Protein Structure Prediction Using a Combination of Sequence Homology and Global Energy Minimization I. Global Energy Minimization of Surface Loops

Michael J. Dudek and Harold A. Scheraga
Baker Laboratory of Chemistry, Cornell University
Ithaca, New York 14853-1301, U.S.A.

This is a preprint of a published article: *J. Comp. Chem.*, **11**, 121–151 (1990).

Abstract

A procedure has been developed for global energy minimization of surface loops of proteins in the presence of a fixed core. The ECEPP potential function has been modified to allow more accurate representations of hydrogen bond interactions and intrinsic torsional energies. A computationally efficient representation of hydration free energy has been introduced. A local minimization procedure has been developed that uses a cutoff distance, minimization with respect to subsets of degrees of freedom, analytical second derivatives, and distance constraints between rigid segments to achieve efficiency in applications to surface loops. Efficient procedures have been developed for deforming segments of the initial backbone structure and for removing overlaps. Global energy minimization of a surface loop is accomplished by generating a sequence (or a trajectory) of local minima, the component steps of which are generated by searching collections of local minima obtained by deforming seven-residue segments of the surface loop. The search at each component step consists of the following calculations: 1) A large collection of backbone structures is generated by deforming a seven-residue segment of the initial backbone structure. 2) A collection of *low-energy* backbone structures is generated by applying local energy minimization to the resulting collection of backbone structures (interactions involving side chains that will be searched in this component step are not included in the energy). 3) One low-energy side chain structure is generated for each of the resulting low-energy backbone structures. 4) A collection of low-energy local minima is generated by applying local energy minimization to the resulting collection of structures. 5) The local minimum with the lowest energy is retained as the next point of the trajectory. Applications of our global search procedure to surface segments of bovine pancreatic trypsin inhibitor (BPTI) and bovine trypsin suggest that component-step searches are reasonably complete. The computational efficiency of component-step searches is such that trajectories consisting of about ten component steps are feasible using an FPS-5200 array processor. Our procedure for global energy minimization of surface loops is being used to identify and correct problems with the potential function and to calculate protein structure using a combination of sequence homology and global energy minimization.

1 INTRODUCTION

This paper is the first in a series devoted to protein structure prediction using a combination of sequence homology and global energy minimization. In this paper, we describe a computer program for global energy minimization of surface loops. The program should enable systematic improvement of the accuracy of the potential function and applications of global energy minimization to structure prediction that would otherwise have been impossible. In paper II of this series,[(1)] we will describe an attempt to improve the accuracy of the potential function. In paper III of this series,[(2)] we will describe an attempt to calculate the structure of a protein for which the crystal structure has been determined using both the crystal structure of a homologous protein and global energy minimization.

The basic concepts of protein structure prediction using a combination of sequence homology and global energy minimization are a) the structure of the core can be determined using the crystal structure of a homologous protein, and b) the structure of the surface can (in principle) be determined using global energy minimization. In practice, obtaining structural information from global energy minimization is difficult both because the potential energy function must be accurate and because large collections of local minima must be examined.[3] These difficulties have limited the success of several recently implemented procedures for global energy minimization of surface loops.[4–8]

The primary goal of this project is to develop methods for calculating the structure of a protein using both the crystal structure of a homologous protein and global energy minimization. The most difficult component of this goal is to develop methods for obtaining a significant amount of structural information from an empirically parameterized function representing energy. There are probably alternative procedures for obtaining reliable structural information about surface loops. We have not attempted to use experimental information from sources other than homology, and we have not attempted to push the concept of homology to give structural information about surface loops.[9,10]

Calculation of protein structure using the crystal structure of a homologous protein and global energy minimization is closely related to calculation of protein structure using NMR distance constraints and global energy minimization. Knowing the crystal structure of a homologous protein is equivalent to knowing distance constraints for some subset of atom pairs. In both cases, methods for imposing the experimentally determined distance constraints are considerably less difficult than methods for obtaining structural information from an empirically parameterized function representing energy. Therefore, methods developed for the former application are useful for the latter.

The description of METHODS is composed of several sections. The primary result of our research is the procedure for global energy minimization of surface loops, which is described in section 2.5. To test the efficiency of this procedure, global energy minimization has been applied to surface segments of protein crystal structures, as described in section 2.6. A precise measure of the efficiency of a procedure for global energy minimization would be the amount of computer time that is required to obtain the global minimum of a specified potential function. However, since the completeness of our global searches cannot be established conclusively, we are forced to use a less precise measure of efficiency, the amount of computer time that is required to accomplish global searches for which available evidence suggests completeness. The functions that were used to represent the vacuum potential energy and the hydration free energy are described in sections 2.1 and 2.2, respectively. Our procedure for local minimization with respect to subsets of degrees of freedom, which is described in section 2.4, is the primary component of our procedure for global energy minimization of surface loops. The hardware on which our computer program has been developed is described in section 2.3. In section 2.7, notation is introduced that allows concise description of backbone structure. This notation is used to describe the low-energy structures that result from applications of global energy minimization to surface segments of protein crystal structures.

Our evaluation of efficiency is independent of the accuracy of the potential function. The accuracy of the potential function is relevant to an evaluation of efficiency only if the potential function is sufficiently accurate that the global energy minimum can be predicted with confidence to be the experimentally determined crystal structure. Such accuracy would enable the global energy minima of surface loops to be determined experimentally, which would enable the completeness of global searches to be established with minimal effort. Since the global minimum of the potential function that was used to evaluate efficiency often differs significantly from the crystal structure, none of our conclusions concerning efficiency is based on the assumption of an accurate potential function. We do not attempt to examine the accuracy of the potential function that was used to evaluate efficiency. The potential function is evolving, and the accuracy of the current potential function is not sufficient to allow reliable prediction of structure.

For reasons that are discussed in section 2.4, procedures for generating localized deformations of large molecules and procedures for efficient local minimization of these deformed structures with respect to subsets of degrees of freedom are the primary components of what appear to be the most promising procedures for global minimization of large molecules. Our procedure for global energy minimization

of surface loops is a procedure of this type. In Appendix A, we describe an algorithm for generating backbone deformations. This algorithm is computationally more efficient and can be implemented with much less effort than previously described algorithms for generating backbone deformations.[11] In Appendix B, we describe a simple yet extremely effective algorithm for removing overlaps from starting conformations. This algorithm significantly reduces the amount of computation that is needed for local minimization with respect to subsets of degrees of freedom.

Our computer program was originally developed to allow global energy minimization of the surface loops of human thrombin. However, our attempt to predict the structure of thrombin is not yet complete.

2 METHODS

2.1 Representation of Vacuum Potential Energy

The protein is modeled using classical mechanics. All hydrogen atoms are represented explicitly as classical particles. The values of bond lengths and bond angles are held fixed. The potential energy function is a modification of the ECEPP potential energy function.[12,13] Modifications include an altered representation of hydrogen bond interactions, an altered representation of the intrinsic torsional energy, and an altered proline geometry.[14]

Fixing the values of bond lengths and bond angles decreases the number of degrees of freedom by a factor of about eight without significantly increasing the number of local minima. This decrease in the number of degrees of freedom allows more efficient energy minimization procedures to be applied.

The computationally intensive part of the ECEPP potential function is the sum of pairwise interaction energies which, when expressed in units of kcal/mol, has the form

$$\sum_{(a,b)\in \mathrm{E}} \left(\epsilon_{(T_a,T_b)} \left\{ n \left(\frac{\rho_{(T_a,T_b)}}{r_{(a,b)}} \right)^{12} - 2 \left(\frac{\rho_{(T_a,T_b)}}{r_{(a,b)}} \right)^{2e} \right\} + \frac{332.0 q_a q_b}{D r_{(a,b)}} \right) \tag{1}$$

where E is the set of atom pairs for which interaction energies are calculated; $r_{(a,b)}$ is the distance of atom pair (a,b); $e=3$ and 5 for non-hydrogen bonding and hydrogen bonding atom pairs, respectively; $n=0.5$ and 1.0 for non-hydrogen bonding, 1–4 atom pairs and for all other atom pairs, respectively; T_a is the atom type of atom a; $\epsilon_{(T_a,T_b)}$ and $\rho_{(T_a,T_b)}$ are parameters that control the depth and the position of the minimum, respectively, for the 12-2e interaction between a pair of atoms having atom types T_a and T_b; q_a is the partial charge of atom a; and D is the dielectric constant.

Some of the modifications to the ECEPP potential function were suggested by analysis of the low-energy structures obtained by applying our global search procedure to surface loops of thrombin using the unmodified ECEPP potential function. These low-energy structures were often inconsistent with structural properties [such as compactness, linearity of hydrogen bonds, and the distribution of (ϕ,ψ) values for each amino acid] observed in crystal structures of globular proteins.[15] Largely because of the unknown accuracy of the hydration free energy contribution to the total potential function, the exact causes of the observed problems have been difficult to identify. However, careful analysis of the potential function has identified possible causes.

The altered representation of hydrogen bond interactions results in a greater energetic preference for linear geometry. The altered representation of the intrinsic torsional energy, which includes a component that depends on the value of (ϕ,ψ), results in a predicted distribution of (ϕ,ψ) values that is more consistent with the distribution that is observed in crystal structures of globular proteins. For a residue preceding proline, the altered geometry of proline results in an energy difference between the α-helical and extended regions of the (ϕ,ψ) map that is more consistent with the corresponding distribution of (ϕ,ψ) values observed in crystal structures of globular proteins. The exact form of the potential function and the parameters that were used for the calculations reported in this paper are presented elsewhere.[14]

2.2 Representation of Hydration Free Energy

Conceptually, the simplest procedures for calculating the difference in hydration free energy between two conformations of a protein are those that represent water molecules explicitly as collections of classical particles.[16] For a system consisting of water molecules in an external field created by a fixed conformation of the protein, these procedures calculate the ratio of the partition function for this system evaluated for two different conformations of the protein, a quantity that is directly related to the difference in hydration free energy between the two conformations. We have not attempted to represent water molecules explicitly because of the computational time required by these procedures. Global searches of surface loops often require tens of thousands of minimizations, and it is desirable that hydration free energy be included at each step of each minimization. Therefore, we have chosen to represent hydration free energy as an empirically parameterized function (defined on the space of protein conformations) that can be evaluated without the need for integration with respect to solvent degrees of freedom.

The function that we use to represent hydration free energy is less complex computationally than functions that determine areas of exposed surfaces[17] or volumes of unoccupied hydration shells.[18,19] The computer time that is required to calculate the function and derivatives of our hydration free energy is about the same as the computer time that is required to calculate the function and derivatives of our vacuum potential energy. The exact form of the potential function and the parameters that were used for the calculations reported in this paper are presented elsewhere.[14] In paper II of this series,[1] the parameters of the potential function (both the vacuum potential energy and the hydration free energy) will be adjusted to optimize the fit between predicted and observed structures for surface loops of proteins having known crystal structures.

2.3 Hardware

Our computer program for global energy minimization of surface loops has been developed on a relatively inexpensive computer system consisting of a Prime 550 minicomputer, an attached FPS-5200 array processor, and an attached Evans and Sutherland PS330 graphics system. All of the calculations completed thus far have been performed on this system. The FPS-5200 has a theoretical peak performance of 12 megaflops. The subprograms that evaluate function and derivatives (first and second) for the vacuum potential energy, the hydration free energy, and the sum of harmonic distance constraints, the subprograms for minimization, and the subprograms for removing overlaps have all been programmed in assembly language to run efficiently on the FPS-5200.

