short runs (Fig. 4b–c), but similar to the ener-

gies obtained in a single long run with the same total number of steps (Fig. 4e).

Acceptance ratio

The acceptance ratio falls off exponentially as α increases and the (uphill) energy difference resulting from a move matters more and more. The total acceptance ratio is about 10%, but in the final stages (α approaching $\alpha_{max} = 10.0$) it has dropped down to 2–5% (Fig. 2). Although the acceptance ratio may seem low, it is compensated for by the extremely fast calculation of energies.

Computation times

The initialization step, loading of rotamers and calculation of lookup tables, takes a few minutes for the larger proteins. Examples (protein, number of residues, CPU time for initialization): 1PPT/36/20 sec, 3FXN/138/1.5 min, 2APP/323/3 min. Monte Carlo optimization takes 0.2 msec CPU time per step, independent of protein size.

Standard protocol

How much effort would be required for convergence to a single rotamer configuration, i.e., the global minimum? The rotamer entropies of the test series in Figure 3 extrapolate to zero between $\delta = 10^{-7}$ and 10^{-8}, requiring 5–50 hr of CPU time per run. Our standard protocol consists of a single Monte Carlo run with $\delta = 0.00001$ and $\alpha_{max} = 10.0$, chosen for reasonable reproducibility and swift execution time (200 sec).

Prediction of Side Chain Conformations

The key physical problem is to what extent the calculated potential energy of a model correlates with correctness of structure. Rigid rotamers and simple steric potential energy are a surprisingly good description of side chain packing: in the final models on average 72% of all residues and 81% of core residues have correctly predicted χ_1 angles in proteins of better than 2.5 Å crystallographic resolution (Table II). Side chain–backbone clashes are eliminated already at the initialization stage. The removal of side chain–side chain clashes by energetic optimization drives the rotamer configurations even closer to the native packing although the relationship between energy and correctness of a model is not strictly monotonic (Fig. 5).

The large variation in prediction success is partly explained by the sharper distribution of sidechain χ angles in highly refined proteins compared to low-resolution structures.[14] For example, the structure of human lysozyme (1LZ1, 1.5 Å resolution, 130 residues) is well refined and has only two lysine side chains in a conformation which is not present in the rotamer library; for this structure 88% of χ_1 angles and 76% of both χ_1 and χ_2 angles are predicted within 30°. At the other extreme, in the structure of ribonuclease-S (1RNS, 2.0 Å resolution, 124 residues) 22 out of the 102 side chains clearly visible in the electron density map (see comments in the PDB file) have a conformation more than 30° away from the main staggered rotamer conformations; here, the accuracy for χ_1 drops to 47%, that of both χ_1 and χ_2 to 45%.

As the rotamer conformations are built by varying the χ angles, the orientation of the side chain is sensitive to the accuracy in the position of the C_β atom. This can be fatal in a model building situation in which the backbone coordinates are inaccurate. For example, the prediction (rebuilding) of cytochrome b_{562} using the X-ray backbone improved drastically as the crystallographic resolution increases from 2.5 to 1.4 Å (data sets 156B and 256B, Table II). In another test of the influence of backbone accuracy on side chain conformation, backbone coordinates reconstructed from C_α traces using the program MaxSprout[11] had a positional error of about 0.5 Å for backbone atoms. As a consequence, the level of correctly predicted χ_1 angles decreased by 12 percentage points relative to reconstruction using the original X-ray backbones (Table II).

Figure 6 shows a comparison of the optimized and X-ray model of flavodoxin. The main constraints are felt by larger side chains in the hydrophobic core. The conformations of aromatic side chains are particularly well predicted. Typical errors include missing hydrogen bonds, misoriented charged and hydrophobic side chains on the border between core and surface due to the lack of a solvation term, and missing ion pairs due to the lack of an electrostatic term in the energy function.

Relaxation

Energy minimization in continuous Cartesian coordinates using GROMOS[21] after side chain optimization had very little effect on the quality (rmsd, % correct χ_1) of optimized models. This is expected as energy minimization is incapable of leaving local minima and the rotamer states are separated by high potential energy barriers.

