A stochastic algorithm for global optimization and for best populations: A test case of side chains in proteins

Meir Glick, Anwar Rayan, and Amiram Goldblum†

Department of Medicinal Chemistry and the David R. Bloom Center for Pharmacy, School of Pharmacy, Hebrew University of Jerusalem, Jerusalem 91120, Israel

Edited by Alan Fersht, University of Cambridge, Cambridge, United Kingdom, and approved November 30, 2001 (received for review August 8, 2001)

The problem of global optimization is pivotal in a variety of scientific fields. Here, we present a robust stochastic search method that is able to find the global minimum for a given cost function, as well as, in most cases, any number of best solutions for very large combinatorial ''explosive'' systems. The algorithm iteratively eliminates variable values that contribute consistently to the highest end of a cost function's spectrum of values for the full system. Values that have not been eliminated are retained for a full, exhaustive search, allowing the creation of an ordered population of best solutions, which includes the global minimum. We demonstrate the ability of the algorithm to explore the conformational space of side chains in eight proteins, with 54 to 263 residues, to reproduce a population of their low energy conformations. The 1,000 lowest energy solutions are identical in the stochastic (with two different seed numbers) and full, exhaustive searches for six of eight proteins. The others retain the lowest 141 and 213 (of 1,000) conformations, depending on the seed number, and the maximal difference between stochastic and exhaustive is only about 0.15 Kcal/mol. The energy gap between the lowest and highest of the 1,000 low-energy conformers in eight proteins is between 0.55 and 3.64 Kcal/mol. This algorithm offers real opportunities for solving problems of high complexity in structural biology and in other fields of science and technology.

Many problems in life sciences and in other fields of science and technology are of high complexity, thus requiring sophisticated methods of searching and scoring to achieve the ability to study and to simulate them by means of a computer simulation. An excellent search method coupled with a highly reliable scoring method should allow comparisons to some natural phenomena. In this article, we have taken the approach of comparing best populations found by a stochastic search method to a full, exhaustive search, as the crucial test of this method. However, comparisons to experimental results also are included. The problem chosen to exemplify this method is the positions of side chains in proteins, which is essential for both theoretical and experimental purposes. On the theoretical side, it is a subproblem in *de novo* protein structure prediction. It is essential for structure-based drug design (1), for inverse folding and threading algorithms (2), for predicting the effect of mutations on structure (3), for *ab initio* predictions of tertiary structure (4), for homology-based modeling (5), and others. From the x-ray crystallographer's point of view, it could speed the placement of side chains using the electron density maps of the main chain before refinement calculations. The main limitation is the large amount of possible conformations that each side chain may adopt (6). An exhaustive search of all possible conformations is beyond the scope of state of the art computers.

Current strategies for side chain addition to a given backbone differ in three categories. The first is the conformational space of each side chain. In continuous space methods (7, 8), any side chain torsion angle may be sampled. Discrete space methods are based on the assumption that side chains exist in energetically preferred conformations called rotamers, which are local minima conformers that have been sampled by statistical analysis of known structures (9–14). Discrete space methods cannot predict conformations that are not present in the rotamer database. There is no agreement regarding the optimal size of a rotamer library. Several groups showed that large rotamer databases that contain very rare conformations do not necessarily yield better predictions than smaller databases (15–18). On the other hand, Xiang & Honig (19) have recently extended the accuracy of predictions with an extensive rotamer library. Rotamer databases also can be classified as backbone dependent and backbone independent. The former are based on a relationship between the side chain conformation and the local backbone conformation (20–21), whereas the latter are not (7, 16, 22).

The second category is the scoring or cost function for evaluating solutions. Energy-based methods rely on nonbonding terms (6, 15, 16, 18, 23–25). The assumption is that the lower the energy, the more accurate the prediction. Knowledge-based methods also were proposed: Sutcliffe *et al.* (26) suggested a procedure for building side chains using spatial information from side chains in topologically equivalent positions—as far as such a correlation may be observed—and most probable conformations of the side chains in the respective secondary structure type. Sali & Blundell (27) described a comparative protein-modeling method designed to find the most probable structure for a sequence, given its alignment with related structures. Bower *et al.* (28) located residues in their most favorable backbone-dependent rotamers and systematically resolved the conflicts that arise from that structure.

Accurate computer location of protein side chains is a complicated task because of the large number of minimum energy conformers on the potential energy surface, even with a rigid backbone. Conventional methods for side chain addition usually result in a single structure of the protein, which is then compared with an experimental structure, if available. The conformational space is disregarded, although protein function and molecular recognition depend on structural plasticity (29), and conformational flexibility of receptor proteins is considered to be one of the major factors that affect ligand docking (30).