2.4 Local Minimization of Large Molecules

Our local minimization procedure uses the following four techniques to increase efficiency in applications to large molecules: a cutoff distance, minimization with respect to subsets of degrees of freedom, analytical second derivatives, and distance constraints between rigid segments. The interaction energies of atom pairs separated by a distance greater than some cutoff distance are not included in the function that is minimized. The cutoff distance reduces the number of interactions that are calculated at each step of the minimization. The energy is minimized with respect to subsets of degrees of freedom. For a given subset, the interaction energies of atom pairs whose distance is independent of the degrees of freedom contained in the subset are not included in the function that is minimized. Minimization with respect to subsets of degrees of freedom further reduces the number of interactions that are calculated at each step of the minimization. The memory size of the FPS-5200 ($1/2$ megawords) limits the number of degrees of freedom within a subset to 250. This is the only restriction on subsets used for local minimization. The program calculates analytical first and second derivatives with respect to the subset of degrees of freedom that is being used. Minimization steps are calculated using Newton's method. A Newton minimization procedure gives better convergence than a quasi-Newton procedure and requires fewer steps than a conjugate gradient procedure. The program provides the option of including in the total energy a sum of harmonic constraints on the distances of atom pairs. These distance constraints

are often used to fix the relative position and orientation of a pair of rigid segments (segments containing no variable backbone dihedral angles). When a pair of rigid segments is fixed by distance constraints, interactions between the rigid segments are not included in the function that is minimized. Distance constraints between rigid segments further reduce the number of interactions that are calculated at each step of the minimization. These techniques allow us to work with proteins of arbitrary size.

The following procedure is used for local minimization of a protein with respect to all degrees of freedom. A collection of subsets is chosen such that the union of the subsets contains all degrees of freedom and such that the subsets are highly overlapped. This collection is then cycled through several times, partially minimizing with respect to the degrees of freedom within a subset before moving to the next. This procedure is most efficient when the subsets contained in the collection control the motions of spatially localized regions of the protein.

For large molecules, an efficient procedure for local minimization with respect to all degrees of freedom is not very useful for global minimization. For small molecules, the most successful attempts at global energy minimization have been buildup methods and methods that generate a sequence (or a trajectory) of local minima.[3] A procedure for local minimization with respect to all degrees of freedom is the primary component of both approaches. However, local minimization with respect to all degrees of freedom requires computer time that is at best proportional to the square of the size of the molecule. Therefore, the number of local minima that can be examined by these methods decreases rapidly as the size of the molecule increases.

The term local minimum will be used to refer both to the result of a local minimization with respect to all degrees of freedom and to the result of a local minimization with respect to a subset of degrees of freedom. However, in all cases, the subset of degrees of freedom with respect to which the energy was minimized will be clear from the context. Local minima that are obtained when energy is minimized with respect to a subset of degrees of freedom are not local minima when energy is minimized with respect to all degrees of freedom. However, when a collection of local minima is obtained by minimizing with respect to a subset that restricts motion to a spatially localized region of the molecule, further minimization of the collection with respect to all degrees of freedom would be expected to produce only small changes both in the values of dihedral angles and in energy differences between conformations.

For large molecules, an efficient procedure for local minimization with respect to subsets of degrees of freedom provides a mechanism for efficient examination of the collection of local minima that can be obtained by deforming a spatially localized region of some initial structure. When rigid segments are fixed, local minimization with respect to subsets of degrees of freedom requires computer time that is approximately independent of the size of the molecule. For more general cases in which rigid segments are not fixed, the required computer time is approximately proportional to the size of the molecule. Therefore, methods for global minimization of large molecules that are based on local minimization with respect to subsets of degrees of freedom allow examination of a much larger collection of local minima than methods based on local minimization with respect to all degrees of freedom.

Our procedure for local minimization with respect to subsets of degrees of freedom is the primary component of our procedure for global energy minimization of surface loops. The subset of degrees of freedom that is currently being used for global energy minimization of surface loops consists of all backbone degrees of freedom (including ω) and all side chain degrees of freedom for each residue of the surface loop and of all side chain degrees of freedom for each side chain that could possibly contact the surface loop (for some conformation of the surface loop and some conformation of the side chain). Our local minimization procedure allows us to minimize the energy of a large molecule with respect to the subset associated with a surface loop using only about twice as much computer time as that required to minimize the energy of the surface loop in isolation. The rigid segments of the protein are held fixed relative to each other by harmonic constraints (with force constants of 10^3 kcal/mol-Å^2) on the distances of approximately 100 atom pairs, and interactions between these rigid segments are not calculated. The function that is minimized includes only interactions of the surface loop with itself and interactions between the surface loop and the rest of the protein.

2.5 Global Energy Minimization of Surface Loops

Global energy minimization with respect to the subset of degrees of freedom that is associated with a surface loop (which is defined in the previous section) is accomplished by generating a sequence (or a trajectory) of local minima such that each point of the trajectory has lower energy than the previous point. The method that we use to generate a trajectory of local minima for global minimization is analogous to the method that is commonly used to generate a trajectory of conformations for local minimization. In local minimization, a step of the trajectory is generated by examining all of the conformations that are contained within a multidimensional sphere that is centered about the current point of the trajectory and then retaining the conformation with lowest energy as the next point of the trajectory. In global minimization, a step of the trajectory is generated by examining a collection of local minima that surround the current point of the trajectory and then retaining the local minimum with lowest energy as the next point of the trajectory.

The following terms are introduced to simplify the description of our procedure for global energy minimization. A trajectory of local minima that is generated to accomplish global energy minimization with respect to the subset of degrees of freedom that is associated with a surface loop will often be referred to as a surface-loop trajectory. This will enable us to distinguish in a concise way a surface-loop trajectory from a trajectory of local minima that is generated to accomplish global energy minimization with respect to some other type of subset of degrees of freedom. The steps of a surface-loop trajectory will often be referred to as component steps. This will enable us to distinguish in a concise way the steps of a surface-loop trajectory from the steps of a local minimization or the steps of some other type of calculation. At each component step, the current point of the trajectory will be referred to as the initial structure. This will enable us to simplify the description of a single component step. Minimization with respect to the side chain degrees of freedom that are contained in the subset of degrees of freedom that is associated with a surface loop will be referred to as minimization with respect to side chain degrees of freedom (as if no other side chain degrees of freedom existed). This will enable us to simplify the description of part C of the component step.

The following procedure is used to generate a collection of low-energy local minima that surround the current point of a surface-loop trajectory:

A) A large collection of backbone structures is generated by deforming a seven-residue segment of the initial backbone structure.

B) A collection of *low-energy* backbone structures is generated by applying local energy minimization to the backbone structures obtained in part A (interactions involving the side chains that will be searched in part C are not included in the energy).

C) One low-energy side chain structure is generated for each of the low-energy backbone structures obtained in part B.

D) A collection of low-energy local minima is generated by applying local energy minimization to the structures obtained in part C.

In part A of this procedure, a large collection of backbone structures is generated by deforming a seven-residue segment of the initial backbone structure. A seven-residue segment of the initial backbone structure is deformed by obtaining alternative values for six consecutive (ψ, ϕ) pairs and the six intervening ω dihedral angles such that the ω values are set at 180° and such that the relative position and orientation of the nondeformed segments of the initial backbone structure are preserved. We group backbone dihedral angles into (ψ, ϕ) pairs, where ψ and ϕ refer to dihedral angles of residues $i-1$ and i, respectively, because this choice simplifies the description of our procedure for generating deformations of the backbone.

Three (ψ, ϕ) pairs are assigned values from discrete collections. Values for the remaining three (ψ, ϕ) pairs are calculated analytically. A more complete description of our procedure for deforming a seven-residue segment of the initial backbone structure is presented in Appendix A.

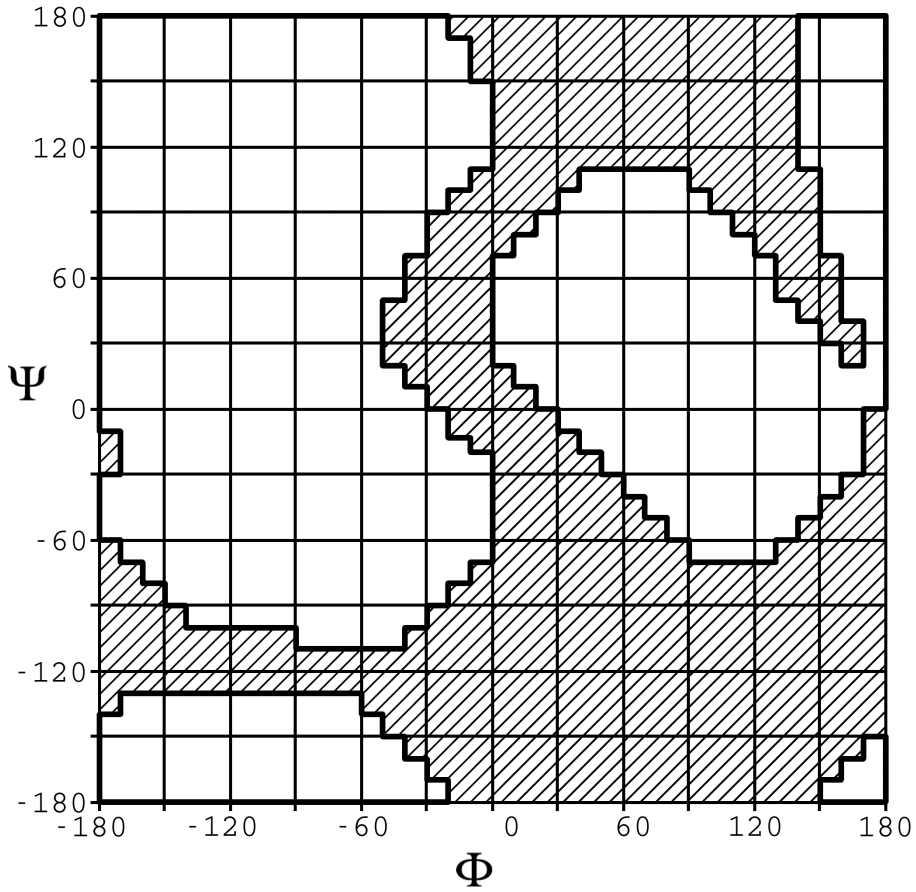

Figure 1: Acceptable (ϕ, ψ) values (in degrees) for alanine. Backbone deformations generated in part A of the component step of the trajectory are discarded without attempting to relieve overlaps if an alanine residue has a (ϕ, ψ) value in the shaded region. An acceptable set of (ϕ, ψ) values has been assigned to each of the 20 amino acids.

The deformations generated by our procedure are not exact in the sense that the relative position and orientation of the nondeformed segments of the structure are not exactly preserved. Our procedure simplifies the system of equations that determines exact deformations by replacing the exact backbone geometry with an approximate geometry. Values for the six (ψ, ϕ) pairs of the deformable segment that satisfy this simplified system of equations are not exact deformations. When the deformed segment contains a proline residue, the generated deformations differ from exact deformations in another way. A deformation is not accepted when the ϕ value of a proline residue is not within $60°$ of $-75°$. Otherwise, the value of ϕ for proline is moved to $-75°$ (the ECEPP value of ϕ for proline), and the value of the preceding ψ is moved an equal amount in the opposite direction. The deformations become exact during part B of this procedure in which a sum of harmonic distance constraints between rigid segments of the protein is included in the energy.

A deformation is also not accepted when the (ϕ, ψ) value of a residue in the deformed segment lies deeply within a high energy region because minimization starting from such a deformation requires a large structural change and, therefore, often results in a low-energy backbone structure that is more easily reached by minimization starting from another deformation. For example, deformations for which the (ϕ, ψ) value of an alanine residue lies in the shaded region of Figure 1 are eliminated.

To reduce the probability of redundant computation in part B, some of the calculated deformations are eliminated. Clusters of similar backbone deformations are identified, and one representative deformation from each cluster is retained. A description of our clustering procedure is presented elsewhere.[14] The resulting collection typically contains between 200 and 2,000 backbone structures.

In part B of this procedure, a collection of low-energy backbone structures is generated by applying local energy minimization to the backbone structures obtained in part A. Throughout this part of the procedure, interactions involving the side chains that will be searched in part C are not included in the energy. For each backbone structure obtained in part A, the 12-2e potential plus a sum of harmonic distance constraints between rigid segments of the protein is minimized with respect to all backbone degrees of freedom of the surface loop. An overlap is defined to be a 12-2e interaction with energy greater than 3.5 kcal/mol. A structure is retained only if all overlaps have been removed. The resulting collection of overlap-free backbone structures is clustered, and one structure from each cluster is retained. For each of the remaining overlap-free backbone structures, the total energy (the sum of the vacuum potential energy, the hydration free energy, and a collection of harmonic distance constraints between rigid segments of the protein) is minimized with respect to all backbone degrees of freedom of the surface loop. The resulting collection of low-energy backbone structures is clustered, and the lowest-energy structure from each cluster is retained. Structures for which the energy is not within 30 kcal/mol of the lowest energy obtained are eliminated. If the resulting collection contains greater than n structures, where n is either 40 or 80, then the collection is reduced to n structures by eliminating the structures that have the highest energies.