One way of getting around inaccuracies in the backbone is to extend the region of conformational space which is being sampled during Monte Carlo optimization, i.e., to increase the number of degrees of freedom. To test this, $\pm 15°$ wobbles of either the χ_1 or χ_2 angle of the main rotamers were added to the rotamer library, except for residue types K, R, E, Q, and M, which have already 10 or more rotamers in the standard library. The expanded library consisted of 316 rotamers. The optimized energies were naturally lower, compared to those generated using the standard library. The average improvement in quality was on average three percentage points for correct χ_1 angles in the core with X-ray backbones. The improvement was only half as big with C_α-derived backbones. The drawbacks of increasing rota-

TABLE II. Quality of Side Chain Prediction*

			X-ray backbone				Backbone from C_α trace				
			Side chain rms[Å]		% correct χ_1		Side chain rms[Å]		% correct χ_1		
PDB†	Nres	Resol [Å]	Total	Core	Total	Core	Total	Core	Total	Core	Protein
5PTI	58	1.0	1.9	1.0	78	100	2.5	2.9	59	62	Pancreatic trypsin inhibitor
7RSA	124	1.3	1.8	1.2	79	84	2.2	1.4	65	70	Ribonuclease A
4PTP	223	1.3	1.9	1.4	72	82	2.0	1.6	65	76	Bovine trypsin
1PPT	36	1.4	1.4	1.3	91	100	2.4	1.4	70	100	Avian pancreatic polypeptide
256B	106	1.4	2.0	0.9	73	92	2.5	2.3	58	64	Cytochrome b_{562}
1LZ1	130	1.5	1.6	1.2	88	96	2.1	1.6	70	78	Lysozyme
2PRK	279	1.5	2.0	1.7	73	78	2.0	1.8	64	69	Proteinase K
1GCR	174	1.6	1.7	1.5	72	81	2.2	1.6	66	75	γ-Crystallin
1PCY	99	1.6	1.6	1.4	63	76	2.3	2.5	56	72	Plastocyanin
3TLN	316	1.6	1.7	1.4	77	86	2.2	2.0	66	74	Thermolysin
1CTF	74	1.7	1.7	0.3	81	100	2.3	1.9	55	67	L7/L12 ribosomal protein
1PSG	370	1.7	2.7	2.0	75	81	3.0	2.5	63	66	Pepsinogen
2WRP	107	1.7	2.2	0.9	64	73	2.2	1.0	60	73	trp repressor
1UBQ	76	1.8	1.7	0.6	71	87	2.1	1.2	60	78	Ubiquitin
2APP	323	1.8	1.4	1.4	81	85	1.8	1.9	67	73	Penicillopepsin
3APR	325	1.8	1.4	1.1	84	87	2.0	1.9	67	70	Endothiapepsin
3FXN	138	1.9	1.9	1.8	61	65	1.9	1.7	52	52	Flavodoxin
1HMQ	113	2.0	2.0	1.6	62	76	2.3	2.3	59	71	Hemerythrin
1RNS	124	2.0	1.9	2.0	47	54	2.6	1.6	40	46	Ribonuclease S
2LYZ	129	2.0	1.7	1.3	72	84	2.1	1.7	63	70	Lysozyme
7WGA	171	2.0	1.7	1.9	80	78	2.2	2.3	62	59	Wheat germ hemagglutinin
4APE	330	2.1	1.6	1.5	65	79	2.1	2.3	54	61	Rhizopuspepsin
3ADK	195	2.1	2.0	1.3	51	67	2.2	1.6	43	55	Adenylate kinase
1PFK	320	2.4	2.0	2.0	75	77	2.2	1.7	67	74	Phosphofructokinase
2CRO	71	2.4	2.3	1.7	57	67	2.6	1.6	57	61	λ cro repressor
1PMB	153	2.5	1.9	1.6	72	82	2.2	2.2	55	60	Myoglobin
1RHD	293	2.5	2.3	2.2	51	55	2.3	2.5	47	49	Rhodanase
1TIM	147	2.5	2.3	2.0	56	63	2.3	2.1	50	54	Triose phosphate isomerase
156B	103	2.5	2.5	2.3	32	55	2.4	2.1	32	55	Cytochrome b_{562}
1CTS	437	2.7	1.9	1.6	64	69	2.3	2.1	54	60	Citrate synthase
3HVP	99	2.8	2.0	1.4	63	63	2.3	1.9	57	44	HIV protease
2TRM	223	2.8	1.6	1.4	66	73	2.1	1.8	53	61	Rat trypsin
1PYP	285	3.0	2.2	2.2	39	31	2.5	2.5	39	33	Pyrophosphatase
Mean‡	176	1.7	1.8	1.4	72	81	2.2	1.8	60	69	
Mean	184	2.0	1.9	1.5	68	77	2.3	1.9	57	65	
Min	36	1.0	1.4	0.3	32	31	1.8	1.0	32	33	
Max	437	3.0	2.7	2.3	91	100	3.0	2.9	70	100	