Our algorithm focuses on the third category, the search strategy, and not on the energy function or the rotamer library. There are numerous examples of search strategies for highly complex problems. Metropolis Monte Carlo methods (15), Gibbs sampling Monte Carlo (18), Neural networks (31), Conformational Space Annealing (32), Genetic and Evolutionary Algorithms (33–35), Simulated Annealing (6), Mean Field Optimization (23), and Locally Enhanced Sampling (8). Combinatorial Searches (21, 24, 33) are used on discrete conformers and may be followed by a continuous minimization in the final stage of refinement. It should be noted that there is no guarantee that any of the above will converge to a valid solution. Another widely used method is Dead End Elimination (DEE). It is based on the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: DEE, Dead End Elimination; MD, molecular dynamics.

†To whom reprint requests should be addressed. E-mail: amiram@vms.huji.ac.il.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

identification of rotamers that are absolutely incompatible with the global minimum energy conformation, eliminating rotamers that cannot contribute to local energy minima of a certain or higher order. Conformations composing such rotamers can be qualified as dead ending (30, 36, 37). If enough rotamers can be eliminated by recursive applications, the global minimum can be found (38, 39). If no conditions can be established to eliminate further rotamers during the calculation, DEE might not converge. The global minimum can be found by an exhaustive search of the remaining rotamers (38), provided the remaining search space is not prohibitively large. Because of the mechanism of elimination of rotamers in DEE, there is little chance of forming an optimal population of solutions.

However, a combination of DEE with the A* algorithm (40) has been suggested for constructing a population of low-energy side chain conformations in proteins, and was used for constructing partition functions. The A* algorithm approach may find the best N solutions, but it is restricted to relatively small proteins. The largest protein solved by this algorithm so far contained 68 amino acids, which comprise about 10^{43} combinations—depending on the complexity of the rotamer library–whereas proteins with a much larger number of combinations are common. As a "stand alone" algorithm (without the DEE preprocessing stage) the A* algorithm reaches a maximum of 10^{21} combinations. An effective search by the A* algorithm must have a good estimate of the cost to reach a goal node. Estimation is problematic because of interactions between residues that have not yet been assigned. Those limitations raise the need for a robust algorithm that finds the global minimum and the lowest energy conformations in larger systems. Such a search algorithm is presented here.

Methods

The Search Technique. The code uses a backbone-dependent rotamer library (13, 21, 28, 41). We used the August 1997 update of the rotamer library of Dunbrack & Karplus, with united atoms (42). Energy is computed by Eq. **1** with the AMBER nonbonding 12–6 Lennard–Jones and electrostatic energy terms (43), where $A_{i,j}$ is the repulsion parameter for the two (i,j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, $r_{i,j}$ is the distance between atoms, and ε is the dielectric constant. A distance-dependent dielectric constant of $\varepsilon = r$ has been used. The nonbonded energy is calculated for interactions with the backbone and with other residues' rotamers. If the nonbonded energy term exceeds the value of 10 Kcal/mol for a given pair of atoms, it is truncated at 10 Kcal/mol.

$$E_{\text{pot}} = \sum_{i<j}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\varepsilon^* r_{i,j}}\right) + \sum_{\text{side-chains}} -\ln\left(\frac{P_{\text{rotamer}}}{P_0}\right) \quad [1]$$

As suggested by Bower *et al.* (28) and implemented in the SCWRL (Side Chains With a Rotamer Library) algorithm, every rotamer is given a local energy based on its probability in the backbone-dependent rotamer library. Energies are taken from the probabilities of the backbone-dependent rotamer library, as $-\ln(p_{\text{rotamer}}/p_0)$, where p_0 is the probability of the most probable rotamer, and p_{rotamer} is the probability of a given rotamer (assuming $kT = 1$). The search strategy includes several steps:

(I) *Steric clashes elimination stage and preliminary rotamer location.* The input for the calculation are the backbone (N, C_α, C, O) coordinates of a protein with a known, high-resolution structure. Those, together with standard bond lengths and bond angles from AMBER 4.1 (43) and with φ and ψ angles of the backbone are used to create the initial placement of possible rotamers for each residue. Possible disulfide bonds between cysteine residues are calculated by the distance between sulfur atoms. All rotamers that clash with the backbone are excluded by a threshold value of 18 Kcal/mol. If all rotamers of a residue clash with the backbone, the rotamer with the lowest "clash energy" remains. The algorithm treats single rotamers as part of the backbone; i.e., other rotamers that clash with those residues also will be excluded. The algorithm also searches for all side chain clashes between rotamer i of amino acid j and rotamer k of amino acid l. The algorithm excludes such pairs from being part of the solution, and, therefore, they are not sampled in the stochastic stage (see below).

(II) *Stochastic stage.* It is obvious that in the case of a large biological system such as a protein, a very large combinatorial problem results. In Hydrolase (1arb; ref. 44), for example, there are 2.4×10^{105} alternative positioning options after step I. A stochastic algorithm is used to reduce the size of the problem. In the protein, the side chain rotamers in d_0 amino acids are unknown. For each amino acid there is usually more than one rotamer, but only one would give the lowest energy. Let $X_j = (x_{j1}, x_{j2} \ldots x_{jd0})$ be a conformation of the protein that includes randomly picked rotamers for d_0 amino acids. For each conformation X_j, the energy $E_j = E(X_j)$ may be calculated according to the energy function described above. The objective is to find the conformation that minimizes E. Because it is impossible to evaluate all of the alternative conformations because of the large number of combinations, the following steps are taken: (*i*) Sample at random n conformations of the large population of combinations $X_1 = (x_{11}, x_{12}, \ldots x_{1d0}), \ldots, X_n = (x_{n1}, x_{n2}, \ldots, x_{nd0})$, where x_{11} is a randomly picked rotamer for the first amino acid in the first conformation, and x_{n1} is a randomly picked rotamer for the same amino acid in the n^{th} conformation. We use $n = 1{,}000$ to create a large enough number of protein conformations and compute the corresponding energy values: $E_1 = E(X_1)$ to $E_n = E(X_n)$.