The total energy of a structure can be minimized with much less computer time after the overlaps have been relieved than would have been required had the overlaps not been previously relieved. Even though the movement required to relieve overlaps is usually much greater than the movement required to minimize the total energy of the resulting overlap-free structure, the computer time required to relieve overlaps is usually much less than the computer time required to minimize the total energy of the resulting overlap-free structure. Our procedure for relieving overlaps is simple yet extremely efficient. A description of this procedure is presented in Appendix B.

In part C of this procedure, one low-energy side chain structure is generated, by global energy minimization with respect to side chain degrees of freedom, for each of the low-energy backbone structures obtained in part B. For each low-energy backbone structure, global energy minimization with respect to side chain degrees of freedom is accomplished by generating a trajectory of low-energy side chain structures (which will be referred to as a side chain trajectory to distinguish it from the surface-loop trajectory) such that each point of the trajectory has lower energy than the previous point. The initial point of the side chain trajectory is the side chain conformation of the current point of the surface-loop trajectory (the side chain conformation of the initial structure whose backbone was deformed in part A). A trajectory of low-energy side chain conformations is generated by 1) selecting a collection of subsets of degrees of freedom such that the union of the subsets in the collection is the subset of side chain degrees of freedom that is being searched, and 2) cycling through this collection once, globally minimizing with respect to the degrees of freedom within a subset before moving to the next subset.

For each low-energy backbone structure, the following procedure is used to select a collection of subsets of side chain degrees of freedom. We first examine all pairs of the side chains that are being searched. If contact is possible between a pair of side chains (for some conformation of the two side chains), then the subset consisting of all the degrees of freedom contained in that pair of side chains is included in the collection. We then examine individually all of the side chains that are being searched. If a side chain is isolated in the sense that no contact is possible with any other side chain that is being searched, then the subset consisting of all the degrees of freedom contained in that side chain is included in the collection. Since contact between a pair of side chains may be possible for some backbone structures but not others, the collection of subsets that is selected by this procedure is often different for different low-energy backbone structures.

For each low-energy backbone structure, global energy minimization with respect to the subsets that are contained in the associated collection is accomplished by a buildup procedure.[20] For each side chain

dihedral angle of each amino acid, between one and six values have been selected as representative of the observed distribution of values for that degree of freedom. At each step of the buildup, a) one, two, or three degrees of freedom are selected from the subset with respect to which energy is being (globally) minimized, b) all combinations of representative values for these degrees of freedom are combined with the low-energy conformations that resulted from the previous step of the buildup to create starting points for energy minimization, c) for each starting conformation, first the 12-2e potential and then the total energy is minimized with respect to the current subset of degrees of freedom, and d) conformations for which the energy is within 16 kcal/mol of the lowest energy are retained. Interactions involving atoms within the pair of side chains (or the individual side chain) that is being searched whose position depends on degrees of freedom that have not yet been assigned representative values are not included in the energy. Otherwise, all of the interactions of the pair of side chains (or the individual side chain) with itself and with the rest of the protein are included.

In part D of this procedure, a collection of low-energy local minima is generated by applying local energy minimization to the structures obtained in part C. For each structure obtained in part C, the total energy is minimized with respect to the subset of degrees of freedom that is associated with the surface loop. To assure that each point of the trajectory will have lower energy than the previous point, local energy minimization is applied to the structure that corresponds to the current point of the trajectory, and the resulting local minimum is added to the previously obtained collection of local minima. A cutoff distance of 18 Å was used in all minimizations previous to this point. At this point, the 10 lowest-energy local minima are used as starting points for minimization of the total energy, with respect to the subset of degrees of freedom that is associated with the surface loop, using no cutoff distance.

The current point of the trajectory was the result of minimization with no cutoff distance in the previous component step. Since the current point of the trajectory is reinserted at the start of part D, it must be reminimized with a cutoff distance of 18Å to allow comparison of its energy with the energies of other structures.

In Figure 2, the algorithm that is used to generate a surface-loop trajectory is summarized by a flow diagram. In Figure 3, part C of the component step of a surface-loop trajectory (the algorithm that is used to generate a side chain trajectory) is summarized by a more detailed flow diagram.

In most cases, the subset of degrees of freedom that is associated with a surface loop includes both backbone and side chain degrees of freedom that are not directly searched in the current step of the surface-loop trajectory in the sense that these degrees of freedom are never assigned alternative values to create starting points for energy minimization. These degrees of freedom may be searched in succeeding steps of the trajectory, or they may serve only to provide additional flexibility during minimization.

The input to our program for global energy minimization of surface loops is quite simple. User interaction with the program occurs only between component steps of the surface-loop trajectory. This interaction consists of observing the results of the previous step, modifying the input such that the current step will differ from the previous step, and restarting the program. The input for each component step consists of the protein sequence; initial dihedral angles for the protein; and sets of residues for which backbone motion will be allowed, for which side chain motion will be allowed, for which backbone degrees of freedom will be searched, and for which side chain degrees of freedom will searched. The global minimization program is currently set up to handle backbone motion in only one segment, but only minor modifications would be needed to allow backbone motion in two or more segments simultaneously.

2.6 Test Calculations on Protein Crystal Structures

To test the efficiency of our global search procedure, calculations have been carried out on proteins for which crystal structures have been determined. The following nine high-resolution protein crystal structures have been selected from the Brookhaven protein data bank:[21] 1PPT (avian pancreatic polypeptide), 1CRN (crambin), 5PTI (trypsin inhibitor), 2EBX (erabutoxin B), 2RHE (immunoglobulin B-J fragment), 5RSA (ribonuclease A), 1LZ1 (lysozyme), 1PPD (papain D), and 2PTN (trypsin). To allow global energy minimization using dihedral angle degrees of freedom, regularized structures (having ECEPP val-

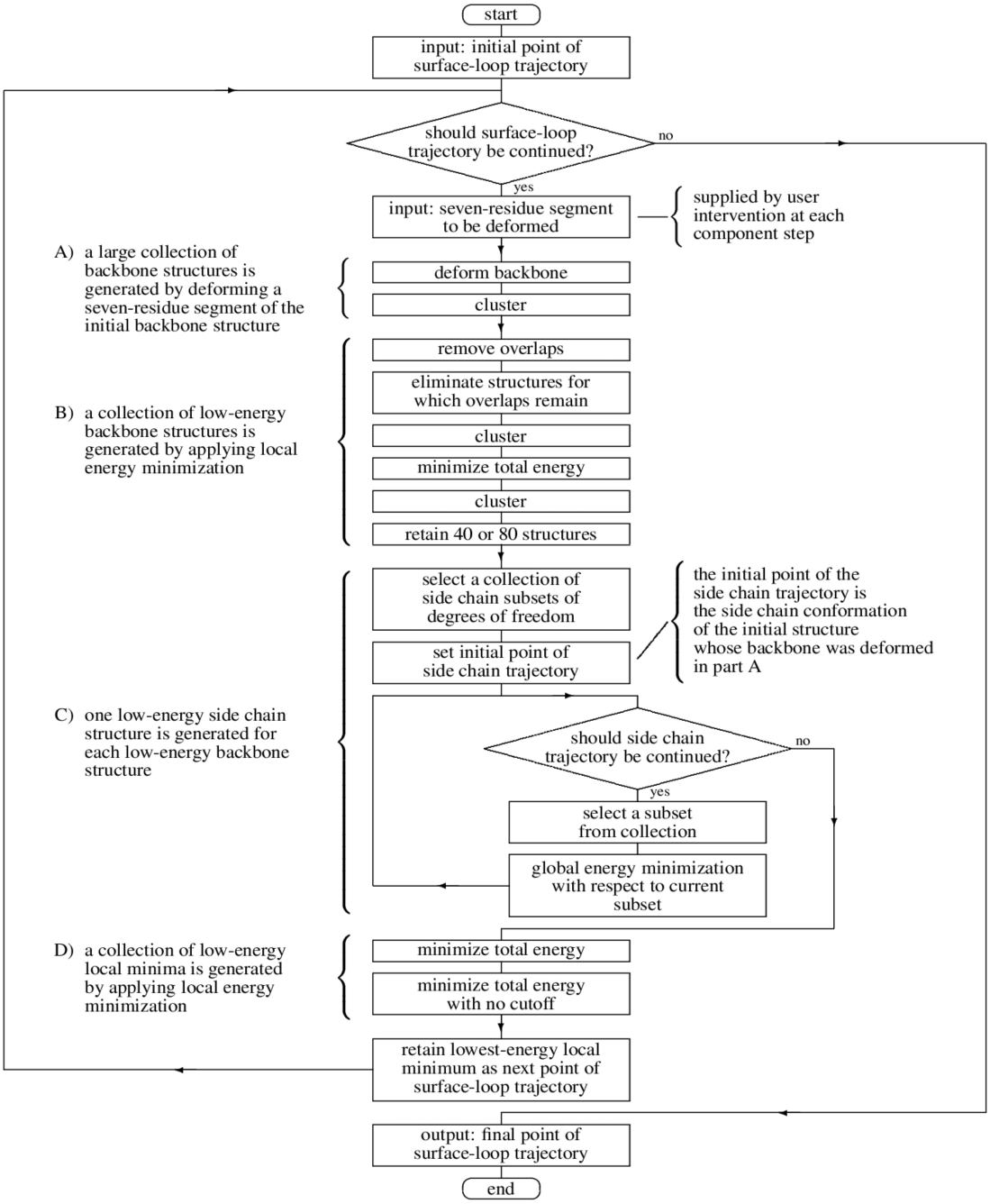

Figure 2: Flow diagram of the algorithm used to generate a surface-loop trajectory. A component step of the trajectory corresponds to one complete cycle of the diagram. At each component step, the algorithm generates a collection of low-energy local minima then retains the local minimum with lowest energy as the next point of the trajectory. In part A, the initial backbone structure is the backbone structure of the current point of the surface-loop trajectory. In part C, a side chain trajectory is generated separately for each of the low-energy backbone structures that was obtained in part B. The initial point of the side chain trajectory is the side chain conformation of the current point of the surface-loop trajectory.