*Two series of side chain construction and optimization: (1) using the correct experimental backbone and (2) using model backbone coordinates as reconstructed from the experimental C_α trace using the program MaxSprout,[11] with an average rms deviation of C_α positions of 0.2 Å relative to the experimental structure (0.5 Å for backbone atoms). Note that the quality of side chain prediction is better when the experimental backbone is used and that it is better in the core than on the solvent exposed surface.
†Full references to the crystallographic work can be found in the Protein Data Bank headers. They are omitted here for space reasons.
‡Mean, average for proteins determined at better than 2.5 Å resolution; mean, average for all proteins; averages are calculated giving unit weight to each protein; Nres, number of residues; resol, nominal crystallographic resolution; total, all side chains, core plus surface.

mer choice are increased CPU time and increased memory requirements since the number of possible rotamer–rotamer interactions grows roughly as the square of the number of rotamers per residue.

Another way to compensate for inaccuracies in backbone coordinates was tried by modifying the potential function to reflect flexibility of side chain conformation. This was done by replacing rotamer–rotamer energies by the minimum over a set of conformations around the standard rotamers. However, no improvement in correctness of the lowest energy protein conformation resulted.

Model Building by Homology

The accuracy of models built by homology depends on sequence similarity. For example, the strongly homologous pair bovine and rat trypsin has 74% sequence identity, no gaps in the alignment, and sim-

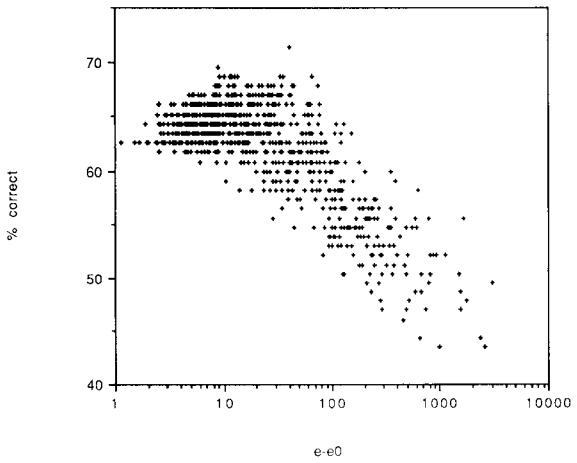

Fig. 5. Energy optimization drives side chains toward the native packing. At different stages during a single run on flavodoxin (3FXN), the percentage of correct χ_1 angles (excluding Gly, Ala, and Pro from the statistics) was sampled and plotted against a logarithmic scale of the energy difference relative to the "global" minimum (−475 energy units, i.e., the lowest energy found during this particular Monte Carlo run). The cooling rate was $\delta = 0.00001$, and every 500th accepted step was sampled. The "global" minimum configuration (outside the sample) had 63% correct χ_1 angles. Higher quality configurations (up to 71%) were reached during the run, but were not retained because of higher energy. The native packing (100% correct χ_1 angles) was not reached, but on the average lower energy structures tend to be closer to the native one.