(*ii*) Construct the distribution $F_E^n (n = 10^3)$. F_E^n is the set of energies of all of the N-sampled conformations for the full protein. Define cutoff points H and L in F_E^n. H contains all variable values satisfying $E_i \geq F_E^n(1 - \alpha)$, where $F_E^n(\alpha)$ is the $\alpha-$th percentile of F_E^n, and L contains all variable values satisfying $E_i \leq F_E^n(\alpha)$. The number of conformations in each of H and L is $n_0 = n \times \alpha$. When $n = 1{,}000$ conformations and $\alpha = 0.01$ (1%) for highest and lowest energy conformations, $n_0 = \alpha \times n = 0.01 \times 1{,}000 = 10$, so $L = 10$ and $H = 10$. In other words, H stands for the 10 highest energy conformations, and L stands for the 10 conformations with the lowest energy. (*iii*) Construct the vector h for all rotamer variables corresponding to the conformations in H. The vector h is the element-wise intersection of all of the rotameric states in H, in the following manner: if all rotameric states in H share the same rotamer at component j (corresponding to x_{nj} of conformation X_n), then $h_j =$ rotamer_number; otherwise, $h_j = 0$ (no common rotamer for j in all high-energy conformations.) (*iv*) Construct the vector l for rotamer variables corresponding to the conformations in L. Unlike vector h, more than one rotamer may appear for each amino acid j up to a maximum of n_0 values in l_j. It is the union of all rotamers of component j that appear in the low-energy conformations of L. (*v*) Compare h and l. If both h_j and l_j have a similar rotamer, it will remain as a viable rotameric state, because it contributes also to low-energy values. However, if h_j does not correspond to any element of l_j, then the corresponding rotamer h_j will be evicted from subsequent iterations. If an amino acid has only one rotamer, it will not be evicted from subsequent iterations because it is the only remaining solution. (*vi*) Repeat steps *i* to *iv* for the reduced set of variables' values until the number of possible combinations of all variables is smaller than a user-defined "end of stochastic stage criteria".

The value of α that is used to determine n_0 should be selected with care. If n_0 is too large, no rotamers will be eliminated. If n_0 is too small, an unjustified elimination of rotamers might occur. At best, n_0 should be adjusted by the number of possible

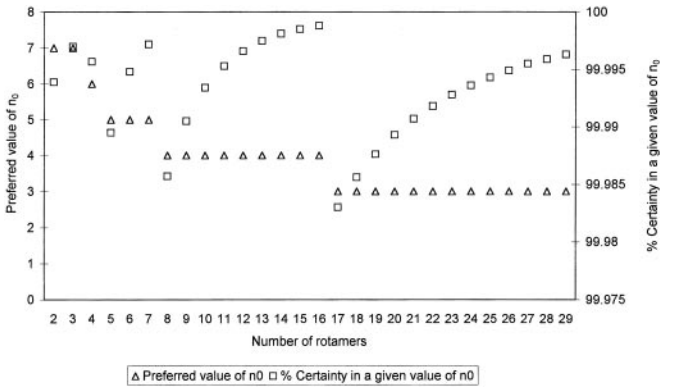

Fig. 1. Values of n_0 for 2 to 29 possible rotamers of a single residue that lead to elimination with high certainty. Each number of rotamers has an associated value of n_0 (△). The larger the number of rotamers, the smaller is n_0. For each given number of rotamers and n_0, the percentage of certainty is calculated (□).

rotamers of each amino acid, to allow an equal probability for the elimination of rotamers. To explain the determination of α, let us assume that each rotamer is not affected by interactions with any other amino acid in its environment. The n_0 values for 2 to 29 possible rotamers of a single residue that would lead to the correct rotamer elimination with a certainty >99.983% are presented (Fig. 1). Those values were calculated in the following manner. Given a residue with three rotamers, if we want to remove one rotamer with a certainty ($P_{correct}$) higher than 99.99%, the error probability (P_{error}) must be smaller than 0.01% (0.0001). For erroneously evicting a rotamer, it must first appear in all of the high-energy conformations. In this case, the probability is $(1/3)^{n0}$. In addition, this rotamer must not appear in any low-energy conformation. In this case the probability is $(2/3)^{n0}$. The total error probability is $P_{error} = (1/3)^{n0}(2/3)^{n0}$. Thus, one may tune the calculation to nearly 100% confidence by employing the general formula in Eq. **2,** where m is the number of variable values (rotamers).

$$P_{\text{error}} = \left(\frac{1}{m}\right)^{n0}\left(\frac{m-1}{m}\right)^{n0} \qquad [2]$$

When $m = 1$ (there is one rotamer) $P_{error} = 0$. Assigning a value of $P_{error} = 0.0001$ and solving the equation for $m = 3$ leads to a value of $n_0 = 6.12$. When n_0 is very large, $P_{error} = 0$, but the odds of evicting any variable value are very low. Thus, we employ the n_0 values from Fig. 1, which allow eviction of variable values, with $P_{correct} = 99.983–99.9988\%$.