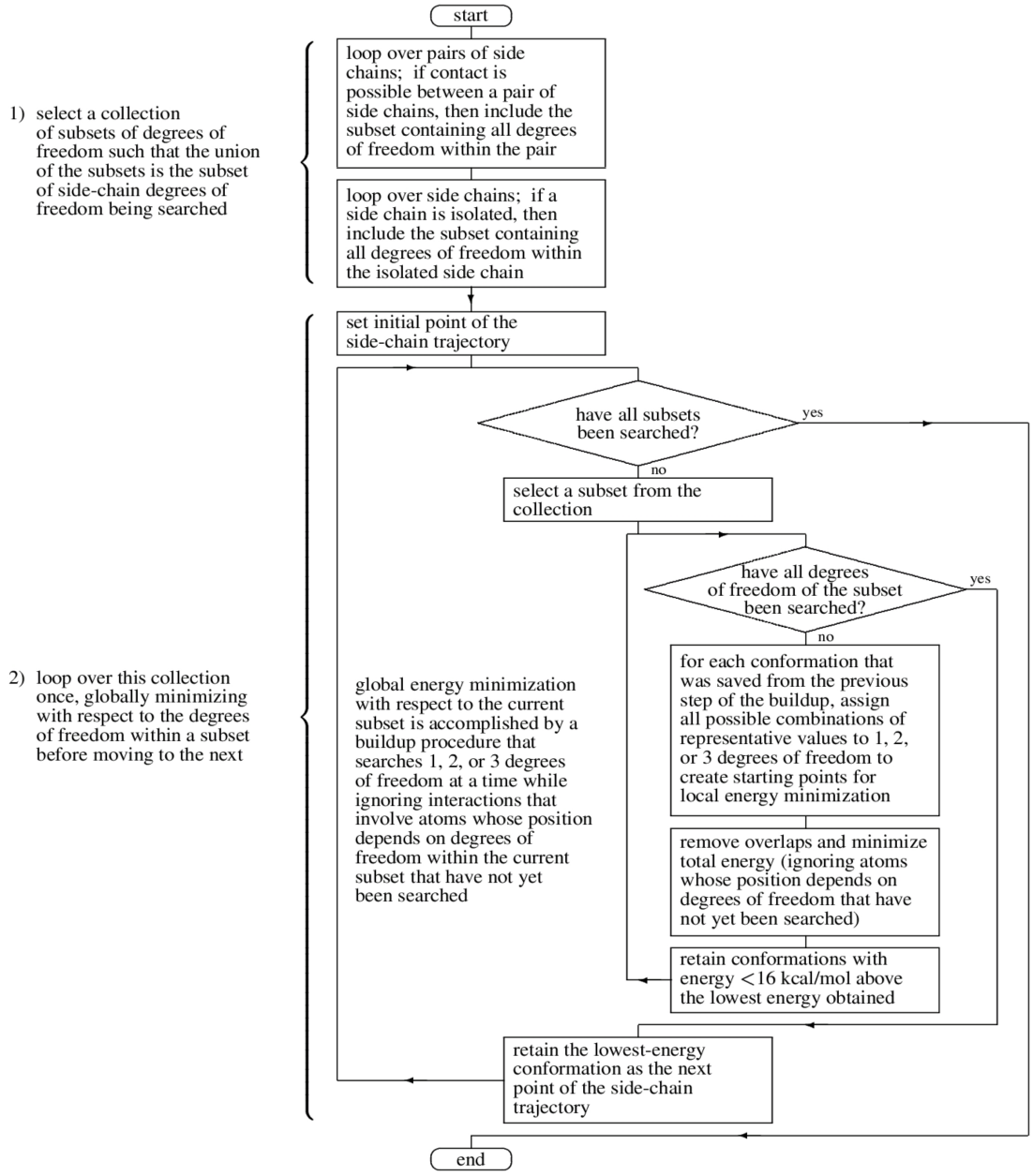

Figure 3: Flow diagram of the algorithm used to generate a side chain trajectory. This algorithm is applied separately to each of the low-energy backbone structures that result from part B of the component step of a surface-loop trajectory. Values of backbone degrees of freedom are held fixed throughout the algorithm. The initial point of the side chain trajectory is the side chain conformation of the current point of the surface-loop trajectory. A step of the side chain trajectory corresponds to one complete cycle of the diagram. At each step of the side chain trajectory, the algorithm generates a collection of low-energy local minima, and then retains the local minimum with lowest energy as the next point of the trajectory. The points of a side chain trajectory are local minima in the sense that each point results from local energy minimization with respect to a subset of degrees of freedom. For a surface-loop trajectory, the subset with respect to which energy is a minimum is the same for each point, whereas for a side chain trajectory, the subset with respect to which energy is a minimum is different for each point.

ues of bond lengths and bond angles) have been fit to the coordinates of these crystal structures. Our fitting procedure consists of local minimization (with respect to all degrees of freedom) of the vacuum potential energy plus a sum of harmonic constraints on the distances of about 4000 atom pairs. Target distances were obtained from the crystal structure, and force constants were set at 1 kcal/mol-Å^2. The C$^\alpha$ rms deviations between the regularized crystal structures and the actual crystal structures are 0.20, 0.13, 0.18, 0.25, 0.25, 0.28, 0.24, 0.25, and 0.26 for 1PPT, 1CRN, 5PTI, 2EBX, 2RHE, 5RSA, 1LZ1, 1PPD, and 2PTN, respectively. Surface segments have been selected from among these structures, and the associated spaces of conformations have been searched. The resulting collections of low-energy local minima have been useful both for evaluating the completeness of the global search procedure and for identifying problems with the potential function. In this paper, the efficiency of our procedure for global energy minimization of surface loops is evaluated based on the results of searches for surface segments of BPTI and bovine trypsin. In paper II of this series,[1] potential functions will be evaluated based on their ability to distinguish the native local minimum from a collection of low-energy local minima.

2.7 Notation

In order to allow concise description of backbone structure, we have partitioned the (ϕ, ψ) map into regions. The (ϕ, ψ) regions used in this project are more closely correlated with distributions of (ϕ, ψ) values obtained from high resolution crystal structures of globular proteins than those used previously.[22] In Figures 4a and 4b, regions are defined separately for non-glycine and glycine residues, respectively. In Figures 4c and 4d, distributions of the (ϕ, ψ) values that are obtained from a set of high-resolution protein crystal structures are presented for non-glycine and glycine residues, respectively.

3 RESULTS and CONCLUSIONS

The crystal structure of BPTI[23] was obtained from the 5PTI entry of the protein data bank. Five surface segments, each containing seven residues, were selected for global energy minimization. These segments and the subsets of degrees of freedom with respect to which the energies of these segments were minimized are specified in Table I. Since each of these surface segments contains seven residues, and since the search at each component step of a trajectory attempts to examine all of the local minima that can be obtained by deforming a seven-residue segment, the trajectories of local minima that were generated to accomplish global energy minimization of these surface segments consist of a single component step.

The results of these global searches are summarized in Tables II to IV. The number of backbone structures in each of the collections that was generated by the backbone search procedure (consisting of parts A and B of the component step of a trajectory) is presented in Table II. The number of backbone structures that were examined by applying local energy minimization during the backbone search (parts A and B of the component step), the number of side chain structures that were examined by applying local energy minimization during the side chain searches for 40 of the resulting low-energy backbone structures (part C of the component step), and the computer times that were required to accomplish these searches and to accomplish the entire component step are presented in Table III. The energies, sequences of (ϕ, ψ) regions, and RMS deviations from the regularized crystal structure are presented in Table IV for the ten lowest-energy structures that were obtained from the global search.

The crystal structure of bovine trypsin[24] was obtained from the 2PTN entry of the protein data bank. A nine-residue surface segment, the seventh surface loop along the sequence, was selected for global energy minimization. This segment and the subset of degrees of freedom with respect to which the energy of this segment was minimized are specified in Table V. Since this surface segment contains nine residues, the trajectory of local minima that was generated to accomplish global energy minimization of this surface segment consists of more than one component step. Backbone deformations were generated for the seven-residue segments containing residues 123-129, 125-131, and 124-130 in component steps

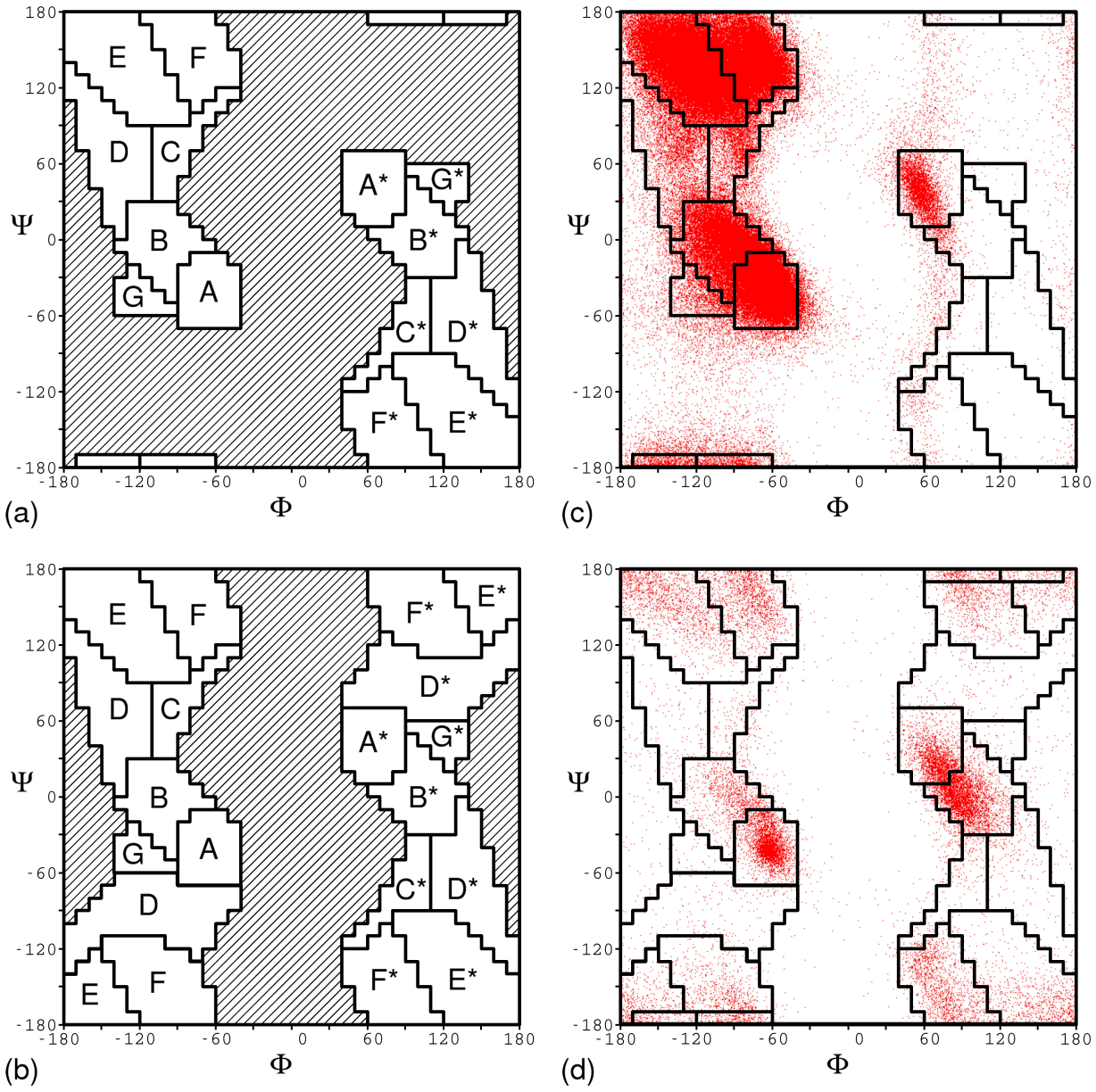

Figure 4: a) Definition of (ϕ, ψ) regions for non-glycine residues. The shaded region will be referred to as the X region. b) Definition of (ϕ, ψ) regions for glycine residues. The shaded region will be referred to as the X region. c) Distribution of (ϕ, ψ) values for non-glycine residues in a collection of high resolution protein crystal structures. d) Distribution of (ϕ, ψ) values for glycine residues in the same collection of protein crystal structures.

Table I: Description of Surface Segments of BPTI and the Associated Subsets of Degrees of Freedom. [a]

Loop 1				Loop 2				Loop 3				Loop 4				Loop 5			
#	aa	b	s	#	aa	b	s	#	aa	b	s	#	aa	b	s	#	aa	b	s
												7	glu	0	1				
												10	tyr	0	1				
								1	arg	0	1	20	arg	0	2				
								22	phe	0	1	35	tyr	0	1	20	arg	0	2
6	leu	2	2	14	cys	2	2	23	tyr	2	1	36	gly	2	0	44	asn	2	2
7	glu	2	2	15	lys	2	2	24	asn	2	2	37	gly	2	0	45	phe	2	2
8	pro	2	0	16	ala	2	1	25	ala	2	1	38	cys	2	2	46	lys	2	2
9	pro	2	0	17	arg	2	2	26	lys	2	2	39	arg	2	2	47	ser	2	2
10	tyr	2	2	18	ile	2	2	27	ala	2	1	40	ala	2	1	48	ala	2	1
11	thr	2	2	19	ile	2	2	28	gly	2	0	41	lys	2	2	49	glu	2	2
12	gly	2	0	20	arg	2	2	29	leu	2	2	42	arg	2	2	50	asp	2	2
41	lys	0	2	34	val	0	2	31	gln	0	2	44	asn	0	2	53	arg	0	2
42	arg	0	2	46	lys	0	1									54	thr	0	2

[a]For each surface segment, the subset of degrees of freedom with respect to which energy was minimized consists of all backbone and side chain degrees of freedom for each residue of the surface segment (the central group of residues) and all side chain degrees of freedom for each side chain that could possibly contact the surface segment (the residues outside of the central group). Column headings have been abbreviated as follows: residue number– #, amino acid– aa, backbone search parameter– b, side chain search parameter– s. For each surface segment, the degrees of freedom with respect to which energy was minimized and the degrees of freedom that were assigned alternative values to create starting points for energy minimization are specified by backbone and side chain search parameters. A value of 0 for the backbone or side chain search parameter indicates that the backbone or side chain dihedral angles of the corresponding residue were fixed during the search procedure. A value of 1 indicates that the corresponding backbone or side chain dihedral angles were variable but were not searched. A value of 2 indicates that the corresponding backbone or side chain dihedral angles were searched.