ilar 3-D structures with C_α atoms within 0.5 Å (rms) of each other after optimal superposition. For this pair the result of sidechain construction of one protein based on the coordinates of the other (Table III) is comparable to that based on its own correct X-ray backbone (see Table II). In a second set of examples side chains were built to rhizopuspepsin (4APE), endothiapepsin (3APR), and pepsinogen (1PSG) models constructed using the backbone of penicillopepsin (3APP). These alignments include gaps and the sequence identities are 51, 36, and 29%, respectively. A framework consisting of 249 residues of 3APP was selected, which deviated from the original C_α traces by 1.4, 1.6, and 1.8 Å (rmsd), respectively. The side chain predictions are more successful the stronger the sequence similarity (Table IV). The backbone of the homologous 3APP structure is an excellent starting point for the models of 4APE and 3APR. The model of 1PSG, which is the least similar to 3APP of the three, suffers from the backbone drift and has a less accurate model. The orientation of aromatic side chains in the core of the homology-derived models is particularly well predicted (Table IV). Even in the 1PSG model three-fourths of aromatic core residues are correctly predicted if conserved side chains are fixed during optimization.

DISCUSSION

Overall Assessment of the Method

We have described here a fast, simple, fully automatic, reproducible, and general algorithm for optimizing the arrangement of side chains on given backbone models. The present optimized algorithm has improved in speed and accuracy relative to our previous implementation.[11]

The algorithm is made extremely efficient by the use of precalculated interaction energy tables which is possible in a system with discrete conformational states. The packing of side chains is optimized in a matter of minutes on a workstation. The quality of the models is surprisingly good considering the simplicity of the physical model. The best results are obtained with high-resolution X-ray backbones (average 81% correct in the core). Inaccuracies in the main chain up to 0.5 Å are tolerated with a reduction in the success rate of only around 10 percentage points. Aromatic residues, which dominate the visual impression of models in comparison to the real structure, are remarkably well predicted even when the backbone coordinates are taken from a weakly homologous protein.

Related Methods

Tuffery et al.[13] use a genetic algorithm with discrete rotamer states and fixed backbones. Lee and Subbiah[12] obtain results of similar quality to ours using Monte Carlo optimization (simulated annealing) on fixed backbones where the elementary step consists of changing all side chain χ angles in 10° increments, at much greater expense of computer time. Correa[22] describes a single example of generating full coordinates from the C_α trace in stepwise fashion using molecular dynamics refinement, at even greater expense of computer time, with an overall result comparable to MaxSprout.[11] For model building by homology, our results compare well to methods which employ energy minimization in continuous space.[8,9] We conclude that the restriction to the subspace of rotamers is a very reasonable approximation.

Limitations

A main limitation of the present algorithm comes from fixing the backbone. The quality of side chain construction decreases if the backbone contains inaccuracies. This is the case if the backbone is taken from a low-resolution X-ray structure or in model building by homology if the total sequence identity is around 30% or lower (five examples tested). A second limitation is in the lack of solvation in the energy function so that few constraints apply to exposed side chains and the lack of proper electrostatics. A third limitation is that the algorithm is not guaranteed to escape from local minima in finite time. However, in long runs (Figs. 3,4) the total en-