(III) End of search. Once there are less than M combinations remaining ($M \approx 10^5$), an exhaustive search is conducted to yield the N lowest energy conformers of the protein.

Results

A Test of the Search Method's Validity. To test the accuracy and efficiency of our method, we impose our stochastic algorithm to find the lowest energy combinations—given the constraints of the energy function and the rotamer library—and compare them to the results of an exhaustive search. We applied the stochastic algorithm to eight high-quality x-ray structures (resolution < 1.5 Å, R factor < 0.17) of proteins taken from the Protein Data Bank (45) with various sizes (54 to 263 residues) that were chosen to cover a range of protein-fold families as shown (Table 1). These proteins are: rubredoxin (5rxn) (46), ovomucoid third domain (2ovo) (47), erabutoxin B (3ebx) (48), ribosomal protein (1ctf) (49), ribosomal protein (1whi) (50), lysozyme (2ihl) (unpublished work), endonuclease (2end) (51) and hydrolase (1arb) (44). We limited the number of rotamers each residue could adopt by employing the most probable rotamers from the SCWRL backbone-dependent rotamer library (28), so that the exhaustive, full search calculation may end in a reasonable computer (CPU) time.

Two stochastic searches were conducted for each test protein. A seed number of 100,000 was used for the first search and was replaced by 8,242,117 in the second. In Fig. 2 *A* and *B*, we compare the energies resulting from the two stochastic searches to the exhaustive one for the 1,000 low-energy conformations. When employing a seed number of 100,000, the low-energy conformations were identical by the stochastic and the exhaustive searches in all of the proteins except 1ctf. In 1ctf (Fig. 2*A*), the first 213 solutions were identical, and the 1,000th solution differed by 0.08 Kcal/mol. In the second stochastic search with a seed number of 8,242,117, the low energy conformations were identical in the stochastic and the exhaustive searches in all of the proteins except 2ovo (Fig. 2*B*), where the first 141 solutions were the same and the 1,000th solution differed by 0.15 Kcal/mol.

The Search Method's Efficiency. By applying our algorithm for rotamer prediction, computing time grows linearly and not exponentially, with an increase in the number of residues. The algorithm was applied to the eight proteins with an initial number of rotamer combinations that range from 1.06×10^{25} to 2.4×10^{105} as shown (Table 1). The ln(number of combinations)

Table 1. Systems selected for comparison

Name	PDB code	Size	Number of combinations*	Number of combinations for comparing exhaustive and stochastic*	Average RMSD† for 1,000 lowest energy conformers, Å	Energy gap† in Kcal/mol between the 1,000th conformer and the global minimum	Residues‡ with different χ_1 among lowest 1,000 energy conformers, %
Rubredoxin	5rxn	54	3.90×10^{27}	1.26×10^{9}	2.20	1.64	14.6
Ovomucoid third domain	2ovo	56	1.06×10^{25}	8.49×10^{7}	2.03	3.64	16.7
Erabutoxin B	3ebx	62	1.50×10^{31}	6.37×10^{8}	2.50	1.35	12.3
Ribosomal protein	1ctf	68	3.23×10^{34}	3.58×10^{8}	2.33	3.28	6.4
Ribosomal protein	1whi	122	4.97×10^{73}	8.49×10^{7}	2.48	3.33	5.9
Lysozyme	2ihl	129	2.17×10^{61}	5.66×10^{7}	2.26	1.98	5.7
Endonuclease	2end	137	1.31×10^{82}	2.01×10^{9}	2.68	3.03	5.9
Hydrolase	1arb	263	2.40×10^{105}	1.61×10^{9}	2.24	0.55	3.0

*After backbone clashes are relieved.
†For a calculation with all the rotamers.
‡Except Gly and Ala.