Table II: Number of Backbone Structures at Each Stage of the Backbone Search for Surface Segments of BPTI. [a]

Stage	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5
deform backbone	467	306	1502	1739	913
cluster	297	302	1447	1643	898
remove overlaps	280	255	1289	922	881
cluster	263	232	1130	752	779
minimize energy	263	232	766	752	779
cluster	178	159	343	528	525
restrict size[b,c]	40	40	40	40	40

[a]The backbone search procedure consists of parts A and B of the component step of a trajectory.

[b]This collection was generated by retaining the 40 lowest-energy backbone structures of the previous collection.

[c]For each surface segment, the backbone structure corresponding to the regularized crystal structure is recovered in this collection of low-energy backbone structures.

Table III: Number of Structures Examined by Local Energy Minimization in Backbone and Side Chain Searches of BPTI Surface Segments and Computer Time Required to Accomplish These Searches. [a]

	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5
backbone search:[b]					
structures examined	297	302	1447	1643	898
time required (hr)	11	11	57	50	30
side chain search:[c]					
structures examined	11,205	13,728	6,552	12,321	30,942
time required (hr)	24	32	11	35	107
entire component step:					
time required (hr)	37	47	72	92	144

[a]Local energy minimization is accomplished in two parts: first overlaps are removed, then the total energy is minimized. The set of structures that was examined by local energy minimization includes those structures for which local minimization was aborted either because overlaps could not be removed or because the resulting overlap-free structure was not retained by our clustering procedure. Computer time refers to elapsed (rather than cpu) time. Approximately 80% of the elapsed time was spent in the FPS-5200 array processor, which is a single-user machine. The cpu of the Prime 550 is shared with other users.

[b]The backbone search procedure consists of parts A and B of the component step of a trajectory.

[c]The side chain search procedure consists of part C of the component step of a trajectory.

1, 2, and 3 respectively. The choice of this particular sequence of seven-residue segments was arbitrary. The results of these component steps are summarized in Tables VI to VIII.

The efficiency of our procedure for global energy minimization of surface loops can be evaluated on two levels, the efficiency with which the component steps of a trajectory are able to locate the global energy minima within the subspaces of conformations searched by these component steps and the efficiency with which the trajectory of local minima is able to locate the global energy minimum within the space of conformations searched by the trajectory as a whole. A useful measure of the efficiency of a procedure for global energy minimization is the amount of computer time that is required to obtain the global minimum of a specified potential function. The amount of computer time required to complete component steps of a trajectory is presented in Tables III and VII. The amount of computer time required to complete a trajectory consisting of three component steps is presented in Table VII. However, since the global minimum of our current potential function often differs significantly from the experimentally determined crystal structure and cannot be determined with absolute confidence by existing procedures for global energy minimization, the evidence for the completeness of these searches is suggestive rather than conclusive.

In all of the component-step searches that have thus far been completed, the backbone structure corresponding to the current point of the trajectory has been recovered in the collection of low-energy backbone structures that is obtained in part B of the component-step search. If the backbone searches at the component steps of a trajectory are not always complete, then the backbone structure corresponding to the current point of the trajectory should sometimes not be recovered. This evidence strongly suggests that the backbone searches at the component steps of a trajectory are complete. For each of the surface segments of BPTI that was selected for global energy minimization, component-step searches result in structures that have lower energy than the native structure. This evidence suggests that component-step searches are reasonably complete. The computational efficiency of component-step searches is such that trajectories consisting of about ten local minima are feasible using an FPS-5200 array processor. On a computer capable of 1200 megaflops, which more accurately represents the limits of current technology, trajectories consisting of about one thousand local minima become feasible.

For the nine-residue surface loop of trypsin, the trajectory of local minima that was generated by our search procedure is successful at 1) lowering the energy and 2) generating a collection of low-energy structures that contains several significantly distinct backbone conformations. Establishing the completeness of searches that are accomplished by generating a trajectory of local minima is quite difficult. Convergence of a trajectory is probably some indication that a search is complete. However, a trajectory

Table IV: Energy, Sequence of (ϕ, ψ) Regions, and RMS Deviation from the Regularized Crystal Structure for Low-Energy Structures of BPTI Surface Segments. [a]

		energy[b] (kcal/mol)	(ϕ, ψ) regions	RMSD[c] (Å)
	1	-703.6	B F F F E A E*	0.27
	2	-693.8	B E F C F E E	1.71
	3	-693.5	B F F C F E E	1.62
	4	-693.5	A E F F F E E	1.57
Loop 1	5	-692.8	A F F C F F E	1.87
	6	-692.2	A F F C A E E	2.30
	7	-691.2	B E F C F F E	1.76
	8	-688.4	C E F C F F E	1.56
	9	-687.4	B F F C F F E	0.95
	10	-686.2	⟶A F F F E A E*	0.27
	1	-633.4	F D A A*E F E	0.91
	2	-633.1	F E E C E E E	1.39
	3	-631.6	F D B A*F F E	0.92
	4	-629.5	F D X D E F E	0.45
Loop 2[d]	5	-629.3	F D F F E E E	0.74
	6	-629.0	F B C F E F E	0.71
	7	-628.5	F A D X E F F	1.97
	8	-627.9	F B F C F E E	0.74
	9	-627.8	F E E D E E E	1.47
	10	-627.1	F B E C F E E	0.66
	1	-714.6	F E A B G B*E	1.15
	2	-714.3	F E B A B B*E	0.27
	3	-712.5	⟶F E A A B A*E	0.28
	4	-710.5	F E A A B A*E	0.97
Loop 3	5	-709.7	F E A C X B*E	0.93
	6	-709.0	F E A G A B*E	0.67
	7	-708.9	F E A A A B*E	0.65
	8	-708.8	F E A X A*A*E	1.72
	9	-708.3	F F B B B A*E	1.06
	10	-707.3	F F X C*A A*E	1.25
	1	-825.4	B B*E A*F F A	0.32
	2	-825.3	B B*E A*F F A	0.31
	3	-824.5	B B*E A*F F A	0.26
	4	-823.5	⟶B B*E A*F F A	0.28
Loop 4	5	-818.6	B B*E E C E G	2.13
	6	-817.9	B B*E E F E G	2.12
	7	-817.7	B B*E A*F F A	0.31
	8	-811.4	B E*A*E E E G	2.53
	9	-810.1	B B*E F D E B	2.04
	10	-802.9	B C*E D X E G	2.11
	1	-806.4	D E B E A A A	0.16
	2	-805.3	D E B E A A A	0.15
	3	-804.7	D E B E A A A	0.15
	4	-803.8	D E B E A A A	0.15
Loop 5	5	-801.1	D E A E A A A	0.12
	6	-799.8	D E B E A A A	0.31
	7	-798.5	D E B E A A A	0.14
	8	-797.5	⟶D E B E A A A	0.15
	9	-797.0	D E B E A A A	0.41
	10	-790.5	D E C D A A A	0.49

[a] The structure that corresponds to the regularized crystal structure is marked by an arrow.

[b] The energies reported in this Table were calculated without a cutoff distance.

[c] The RMS deviations reported in this Table are RMS deviations from the regularized crystal structure (having ECEPP values of bond lengths and bond angles). The C^{α} RMS deviation between the regularized crystal structure and the actual crystal structure is 0.177Å. Deviations are calculated only for the backbone heavy atoms of the surface loop that is being searched.

[d] The sequence of (ϕ, ψ) regions for loop 2 of the regularized crystal structure is F D F D E F E . Thirteen structures were obtained for this surface segment that have lower energy than the regularized crystal structure.

Table V: Description of a Surface Loop of Bovine Trypsin and the Associated Subset of Degrees of Freedom. [a]

Loop 7			
#	aa	b	s
123	asn	2	2
124	thr	2	2
125	lys	2	2
126	ser	2	2
127	ser	2	2
128	gly	2	0
129	thr	2	2
130	ser	2	2
131	tyr	2	2

[a] See footnote a of Table I.

Table VI: Number of Backbone Structures at Each Stage of the Backbone Search for Each Component Step of the Trajectory Generated to Accomplish Global Energy Minimization of Surface Loop 7 of Bovine Trypsin. [a]

Stage	Step 1	Step 2	Step 3
deform backbone	1730	631	534
cluster	1349	516	418
remove overlaps	1295	500	394
cluster	1224	466	377
minimize energy	1224	466	377
cluster	973	362	300
restrict size[b,c]	80	80	80

[a] See footnote a of Table II.

[b] This collection was generated by retaining the 80 lowest-energy backbone structures of the previous collection.

[c] For each component step, the backbone structure corresponding to the previous point of the trajectory is recovered in this collection of low-energy backbone structures.

Table VII: Number of Structures Examined by Local Energy Minimization in Backbone and Side Chain Searches at Each Component Step of the Trajectory for Surface Loop 7 of Bovine Trypsin and Computer Time Required to Accomplish These Searches. [a]

	Step 1	Step 2	Step 3
backbone search:			
structures examined	1349	516	418
time required (hr)	134	46	36
side chain search:			
structures examined	28,818	33,234	34,374
time required (hr)	66	76	71
entire component step:			
time required (hr)	216	137	121

[a] See footnotes a-c of Table III.

Table VIII: Energy, Sequence of (ϕ, ψ) Regions, and RMS Deviation from the Regularized Crystal Structure for Low-Energy Structures Obtained at Each Component Step of the Trajectory for Surface Loop 7 of Bovine Trypsin. [a]

		energy[b] (kcal/mol)	(ϕ, ψ) regions	RMSD[c] (Å)
Step 1[d]	1	-1281.7	F B E F A∗E∗E D E	0.97
	2	-1278.3	F B F A D E∗E E E	1.17
	3	-1278.2	F D E A E E∗E X E	1.10
	4	-1277.4	F B F F A∗E∗E D E	1.05
	5	-1275.5	F B E F A∗E∗E E E	1.35
	6	-1271.8	F B E F A∗E∗E D E	1.32
	7	-1270.4	F A E F A∗E D E E	1.15
	8	-1269.5	F B F C A∗E E D E	1.81
	9	-1268.8	F B A∗F A∗C∗E E E	5.11
	10	-1257.5	F X C A D X∗E E F	6.41
Step 2	1	-1281.7	⟶F B E F A∗E∗E D E	0.97
	2	-1279.3	F B F A E E∗E D F	1.12
	3	-1279.1	F B F F A∗E∗E D E	1.08
	4	-1278.7	F B F A B E∗E D E	0.89
	5	-1278.4	F B F F A∗E∗E D E	1.11
	6	-1278.2	F B E F A∗E E E E	1.04
	7	-1277.9	F B F F A∗E∗E D E	1.06
	8	-1276.3	F B F F A∗E∗E D E	1.09
	9	-1274.8	F B E F A∗E E D E	1.04
	10	-1274.4	F B E F A∗E E D F	1.10
Step 3	1	-1283.0	F B E A A C E D E	0.89
	2	-1281.7	⟶F B E F A∗E∗E D E	0.97
	3	-1280.6	F B E A C E∗E D E	0.90
	4	-1279.3	F B C A C E∗E D E	0.80
	5	-1278.8	F B E A C E∗E D E	0.80
	6	-1278.7	F B C A C E∗E D E	0.80
	7	-1277.8	F B E A E E∗E E E	1.06
	8	-1275.3	F B E A C E∗E E E	1.02
	9	-1274.0	F B E F A∗E D D E	1.50
	10	-1271.2	F B A∗E A∗C∗E E E	3.89

[a] The structure that corresponds to the previous point of the trajectory is marked by an arrow. The initial point of the trajectory is the regularized crystal structure.