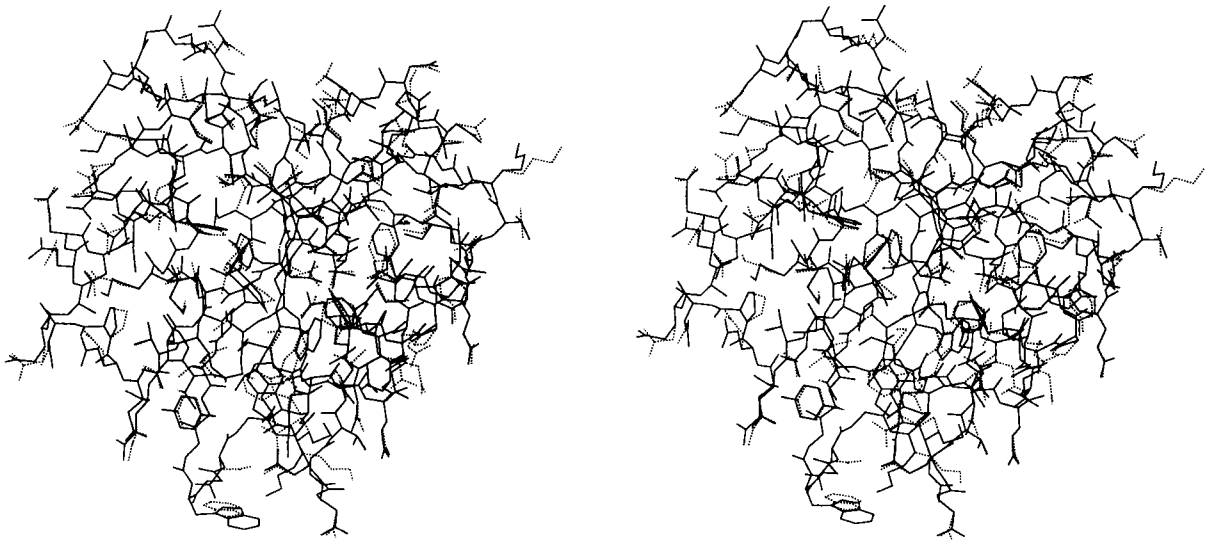

Fig. 6. Predicted and crystal structure side chains overlap remarkably well. Shown here is flavodoxin (3FXN), predicted side chain positions in continuous lines and X-ray structure positions in dashed lines. On the protein surface, prediction and crystal structure differ, but in the core they are very similar. The main constraints are felt by larger side chains in the hydrophobic core. The percentage of correctly predicted sidechains (χ_1 angles) and rms position deviation of sidechain atoms are 65% and 1.75 Å in the core, and 61% and 1.9 Å overall. Side chain packing in most proteins tested was predicted even better than this example (see Table II). Note that the prediction makes use of the correct backbone atom (C_α, C, N, O, C_β) positions of the crystal structure. Plotted using What If.[25]

TABLE III. Accuracy of Model Building by Homology: Rat and Bovine Trypsin*

			Side chain rms [Å]		% correct χ_1	
Model	Template	Fix	Total	Core	Total	Core
2 TRM	4PTP	No	2.0	1.3	55	66
2TRM	4PTP	Yes	1.4	1.0	70	84
4PTP	2TRM	No	1.9	1.5	66	76
4PTP	2TRM	Yes	1.5	1.0	73	85

*Better accuracy is reflected in lower side chain rms (positional deviation) or higher percentage of correct χ_1 angles. In each case, full atomic coordinates of the model protein were built using the backbone of the template. The model was then compared to the correct crystal structure. For example, 2trm was built using the backbone of 4ptp and the model 2trm was then compared with the experimental 2trm. The "fix" column indicates whether conserved side chains were copied from the template (yes) or optimized along with the other side chains (no). Other notation, see Table II. 2TRM, rat trypsin; 4PTP, bovine trypsin.

TABLE IV. Accuracy of Model Building by Homology: Pepsins*

Model	Number of loops	Fix	% correct χ_1 Total	Core	% correct χ^1 aromatic residues in core
4APE	6	No	59	74	86
4APE		Yes	63	80	97 (66)
3APR	13	No	63	62	82
3APR		Yes	68	69	93 (57)
1PSG	14	No	49	54	56
1PSG		Yes	54	56	76 (44)

*Notation as in Table III. Penicillopepsin (3APP) was used as template. The numbers in parentheses show how much is gained trivially by fixing conserved sidechains. 4APE, rhizopuspepsin; 3APR, endothiapepsin; 1PSG, pepsinogen.

ergy as a function of side chain conformation in the core appears to have almost fully converged.

Ongoing and Potential Applications

The algorithm described here is currently being applied in our laboratory to generate models of the cores of the GTP-binding domains of *ras*-related proteins (Valencia and Sander, unpublished; reviewed in reference 23).