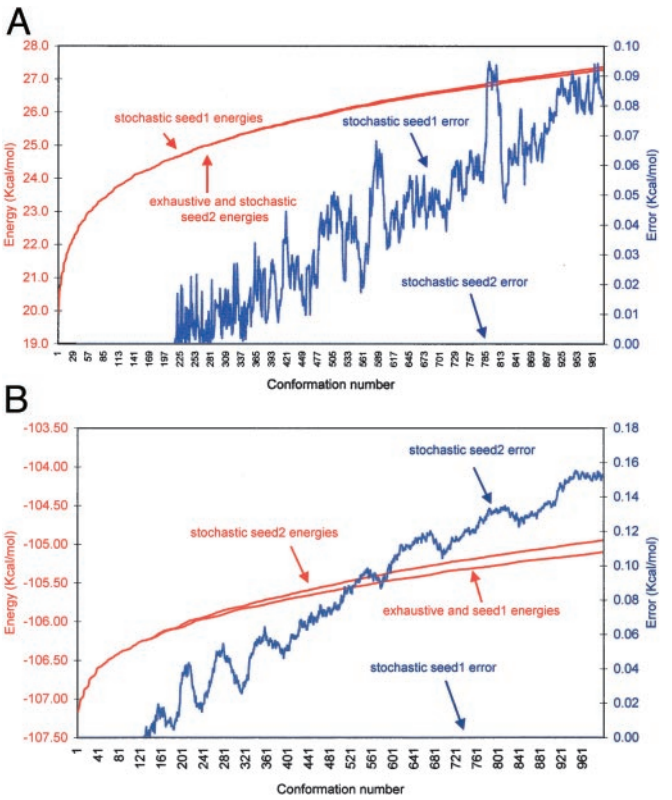

Fig. 2. Comparison of stochastic searches when employing two different seed numbers to an exhaustive search on two test proteins. Lowest energy conformations (1,000) are presented. Error is calculated as the energy difference between the given conformation in the stochastic and exhaustive searches. (*A*) Ribosomal protein (1ctf). (*B*) Ovomucoid third domain (2ovo).

vs. the number of iterations is depicted (Fig. 3). The number of iterations to convergence ranged between 516 for 2ovo (1.06×10^{25} combinations) to 4,441 for 1arb (2.4×10^{105}). The ratio between the combinations for these two proteins is 2.26×10^{80}, whereas the ratio between the iterations was 8.6. The 129 residues of 2ihl required 1,894 iterations to end the stochastic stage, whereas the 263 residues of 1arb needed 4,440 iterations. The number of starting combinations for these two was 2.17×10^{61} vs. 2.40×10^{105}, respectively.

Comparison of the Algorithm to Experimental Results. Results (Table 1) are given for the average RMSD (root mean square deviation) of the lowest energy populations of 1,000 conformations for each

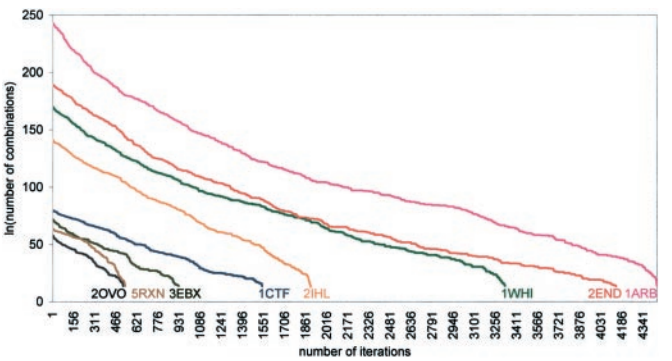

Fig. 3. ln(number of combinations) vs. the number of iterations for eight proteins.

protein. The values are between 2.03 to 2.68. The lowest energy conformations (data not shown) did not have, in most proteins, the lowest RMSD to the x-ray structure. The energy gap between the lowest and highest energies among the low-energy populations ranges between 0.55 for 1arb to 3.64 for 2ovo. Among the thousand results, we find (last column in Table 1) that a relatively small number of residues (expressed as percentage) deviate from the crucial χ_1, which is the angle closest to the backbone and affects most strongly the conformation of the side chain. We evaluated the number of side chains in each protein that adopted multiple positions by calculating the percentage of residues (except Ala and Gly) with different χ_1 among the 1,000 lowest energy conformations found by our search method. An angle χ_1 that deviates by 30° or more from the rest was considered as different. For example, rubredoxin contains (Table 1, line a) 48 residues that are neither Ala nor Gly. We found that seven residues exhibited χ_1 deviations, thus 14.6%.

Discussion

We present a stochastic search technique and an example of its possible applications, exploring a given conformational space of proteins' side chains. The algorithm successfully explores the conformational space of various sizes of proteins and can deal with a large number of combinations after eliminating rotamers that clash with the backbone. The robustness of the stochastic algorithm in handling complex combinatorial searches is clearly demonstrated (Fig. 2 *A* and *B* and Fig. 3). Comparing it to an exhaustive search proves the reliability of the stochastic algorithm in reproducing most of the population of lowest energy conformations. In all proteins, the global minimum has been consistently detected. The 1,000 low-energy conformations were identical in the stochastic and the exhaustive searches in 14 of 16 comparisons, whereas 2 cases had a smaller set of lowest energy conformations that were identical in the two searches. Even in these cases, with 141/1,000 and 213/1,000 identical lowest conformations, the prevailing contributors to the molecular partition function are included, and may subsequently be used to estimate the conformational entropy. Indeed, Leach & Lemon (40) used low-energy rotamer combinations to evaluate the partition function and, thus, calculated the side chain contribution to the conformational entropy of the folded protein. One must bear in mind that both our method and Leach & Lemon's method are conducted in discrete space. A numeric comparison to entropy values obtained from continuous searches may give further insight into the reliability of a discrete search. Full conformational freedom of the backbone is required to extract real entropy values for proteins.