[b] The energies reported in this Table were calculated without a cutoff distance.

[c] The RMS deviations reported in this Table are RMS deviations from the regularized crystal structure (having ECEPP values of bond lengths and bond angles). The C^{α} RMS deviation between the regularized crystal structure and the actual crystal structure is 0.264Å. Deviations are calculated only for the backbone heavy atoms of the surface loop that is being searched.

[d] The sequence of (ϕ, ψ) regions for loop 7 of the regularized crystal structure is F B F A E E∗E E E . In the first component step, 23 structures were obtained that have lower energy than the regularized crystal structure.

of local minima could possibly converge to a local minimum other than the global minimum. For a surface loop containing more than seven residues, some regions of the space of backbone conformations could possibly be accessible to a trajectory only if backbone deformations are generated for segments containing eight or more residues at each component step. Such a situation, if it exists, represents a generalization of the multiple-minima problem to the space of local minima. Convergence of several trajectories to the same conformation is probably a good indication that a search is complete. However, since generation of several trajectories is computationally expensive, and since establishing the completeness of global searches is less relevant to the primary goal of the overall project than increasing the accuracy of the potential function, we have not attempted to generate several trajectories for any surface segment.

Our procedure for global energy minimization of surface loops allows efficient examination of large collections of local energy minima. When applied to protein crystal structures, this procedure provides a mechanism for obtaining structures that have lower energy than the native structure. Even though evidence for the completeness of the resulting global searches is not conclusive, the capabilities of our global search procedure are clearly sufficient for the first intended application, identifying and correcting problems with the potential function, and probably sufficient for the second intended application, predicting protein structure using both the crystal structure of a homologous protein and global energy minimization.

4 DISCUSSION

A clear distinction should be made between structure prediction using a combination of sequence homology and global energy minimization, which we and others are attempting to implement,[4–8] and structure prediction using sequence homology alone, in which energy is not used or is used in a limited way to examine only one (or a few) local minima.[9,10,25–27] Structure predictions using a combination of sequence homology and global energy minimization attempt to obtain a significant amount of structural information from an empirically parameterized function representing energy. In contrast, structure predictions using sequence homology alone obtain very little structural information from energy. The more advanced procedures of the latter type[9,10,26] use sequence homology to obtain information about the structure of the surface.

The procedure for generating one low-energy side chain conformation for each low-energy backbone conformation (part C of the component step of the trajectory) has evolved into its present form mainly because of selective pressure on the required computational time. We have not yet accumulated enough data for a complete evaluation of the effectiveness of this procedure. The possibility exists that in some cases component-step searches are not complete because the values of backbone degrees of freedom are not allowed to vary during the local minimizations of the side chain search. However, even if component-step searches are in some cases not complete, the trajectory as a whole could still be effective for sampling the space.

Our procedure for global energy minimization of surface loops provides a mechanism for utilizing the independence of spatially separated regions of a protein. The following example demonstrates the importance of utilizing independence. Consider a protein for which distance constraints have been obtained from NMR experiments. Assume that the structure of the protein is uniquely determined everywhere with the exception of six surface segments, that these surface segments are spatially separated, and that ten local minima are consistent with the distance constraints for each of these surface segments. One approach to global energy minimization for this system consists of applying one of the various distance geometry procedures[28,29] to satisfy the distance constraints and then minimizing the energy of the resulting structure. Since approximately 10^6 local minima exist for this system, this procedure would need to be applied on the order of 10^6 times to obtain the native conformation. An alternative approach consists of applying the distance geometry program once to satisfy the distance constraints, and then applying our procedure for global energy minimization of surface loops independently to each of the six

surface loops. This alternative approach would require about 60 minimizations.

Procedures for global energy minimization of surface loops offer an unprecedented opportunity for improving the accuracy of the potential function. Previous methods for comparing the accuracies of various potential functions have been based on energy minimization within the native potential well and subsequent analysis of the movement away from the experimental structure.[30–32] Such methods are more sensitive to the radii of the various atom types than to the magnitudes, distance dependences, and angular dependences of the various interactions. Methods for comparing the accuracies of various potential functions based on global searches of proteins have not been possible because a molecule the size of BPTI is too large for current global optimization methods and because the ensemble of solution structures for a molecule significantly smaller than BPTI is difficult to characterize experimentally.

Structure prediction using both the crystal structure of a homologous protein and global energy minimization offers the opportunity for a major advance in the attempt to obtain useful structural information from a function representing energy. The crystal structure of a homologous protein is available in a significant fraction of protein structural problems. The efficiency of our search procedure and the existence of computer hardware capable of gigaflops suggest that in most of these cases the technology for accomplishing the necessary global energy minimization is currently available (although this assessment could change if the complexity of the potential function needs to be significantly increased). Therefore, the development of a more accurate potential function is perhaps the only obstacle preventing numerous successful applications.

Our program has evolved slowly over a period of several years and will probably continue to evolve, although more slowly, throughout the project. It has reached a level of efficiency that is sufficient for many of the intended applications, and therefore, the focus of the project has now shifted from program development to applications. However, an accurate potential function would assist in the continued evolution of the program by allowing direct evaluation of efficiency. A more distant application of our program (after an accurate potential function has been obtained and after the program has evolved to require less structural information from sequence homology) might be to obtain detailed molecular structures from the schematic protein structures predicted using either pattern matching[33,34] or pattern recognition[35] techniques.

5 Comparison to Other Work

An alternative procedure for global energy minimization of surface loops has been described by Levinthal and coworkers.[6,7] Their procedure for searching the space of backbone conformations consists of a large number of local minimizations starting from a uniform distribution of loop deformations, molecular dynamics trajectories starting from the endpoints of the local minimizations, and local minimizations starting from the endpoints of the molecular dynamics trajectories. They report calculations using the CHARMM/GEMM computer program which has been programmed in assembly language to run efficiently on a Star ST-100 array processor. The Star ST-100 has a theoretical peak performance of 100 megaflops, approximately eight times faster than our FPS-5200. While many of the features of their procedure are similar to features of our procedure, there are also several differences.

Levinthal and coworkers have not attempted to account for the effects of hydration, except through the use of a distance-dependent dielectric constant. In contrast, we include a function representing hydration free energy at every stage of the calculations as a component of our total energy. Our potential function includes hydrogen atoms explicitly whereas the CHARMM/GEMM potential function uses united atoms. As a result, the function that we optimize requires at least four times more computation than the function that they optimize.

To obtain overlap-free backbone deformations, Levinthal and coworkers generate random deformations that span the entire loop, resulting in a uniform distribution of deformations, then discard those structures that contain overlaps. In contrast, we generate a nonuniform distribution of deformations which span the entire loop (when considered over several steps of the trajectory), then actively remove

overlaps from these structures. In both procedures, the time required to generate overlap-free structures is negligible in comparison to the time required to minimize the total energy of these structures. However, as a consequence of their conservative definition of overlap, their overlap-free structures will in general have very high energies whereas our overlap-free structures have negative values of the 12-2e energy.

Levinthal and coworkers use a conjugate gradient minimizer, which requires about 1,000 steps to minimize the energy of one conformation for a seven-residue surface loop. In contrast, we use a Newton minimizer, which in general requires fewer than 30 steps to minimize the energy of one overlap-free conformation for a seven-residue surface loop.

We have developed an automated procedure for efficiently generating low-energy side chain conformations whereas Levinthal and coworkers currently rely on an interactive procedure. They have not yet attempted to predict the native conformation based on minimization with respect to all the degrees of freedom associated with a loop. In contrast, after generating one low-energy side chain conformation for each low-energy backbone conformation, we minimize the energy with respect to all of the degrees of freedom associated with the loop to select the next point of the trajectory.

Levinthal and coworkers report global searches of the space of backbone conformations for surface loops containing five, seven, nine, and eleven residues. These four searches require 700 hours of computer time. It is difficult to make a direct comparison of the computer time required by the two procedures to search the space of backbone conformations for a surface loop of a given size because they have not reported the computer times required for individual searches. It is also difficult to make a theoretical comparison of the two procedures. For small peptides, a large number of minimizations from randomly chosen starting conformations is one of the least efficient methods for searching the space of conformations. However, we have no experience with small peptides that is relevant to a comparison of the efficiencies with which a molecular dynamics trajectory and a trajectory of local minima are able to search the space of conformations.

Other procedures for global energy minimization of surface loops have been described by Moult and James[8] and by Bruccoleri and Karplus.[4] A summary of the differences between these procedures and the procedure of Levinthal and coworkers has been presented elsewhere.[6]

A Generating Loop Deformations

In this Appendix, we describe our procedure for deforming a seven-residue segment of the initial backbone structure. A seven-residue segment of the initial backbone structure is deformed by obtaining alternative values for six consecutive (ψ, ϕ) pairs and the six intervening ω dihedral angles such that the ω values are set at 180° and such that the relative position and orientation of the nondeformed segments of the initial backbone structure are approximately unaltered. Backbone deformations that satisfy exactly the condition that the relative position and orientation of the nondeformed segments be unaltered will be referred to as exact loop deformations. Backbone deformations that satisfy this condition approximately will be referred to as approximate loop deformations.

Six dihedral angles can be assigned arbitrary values before the set of exact loop deformations that are consistent with these assigned values becomes discrete. After arbitrary values have been assigned to six dihedral angles, the problem of obtaining exact loop deformations requires solving a system of six equations expressing relationships between the values of the remaining six dihedral angles. In our procedure, three (ψ, ϕ) pairs will be assigned values from discrete collections of ordered pairs, and values for the remaining three (ψ, ϕ) pairs will be calculated analytically. The resulting system of equations is slightly less complicated when the values for three (ϕ, ψ) pairs are calculated analytically rather than the values for three (ψ, ϕ) pairs. A straightforward but lengthy procedure for solving this less complicated system of equations has been described by Gō and Scheraga.[11]

Our procedure introduces an approximate backbone geometry that considerably simplifies the calculations. As a consequence of this approximation, the resulting loop deformations are not quite exact.

However, the purpose of generating loop deformations is to obtain starting points for energy minimization, distance constraints force the relative position and orientation of the rigid segments of the protein to become exactly unaltered during energy minimization, and energy minimization is no more difficult for approximate loop deformations than for exact loop deformations. Generating approximate loop deformations requires less computation than generating exact loop deformations. However, even generating exact loop deformations requires much less computation than minimizing the energy of the resulting structures. We have chosen to generate approximate deformations mainly because the corresponding procedure is easier to understand and to program than the procedure for generating exact deformations.

A.1 Generating Cartesian Coordinates from Internal Coordinates

In this section, we describe the generation of cartesian coordinates from internal coordinates for a chain consisting of $N+3$ points. The cartesian coordinates of this chain are represented by the vector x. The set of indices corresponding to the points of the chain is chosen to be $\{-2,-1,0,\ldots,N\}$. If $\{x_{-2},x_{-1},\ldots,x_{m-1}\}$ have been specified, then x_m can easily be calculated if the length d_m of the $(m,m-1)$ bond, the angle λ_m formed by the $(m,m-1)$ and $(m-1,m-2)$ bonds, and the dihedral angle χ_m about the $(m-1,m-2)$ bond formed by the $(m,m-1)$ and $(m-2,m-3)$ bonds have been specified. The vectors d, λ, and χ are referred to as internal coordinates. The sets of indices of d, λ, and χ are $\{-1,0,\ldots,N\}$, $\{0,1,\ldots,N\}$, and $\{1,2,\ldots,N\}$, respectively. The range of values for each component of d, λ, and χ are $[0,\infty)$, $[0,\pi]$, and $[-\pi,\pi)$, respectively.