The database of homology derived protein structures (HSSP, ver. 0.9) by Sander and Schneider[24] aligns 3,512 sequences to structures in the Protein Data Bank (PDB[15]). The implied 3-D models that could be built based on the alignments in this database should be particularly reliable for the 1,690 sequences which are more than 50% identical to a cousin in the PDB. The PDB currently holds about 150 nonredundant protein chains if 50% sequence identity is used as a cutoff (for alignments of 80 residues or longer).[19] It was estimated that between three and four hundred of the aligned sequences in HSSP are less than 50% identical to other sequences within this set or with any protein in the PDB.[24] Therefore, modeling these new sequences would more than double the effective size of the PDB. The price paid, of course, is limited accuracy of model coordinates, especially in loop regions. As it takes on average 3 to 4 min of CPU time to construct a model with our method (with unoptimized loops), the task would be completed in one day.

CONCLUSION

The protein folding problem is still unsolved, but drawing on the available databases and using the kind of method described here, we already have powerful tools available to build fair 3-D models from sequence in favorable cases. There remain two major open questions. First, we need more reliable energy functions, especially for solvation and electrostatics, in order to more finely discriminate between correct and incorrect conformations. Second, we need optimization algorithms that include relaxation of backbone coordinates and are efficient enough to allow full exploration of the many-dimensional conformational space of a protein chain in finite time.

ACKNOWLEDGMENTS

We thank the crystallography groups who have made their 3-D structures available through the Protein Data Bank, Michael Scharf for discussion, and Gerrit Vriend for help with WHAT IF. L.H. acknowledges an EMBO fellowship.

REFERENCES

1. Pearl, L.H., Taylor, W.R. A structural model for the retroviral proteases. Nature (London) 329:351–354, 1987.
2. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5: 823–826, 1986.
3. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5(4): 819–822, 1986.
4. Moult, J., James, M.N.G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1:146–163, 1986.
5. Bruccoleri, R.E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
6. Bruccoleri, R.E., Karplus, M.A. Conformational sampling using high-temperature molecular dynamics. Biopolymers 29:1847–1862, 1990.
7. Havel, T.F., Snow, M.E. A new method for building p conformations from sequence alignments with homologues of known structure. J. Mol. Biol. 217:1–7, 1990.
8. Summers, N.L., Karplus, M. Construction of side-chains in homology modelling. J. Mol. Biol. 210:785–811, 1989.
9. Schiffer, C.A., Caldwell, J.W., Kollman, P.A., Stroud, R.M. Prediction of homologous protein structures based on conformational searches and energetics. Proteins 8:30–43, 1990.
10. Snow, M.E., Amzel, L.M. Calculating three-dimensional changes in protein structure due to amino acid substitutions: The variable region of immunoglobulins. Proteins 1:267–279, 1986.
11. Holm, L., Sander, C. Database algorithm for building full backbone and sidechain coordinates from a protein Cα trace. Application to model building and detection of coordinate errors. J. Mol. Biol. 218:183–194, 1991.
12. Lee, C., Subbiah, S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217:373–388, 1991.
13. Tuffery, P., Etchebest, C., Hazout, S. Lavery, R. A new approach to the rapid determination of protein side chain conformations. J. Biomol. Str. Dynam. 8:1267–1289, 1991.
14. Ponder, J. W., Richards F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
15. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
16. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A. and Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087, 1953.
17. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. Optimization by simulated annealing. Science 220:671–680, 1983.
18. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
19. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Science. (in press), 1992.
20. Baumann, G., Frömmel, C., Sander, C. Polarity as a criterion in protein design. Prot. Eng. 2:329–334, 1989.
21. van Gunsteren, W.F., Berendsen, H.J.C. GROMOS, Groningen Molecular Simulation Computer Program Package, University of Groningen, The Netherlands, 1987.
22. Correa, P.E. The building of protein structures from α-carbon coordinates. Proteins 7:366–377, 1990.
23. Valencia, A, Chardin P, Wittinghofer F, Sander C. The ras protein family: Evolutionary tree and role of conserved amino acids. Biochemistry 30:4637–4648, 1991.
24. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
25. Vriend, G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8:52–56, 1990.