Table 1 presents the energy gaps between the 1,000th conformer and the global minimum of each protein. These energy differences indicate that the rotamers have a considerable degree of conformational flexibility that varies between the different proteins: it is 0.55 Kcal/mol for 1arb and 3.64 Kcal/mol for 2ovo. The energy gap variations between proteins may reflect the relative flexibilities of their side chains and should be studied further in connection with other indices of flexibility. Also, the lack of relations between protein size and energy gap warrants further examination.

The algorithm presented here belongs to the class of heuristic solutions. One of the tools used to assess the quality of our results is changing the seed number. Like other stochastic heuristic methods, our algorithm is not immune to such an effect. Nevertheless, we demonstrated that the algorithm found the global minimum in all of the proteins, when employing different seed numbers. Thus, the global minimum has been retained and not evicted in any of the large number of iterations for each of the proteins. In these test cases we have demonstrated that by combining two different seed numbers we succeeded in finding all of the required low-energy populations. Also, it should be

noted that no accidental eviction of values is possible: each such eviction is a result of a systematic test. Those values that are not evicted remain for the final exhaustive step, in which all their combinations are evaluated. Thus, each one of the total of initial values must be probed and either evicted or retained for the final full search.

The hub of this work is a search methodology and neither a rotamer library nor a cost function. Most deviations from maximal accuracy may be caused by the size limitation of the rotamer library and the deficiencies of the energy function. Indeed, the current rotamer library's best possible RMSD for the tested proteins (found by positioning the rotamer that is closest to the x-ray structure) is between 0.94 Å to 1.52 Å. A way to overcome the search-space limitation was suggested by Mendes *et al.* (52); it presents a rotamer as a continuous ensemble of conformations that cluster around the classic rigid rotamer. A different approach to expanding the search space was recently devised by Honig and coworkers (19), which achieved accurate predictions by using an extensive rotamer library containing over 7,560 members, in which bond lengths and bond angles were taken from the database rather than simply assuming idealized values. Further, the performance of CHARMM (53) was better than that of AMBER in that work. The limitations of the force field are noticeable mostly in the fact that, in most proteins, the lowest energy conformations did not have the lowest RMSD from the x-ray structure.

Currently, there are four main methods to study the conformational space of a given protein: x-ray crystallography, NMR, molecular dynamics (MD), and rotamer library-based methods. Experimental information of biomolecular structure and conformations has its own limitations. X-ray crystallography usually supplies a single structure which reflects the biomolecule in the highly ordered crystal lattice, as opposed to the more physiologically relevant solution environment of an NMR structure. The former might be biased toward specific conformational substates in the crystal, which may not be among the ensemble of conformations in solution (54). Observation of alternate rotamers is beyond the detection limits of conventional x-ray crystallographic techniques, except at the very highest resolution. At least 10% of all side chains in proteins adopt multiple, discrete conformations in carefully refined crystal structures (55). MacArthur & Thornton (56) found a significant and unexpected correlation between χ_1 mean values and resolution, mainly for small flexible side chains. All of the data support the hypothesis that this observation reflects local conformational flexibility and disorder, which at low resolution might be interpreted as a single distorted conformer.

We used the algorithm to explore the side chain conformational space of *E. coli* ribonuclease HI (57) and compared the results to experimental and theoretical methods that offer an insight into the multiple conformations that each side chain may adopt under different conditions: x-ray crystallography, NMR, and MD. Our algorithm found 82% of the multiple side chain conformers in this case (data not shown). The advantage of our algorithm is straightforward: it extends the single conformation into a population of viable conformations.

Unlike x-ray crystallography, NMR suggests alternative conformations by deciphering the two-dimensional and three-dimensional coupling maps (57, 58). NMR does not teach us about the shape of the energy minima on the potential energy surface. NMR of proteins is a long and tedious experiment limited by the time scale of conformational variations, especially in large proteins. In this case, our algorithm may be an additional tool for suggesting alternative conformations. When NMR structures are available, our algorithm may be used to extend this information by allowing the determination of the conformations' energy weights, thus enabling an assessment of their contribution to the overall population at equilibrium.

Classical MD simulations suggest conformations that may not be detected by NMR or by x-ray crystallography. With current technology, MD simulations of systems consisting of tens of thousands of atoms for a few nanoseconds are becoming more common (59). However, relevant time scales for biomolecular functions range from nanoseconds to more than seconds. The time required to reach an equilibrium between different conformers of a protein by MD is prohibitive for such simulations, and we may acquire only a glimpse of the protein's behavior in its surrounding. As a result, the ability of MD to detect the global minimum or the population of lowest-energy conformations in large biomolecules is limited. The reliability of our stochastic algorithm in finding both has been demonstrated in this article. Whereas MD trajectories imply a mechanism of conformational interconversions, our stochastic approach, like Monte Carlo, concentrates on products and not pathways, because of the employment of discrete values and its nondeterministic nature.