Consider a chain consisting of $N+3$ points for which the $3(N+3)-6$ internal coordinates have been specified. Since the internal coordinates are independent of translation and rotation of the chain as a whole, we can assume that the chain has been positioned and oriented such that

$$\begin{aligned} x_0 &= (\,0,\,0,\,0) \\ x_{-1} &= (\,-d_0,\,0,\,0) \\ x_{-2} &= (\,-d_0+d_{-1}\cos\lambda_0,\;d_{-1}\sin\lambda_0,\;0)\,. \end{aligned} \tag{A1}$$

The conformation that is defined by the specified bond lengths, by the specified value of λ_0, and by

$$\left.\begin{array}{rl} & \lambda_j = 180^\circ \\ \text{and} & \chi_j = 0^\circ \end{array}\right\} \quad \text{for} \quad j\in\{1,2,\ldots,N\} \tag{A2}$$

is a convenient reference conformation because cartesian coordinates can easily be determined for this conformation. For a conformation with arbitrary values of bond angles and dihedral angles, the cartesian coordinates can be generated from the cartesian coordinates of the reference conformation using two basic operations, rotation about the z-axis by $\pi-\lambda$ to set bond angles,

$$U(\lambda)=\begin{pmatrix} \cos(\pi-\lambda) & -\sin(\pi-\lambda) & 0 \\ \sin(\pi-\lambda) & \cos(\pi-\lambda) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{A3}$$

and rotation about the x-axis by χ to set dihedral angles,

$$R(\chi)=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\chi & -\sin\chi \\ 0 & \sin\chi & \cos\chi \end{pmatrix}. \tag{A4}$$

Conceptually, rotations begin with point N and work backward to point 1. For point m, the rotations $R(\chi_m)U(\lambda_m)$ about point $m-1$ are applied to points $\{m,m+1,\ldots,N\}$. It would be computationally inefficient to calculate cartesian coordinates using this sequence of rotations. However, once the equations have been determined, the computational efficiency can be improved by performing the component operations in a different order. The mapping of internal coordinates into cartesian coordinates for points $\{1,2,\ldots,N\}$ is described by the equations

$$
\begin{aligned}
x_1 &= R(\chi_1)U(\lambda_1)\begin{pmatrix} d_1 \\ 0 \\ 0 \end{pmatrix} \\
x_2 &= R(\chi_1)U(\lambda_1)\left\{\begin{pmatrix} d_1 \\ 0 \\ 0 \end{pmatrix} + R(\chi_2)U(\lambda_2)\begin{pmatrix} d_2 \\ 0 \\ 0 \end{pmatrix}\right\} \\
&= x_1 + R(\chi_1)U(\lambda_1)R(\chi_2)U(\lambda_2)\begin{pmatrix} d_2 \\ 0 \\ 0 \end{pmatrix} \\
x_N &= x_{N-1} + R(\chi_1)U(\lambda_1)\ldots R(\chi_N)U(\lambda_N)\begin{pmatrix} d_N \\ 0 \\ 0 \end{pmatrix} .
\end{aligned}
\tag{A5}
$$

For convenience, we define

$$
T_m = R(\chi_1)U(\lambda_1)\ldots R(\chi_m)\,. \tag{A6}
$$

We refer to T as the rotation matrix.

A.2 Expressing Relationships Between the Six (ψ, ϕ) Values

Consider the protein backbone shown in Figure A1. An approximate loop deformation will be generated by obtaining alternative values for the six labeled (ψ, ϕ) pairs and the six intervening ω dihedral angles such that the ω values are set at 180° and such that x_{16}, x_{17}, and x_{18} are approximately unaltered. The relationships between x_{16}, x_{17}, and x_{18} can be expressed as

$$
\begin{aligned}
x_{17} &= x_{16} + T_{17}U(\lambda_{17})\begin{pmatrix} d_{17} \\ 0 \\ 0 \end{pmatrix} \\
&= x_{16} + T_{18}\begin{pmatrix} d_{17} \\ 0 \\ 0 \end{pmatrix} \\
x_{18} &= x_{17} + T_{18}U(\lambda_{18})\begin{pmatrix} d_{18} \\ 0 \\ 0 \end{pmatrix} .
\end{aligned}
\tag{A7}
$$

Therefore, the six conditions that x_{16}, x_{17} and x_{18} remain unaltered are equivalent to the six conditions that x_{17} and T_{18} remain unaltered. For convenience, we define

$$
\begin{aligned}
x^{(1)} &= x_{17} \\
\text{and} \quad T^{(1)} &= T_{18}\,.
\end{aligned}
\tag{A8}
$$

These quantities, which will be referred to as the initial target coordinates and the initial target rotation matrix, are calculated using the ECEPP values for bond lengths and bond angles and the dihedral angles of the initial backbone structure.

The (ψ, ϕ) pairs labeled (ψ'_1, ϕ'_1) and (ψ''_1, ϕ''_1) in Figure A1 will be assigned values from discrete collections of (ψ, ϕ) values. These discrete collections consist of the grid points of a two-dimensional grid that has 40° separations between neighboring grid points intersected with the regions of the (ψ, ϕ) map that are observed to exist in protein crystal structures for the corresponding pair of amino acids. The (ψ, ϕ) pair labeled (ψ_4, ϕ_4) in Figure A1 will be assigned values from a discrete collection that depends on the previously assigned values of (ψ'_1, ϕ'_1) and (ψ''_1, ϕ''_1). This discrete collection consists of the grid points of a two-dimensional grid that has 20° separations between neighboring grid points intersected with the set of points for which solutions exist for the remaining three (ψ, ϕ) pairs.

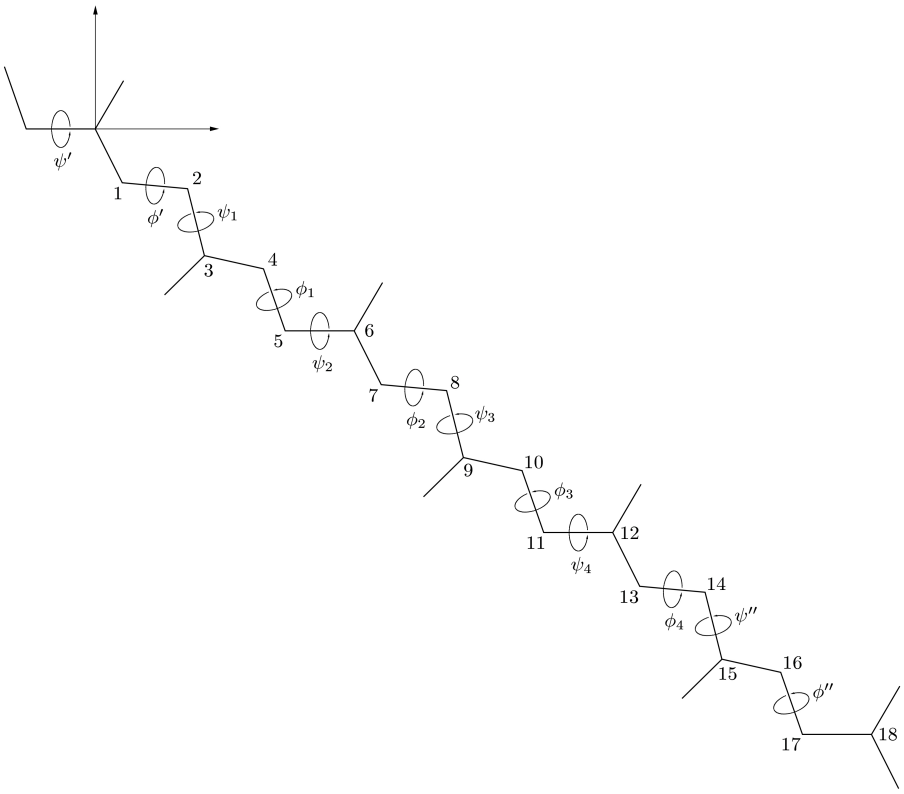

Figure A1: Definition of symbols used to denote dihedral angles contained in a collection of six consecutive (ψ, ϕ) pairs.

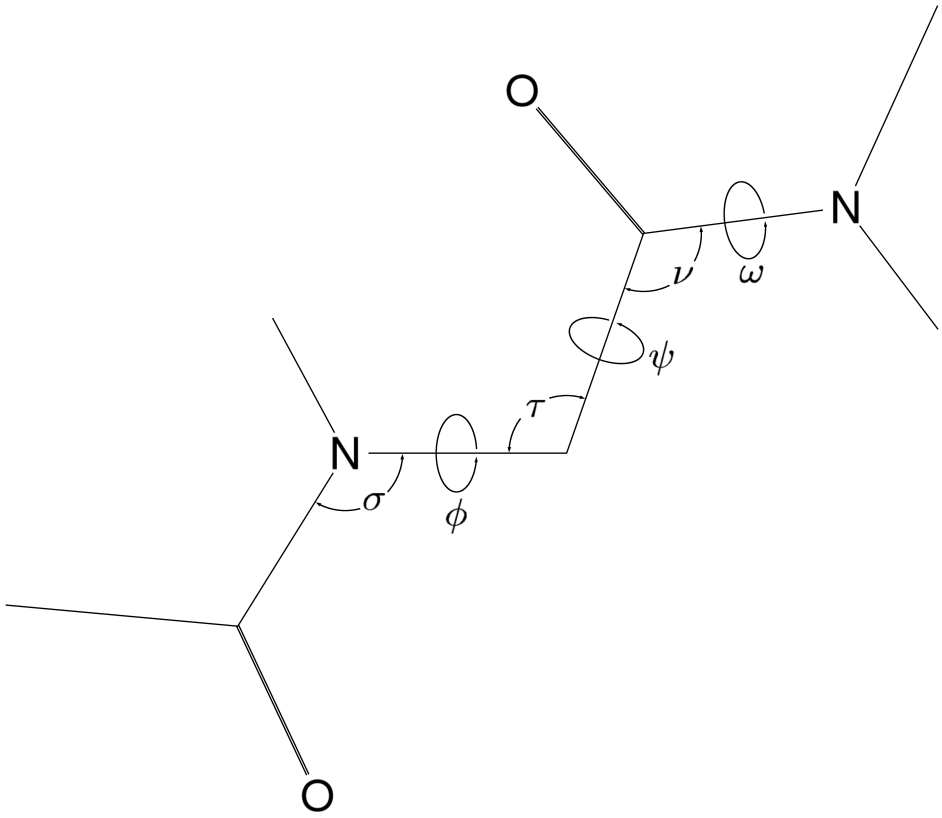

Figure A2: Definition of symbols used to denote bond angles of the protein backbone.

The notation that is used for the bond angles and dihedral angles of the protein backbone is defined in Figure A2. The contribution of each (ψ, ω, ϕ) triplet to the rotation matrix is

$$R(\psi_m)U(\nu)R(\omega_m)U(\sigma)R(\phi_m)U(\tau_m)\,. \tag{A9}$$

To simplify this sequence of rotations, we introduce the approximations that $\sigma = \nu = 118^\circ$ and that $\tau = 110^\circ$ for each amino acid. The ECEPP values for these bond angles are 121° for σ, 115° for ν, and between 108° and 111° for τ. Using these approximations, the contribution of each (ψ, ω, ϕ) triplet to the rotation matrix becomes

$$R(\psi_m + \phi_m + \pi)U(\tau)\,. \tag{A10}$$

For convenience, we define

$$\theta_m = \psi_m + \phi_m + \pi\,, \tag{A11}$$

$$\begin{aligned} U &= U(\tau) \\ R_j &= R(\theta_j)\,, \end{aligned} \tag{A12}$$

$$\begin{aligned} \text{and} \qquad b &= U(\tau) \begin{pmatrix} d_{(C^\alpha - C')} \\ 0 \\ 0 \end{pmatrix} \\ b' &= U(\nu) \begin{pmatrix} d_{(C' - N)} \\ 0 \\ 0 \end{pmatrix} + U(\nu) R(\pi) U(\sigma) \begin{pmatrix} d_{(N - C^\alpha)} \\ 0 \\ 0 \end{pmatrix} . \end{aligned} \tag{A13}$$

We use the symbol $\equiv$ to indicate that quantities following this symbol are being introduced and defined by quantities preceding this symbol that are grouped by parenthesis or braces. We use boldface type to distinguish vectors and matrices that depend only on degrees of freedom for which values have been assigned at some previous point in the procedure from vectors and matrices that depend on degrees of freedom for which values have not yet been assigned. We use a dagger to indicate a transposed matrix.