Dill and Chan (60, 61) suggested that the native state of a given protein corresponds to the global minimum in free energy, which is not necessarily the computed global minimum potential energy, even with a reliable function. The missing entropy evaluation may be contributed partially by our algorithm, as it yields most of the low-energy conformers. Our search offers, in addition to finding the global minimum, the next N best solutions for rotamers in large proteins without any mean field approximation and is unique in that sense. Thus, it may be used for studying thermodynamic properties of complex molecular systems. The stochastic algorithm can treat more than 250 residues (the maximum at this stage has been 2.29×10^{105} combinations, with no optimization of the CPU time), which is more than any algorithm known to us that is able to generate side chain populations and not single minima. Another advantage is in its ability to form populations by employing the stochastic algorithm in a stand-alone mode without any preprocessing algorithm (such as DEE, in the case of the A* algorithm). Also, one should note that the numbers of combinations presented (Table 1) for the stochastic algorithm refer to possible numbers of combinations that remain after evicting rotamers that clash with the backbone. Hence, the real number of possible combinations is much higher. This algorithm can be applied to other issues (62) of complex optimization.

It may be possible to simplify the combinatorial nature of the side chain problem and reduce it to pairwise (36, 37, 38) or to self-consistency (19) methods. However, such approaches cannot produce an accurate or close approximation to the ensemble of structures, the "best population" that may be crucial for the physical and biological characteristics of a protein. Our method, however, transcends the side chain issue that was used here as a test case. We regard our comparison of these heuristic search results to full exhaustive results as the most significant test of this method's performance and suggest it as a yardstick for future comparisons of methodologies in this field and others.

Our approach for finding low-energy minima of a complex biomolecular system is not necessarily limited to the life sciences. After adjusting the number of sampled solutions in each iteration (n) and cutoff points H and L in F_E^n to the specific nature and complexity of the problem (i.e., the number of variable values, which is the number of rotamers in this example), this strategy may be used in other problems as long as the search space is discrete and a reliable or reasonable cost function may be used. This algorithm thus may evolve to be useful for other fields such as telecommunications (to design efficient networks), transportation, and economics.

We thank Dr. Andrew Leach from GlaxoSmithKline for his prompt response and advice. This project was supported in part by the Israel Ministry of Trade and Industry, in the framework of the Daat ("knowledge") consortium (Magnet project) and by a grant of the Israel Science

Foundation established by the Israel Academy of Sciences and Humanities. Equipment was supplied by the Alex Grass Center for Drug Design and Synthesis of Novel Therapeutics at the School of Pharmacy, Hebrew University of Jerusalem.

1. Defay, T. & Cohen, F. E. (1995) *Proteins Struct. Funct. Genet.* **23,** 431–445.
2. Bahar, I. & Jernigan, R. (1997) *J. Mol. Biol.* **266,** 195–214.
3. Wong, K. B., DeDecker, B. S., Freund, S. M. V., Proctor, M. R., Bycroft, M. & Fersht, A. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 8438–8442.
4. Huang, E. S., Koehl, P., Levitt, M., Pappu, R. V. & Ponder, J. W. (1998) *Proteins Struct. Funct. Genet.* **33,** 204–217.
5. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987) *Nature (London)* **326,** 347–352.
6. Lee, C. & Subbiah, S. (1991) *J. Mol. Biol.* **217,** 373–388.
7. Eisenmenger, F., Argos, P. & Abagyan, R. (1993) *J. Mol. Biol.* **231,** 849–860.
8. Roitberg, A. & Elber, R. (1991) *J. Chem. Phys.* **95,** 9277–9287.
9. Chandrasekaran, R. & Ramachandran, G. N. (1970) *Int. J. Protein Res.* **2,** 223–233.
10. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000) *Proteins Struct. Funct. Genet.* **40,** 389–408.
11. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
12. Gelin, B. R. & Karplus, M. (1979) *Biochemistry* **18,** 1256–1268.
13. Dunbrack, R. L., Jr., & Karplus, M. (1994) *Nat. Struct. Biol.* **1,** 334–340.
14. Cheng, B., Nayeem, A. & Scheraga, H. A. (1996) *J. Comput. Chem.* **17,** 1453–1480.
15. Holm, L. & Sander, C. (1992) *Proteins Struct. Funct. Genet.* **14,** 213–223.
16. Laughton, C. A. (1994) *J. Mol. Biol.* **235,** 1088–1097.
17. Tanimura, R., Kidera, A. & Nakamura, H. (1994) *Protein Sci.* **3,** 2358–2365.
18. Vasquez, M. (1995) *Biopolymers* **36,** 53–70.
19. Xiang, Z. & Honig, B. (2001) *J. Mol. Biol.* **311,** 421–430.
20. McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **198,** 295–310.
21. Dunbrack, R. L., Jr., & Karplus, M. (1993) *J. Mol. Biol.* **230,** 543–574.
22. Levitt, M. (1992) *J. Mol. Biol.* **226,** 507–533.
23. Koehl, P. & Delarue, M. (1994) *J. Mol. Biol.* **239,** 249–275.
24. Wilson, C., Gregoret, L. M. & Agard, D. A. (1993) *J. Mol. Biol.* **229,** 996–1006.
25. Vasquez, M. (1996) *Curr. Opin. Struct. Biol.* **6,** 217–221.
26. Sutcliffe, M. J., Hayes, F. R. & Blundell, T. L. (1987) *Protein Eng.* **1,** 385–392.
27. Sali, A. & Blundell, T. L. (1993) *J. Mol. Biol.* **234,** 779–815.
28. Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997) *J. Mol. Biol.* **267,** 1268–1282.
29. Garcia, K. C., Degano, M., Pease, L. R., Huang, M., Peterson, P. A., Teyton, L. & Wilson, I. A. (1998) *Science* **279,** 1166–1172.
30. Desmet, J., Wilson, I. A., Joniau, M., De Maeyer, M. & Lasters, I. (1997) *FASEB J.* **11,** 164–172.
31. Hwang, J. K. & Liao, W. F. (1995) *Protein Eng.* **8,** 363–370.
32. Pillardy, A., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D. R., Kazmierkiewicz, R., Oldziej, S., Wedemeyer, W. J., Gibson, K. D., Arnautova, Y. A., *et al*. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2329–2333. (First Published February 20, 2001; 10.1073/pnas.041609598)
33. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991) *J. Biomol. Struct. Dyn.* **8,** 1267–1289.
34. Bowie, J. U. & Eisenberg, D. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4436–4440.
35. Forrest, S. (1993) *Science* **261,** 872–878.
36. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992) *Nature (London)* **356,** 539–542.
37. Lasters, I. & Desmet, J. (1993) *Protein Eng.* **6,** 717–722.
38. Looger, L. L. & Hellinga H. W. (2001) *J. Mol. Biol.* **307,** 429–445.
39. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000) *J. Mol. Biol.* **299,** 789–803.
40. Leach, A. R. & Lemon, A. P. (1998) *Proteins Struct. Funct. Genet.* **33,** 227–239.
41. Dunbrack, R. L., Jr., & Cohen, F. E. (1997) *Protein Sci.* **6,** 1661–1681.
42. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Jr. & Weiner, P. (1984) *J. Amer. Chem. Soc.* **106,** 765–784.
43. Pearlman D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., III, Ferguson, D. M., Seibel, G. L., Singh, U. C., Weiner, P. K. & Kollman, P. A. (1995) AMBER (Univ. of California, San Francisco), VERSION 4.1.
44. Tsunasawa, S., Masaki, T., Hirose, M., Soejima, M. & Sakiyama, F. (1989) *J. Biol. Chem.* **264,** 3832–3839.
45. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
46. Watenpaugh, K. D., Sieker, L. C. & Jensen L. H. (1980) *J. Mol. Biol.* **138,** 615–633.
47. Empie, M. W. & Laskowski, M., Jr. (1982) *Biochemistry* **21,** 2274–2284.
48. Smith, J. L., Corfield, P. W., Hendrickson, W. A. & Low, B. W. (1988) *Acta Crystallogr. A* **44,** 357–368.
49. Leijonmarck, M. & Liljas, A. (1987) *J. Mol. Biol.* **195,** 555–579.
50. Davies, C., White, S. W. & Ramakrishnan, V. (1996) *Structure (London)* **4,** 55–66.
51. Morikawa, K., Matsumoto, O., Tsujimoto, M., Katayanagi, K., Ariyoshi, M., Doi, T., Ikehara, M., Inaoka, T. & Ohtsuka, E. (1992) *Science* **256,** 523–526.
52. Mendes, J., Baptista, A. M., Carrondo, M. A. & Soares, C. M. (1999) *Proteins Struct. Funct. Genet.* **37,** 530–543.
53. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swamirathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
54. Brunger, A. T. (1997) *Nat. Struct. Biol.* **4** Suppl., 862–865.
55. Smith, J. L., Hendrickson, W. A., Honzatko, R. B. & Sheriff, S. (1986) *Biochemistry* **25,** 5018–5027.
56. MacArthur, M. W. & Thornton, J. M. (1999) *Acta Crystallogr. D* **55,** 994–1004.
57. Philippopoulos, M. & Lim, C. (1999) *Proteins Struct. Funct. Genet.* **36,** 87–110.
58. Ellgaard, L., Riek, R., Herrman, T., Guntert, P., Braun, D., Helenius, A. & Wutrich, K. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3133–3138. (First Published March 6, 2001; 10.1073/pnas.051630098)
59. Sagui, C. & Darden, T. A. (1999) *Annu. Rev. Biophys. Biomol. Struct.* **28,** 155–179.
60. Chan, H. S. & Dill, K. A. (1998) *Proteins Struct. Funct. Genet.* **30,** 2–33.
61. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
62. Wales, D. J. & Scheraga, H. A. (1999) *Science* **285,** 1368–1372.