Using approximate backbone geometry, the relationships between the six (ψ, ϕ) values of a loop deformation can be expressed by the equations

$$\mathbf{T}^{(1)} = \Big(R_1' \mathbf{U}\Big)\Big(R_1 \mathbf{U} R_2 \mathbf{U} R_3 \mathbf{U} R_4\Big)\Big(\mathbf{U} R_1''\Big) \equiv A T^{(2)} B \tag{A14}$$

$$\begin{aligned} \mathbf{x}^{(1)} &= \left.\begin{array}{l} R(\psi_1')\mathbf{b}' \\ +R_1'\mathbf{b} \end{array}\right\} \\ &\quad \left.\begin{array}{l} +AR(\psi_1)\mathbf{b}' \\ +AR_1\mathbf{b} \\ +AR_1\mathbf{U}R(\psi_2)\mathbf{b}' \\ +AR_1\mathbf{U}R_2\mathbf{b} \\ +AR_1\mathbf{U}R_2\mathbf{U}R(\psi_3)\mathbf{b}' \\ +AR_1\mathbf{U}R_2\mathbf{U}R_3\mathbf{b} \\ +AR_1\mathbf{U}R_2\mathbf{U}R_3\mathbf{U}R(\psi_4)\mathbf{b}' \end{array}\right\} \\ &\quad \left.\begin{array}{l} +AT^{(2)}\mathbf{b} \\ +AT^{(2)}\mathbf{U}R(\psi_1'')\mathbf{b}' \end{array}\right\} \\ &\equiv \begin{array}{l} x' \\ +Ax^{(2)} \\ +AT^{(2)}x'' \, . \end{array} \end{aligned} \tag{A15}$$

A.3 Assigning Values of (ψ_1', ϕ_1'), and (ψ_1'', ϕ_1'') from Discrete Collections

Values are assigned to (ψ_1', ϕ_1') and (ψ_1'', ϕ_1'') from the corresponding discrete collections of (ψ, ϕ) values. The required relationships between the four (ψ, ϕ) values that remain to be assigned can be expressed by the equations

$$\mathbf{T}^{(2)} = R_1 \mathbf{U} R_2 \mathbf{U} R_3 \mathbf{U} R_4 \tag{A16}$$

$$\begin{aligned} \mathbf{x}^{(2)} &= R(\psi_1)\mathbf{b}' \\ &+ R_1\mathbf{b} \\ &+ R_1\mathbf{U}R(\psi_2)\mathbf{b}' \\ &+ R_1\mathbf{U}R_2\mathbf{b} \\ &+ R_1\mathbf{U}R_2\mathbf{U}R(\psi_3)\mathbf{b}' \\ &+ R_1\mathbf{U}R_2\mathbf{U}R_3\mathbf{b} \\ &+ R_1\mathbf{U}R_2\mathbf{U}R_3\mathbf{U}R(\psi_4)\mathbf{b}' \, , \end{aligned} \tag{A17}$$

which are obtained from equations A14 and A15 by eliminating terms that depend only on degrees of

freedom for which values have been assigned. Equation A16, which requires that the rotation matrix be unaltered by the deformation, expresses relationships between the values of θ_1, θ_2, θ_3 and θ_4.

A.4 Determining the Range of θ_4

In this section, we determine the set of values of θ_4 for which two distinct values of $(\theta_1, \theta_2, \theta_3)$ can be obtained by solving equation A16. For convenience, we define

$$\left.\begin{array}{rcl} c & = & \cos(\pi - \tau) \\ s & = & \sin(\pi - \tau) \\ c\theta_j & = & \cos(\theta_j) \\ s\theta_j & = & \sin(\theta_j) \end{array}\right\} \quad \text{for} \quad j \in \{1, 2, 3, 4\}\,. \tag{A18}$$

Equation A16 can be expressed in an alternative form as

$$\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger} = R_1 \mathbf{U} R_2 \mathbf{U} R_3\,, \tag{A19}$$

which can be written more explicitly as

$$\begin{pmatrix} t_{11}c - t_{12}sc\theta_4 + t_{13}ss\theta_4 & - & - \\ - & - & - \\ - & - & - \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c\theta_1 & -s\theta_1 \\ 0 & s\theta_1 & c\theta_1 \end{pmatrix} \begin{pmatrix} c^2 - s^2 c\theta_2 & -cs(1 + c\theta_2) & ss\theta_2 \\ cs(1 + c\theta_2) & -s^2 + c^2 c\theta_2 & -cs\theta_2 \\ ss\theta_2 & cs\theta_2 & c\theta_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & c\theta_3 & -s\theta_3 \\ 0 & s\theta_3 & c\theta_3 \end{pmatrix}. \tag{A20}$$

The $(1,1)$ matrix element of the matrix product that constitutes the right hand side of this equation depends only on θ_2.

It is useful to consider a geometric description of the sequence of rotations $R_1 \mathbf{U} R_2 \mathbf{U} R_3$. Figure A3 shows the effect of these rotations on a unit vector e that lies along the x-axis. The rotation R_3 maps e into itself. $\mathbf{U}$ rotates $R_3 e$ about the z-axis by 70°. R_2 rotates $\mathbf{U} R_3 e$ about the x-axis resulting in the set of possible values shown in Figure A3a. $\mathbf{U}$ rotates this set about the z-axis resulting in the set of values shown in Figure A3b. Finally, R_1 rotates this set about the x-axis. Therefore, the set of possible rotations $R_1 \mathbf{U} R_2 \mathbf{U} R_3$ map e onto all points of the unit sphere that have an x-component greater than or equal to $c^2 - s^2$. If the rotation $\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger}$ maps e onto any point of the unit sphere that has an x-component less than $c^2 - s^2$, then no values of $(\theta_1, \theta_2, \theta_3)$ exist that satisfy equation A19. If the rotation $\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger}$ maps e onto any point of the unit sphere that has an x-component greater than $c^2 - s^2$ and less than 1.0, then there exist exactly two values of $(\theta_1, \theta_2, \theta_3)$ that satisfy equation A19. In this case, there exist exactly two values of θ_2 such that the x-component of $R_1 \mathbf{U} R_2 \mathbf{U} R_3 e$ is equal to the x-component of $\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger} e$. For each solution θ_2, there exists a unique value of θ_1 such that the y and z-components of $R_1 \mathbf{U} R_2 \mathbf{U} R_3 e$ equal the y and z-components of $\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger} e$. For each solution (θ_1, θ_2), there exists a unique value of θ_3 such that $R_1 \mathbf{U} R_2 \mathbf{U} R_3$ maps the y and z axes into the same directions as $\mathbf{T}^{(2)} R_4^{\dagger} \mathbf{U}^{\dagger}$.

The set of θ_4 values for which solutions to equation A16 exist can be determined from the equation corresponding to the $(1,1)$ matrix element of equation A20,

$$t_{11}c - t_{12}sc\theta_4 + t_{13}ss\theta_4 \geq c^2 - s^2\,, \tag{A21}$$

which can be written more concisely as

$$\cos(\theta_4 - \alpha) \geq \delta\,, \tag{A22}$$

where

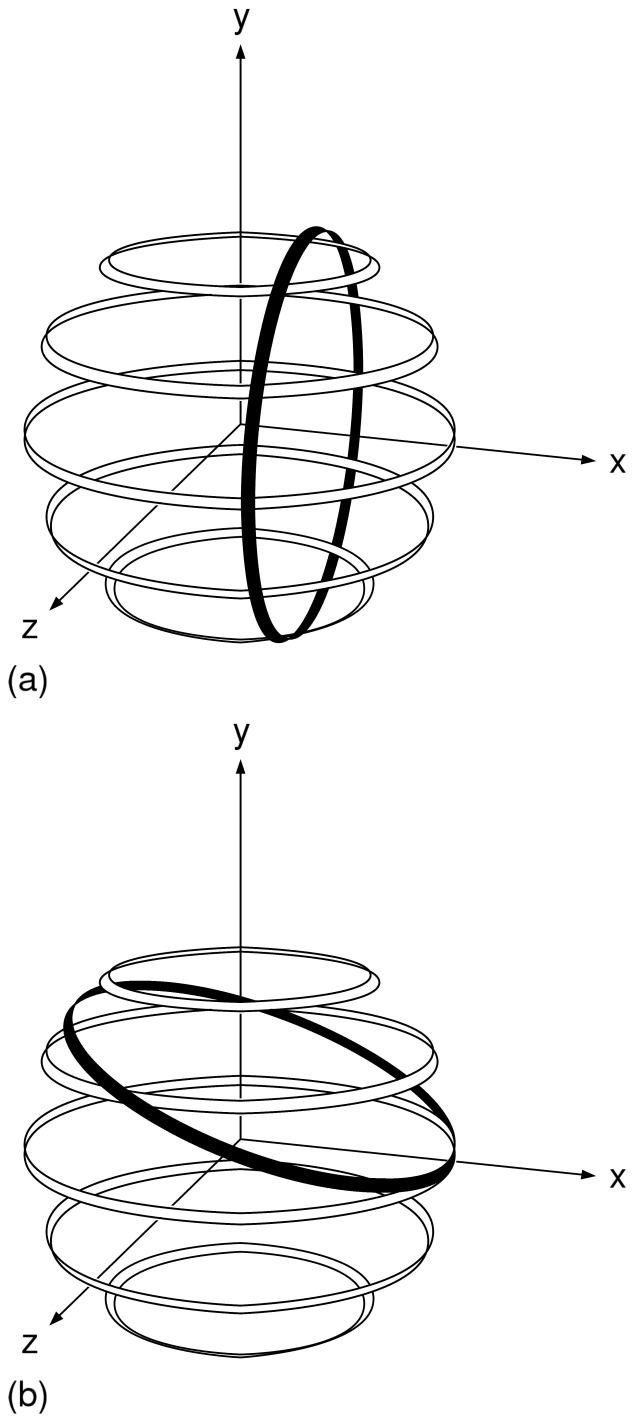

Figure A3: Subsets of points (indicated by heavy lines) on the unit sphere resulting from subsets of rotations applied to the unit vector along the x-axis. The subsets of rotations are described by a) $R_2\mathbf{U}R_3$, and b) $\mathbf{U}R_2\mathbf{U}R_3$ where $(\theta_1, \theta_2, \theta_3) \in [-\pi, \pi)^3$.

$$
\begin{aligned}
a &= \left((t_{12}s)^2 + (t_{13}s)^2 \right)^{\frac{1}{2}} \\
(\cos\alpha, \sin\alpha) &= \left(\frac{-t_{12}s}{a}, \frac{t_{13}s}{a} \right) \\
\delta &= \frac{(c^2 - s^2) - t_{11}c}{a} \, .
\end{aligned}
\tag{A23}
$$

The set of values of θ_4 for which two distinct values of $(\theta_1, \theta_2, \theta_3)$ can be obtained by solving equation A16 is determined using equation A22. A discrete collection of θ_4 values is selected from this set, and a value is assigned to θ_4 from this discrete collection.

A.5 Calculating $(\theta_1, \theta_2, \theta_3)$

Two values of $(\theta_1, \theta_2, \theta_3)$ are obtained by solving equation A16. Let

$$
\mathbf{T}^{(3)} = \mathbf{T}^{(2)} \mathbf{R}_4^{\dagger} \mathbf{U}^{\dagger} \, . \tag{A24}
$$

Then equation A19 can be written in the simplified form

$$
\mathbf{T}^{(3)} = R_1 \mathbf{U} R_2 \mathbf{U} R_3 \, . \tag{A25}
$$

The two solutions for θ_2 are obtained by solving the equation corresponding to the $(1,1)$ matrix element of equation A25,

$$
t_{11} = c^2 - s^2 c\theta_2 \, . \tag{A26}
$$

For each solution that is obtained for θ_2, the unique solution for (θ_1, θ_3) is obtained by solving equation A25.

A.6 Determining the Range of ψ_4

Equation A17, which requires that the position of atom 17 be unaltered by the deformation, expresses relationships between the values of ψ_1, ψ_2, ψ_3 and ψ_4. Since values have now been assigned to θ_1, θ_2, θ_3 and θ_4; equation A17 can be rewritten as

$$
\begin{aligned}
\mathbf{x}^{(2)} &= R(\psi_1)\mathbf{b}' \\
&+ \mathbf{R}_1\mathbf{b} \\
&+ \mathbf{R}_1\mathbf{U}R(\psi_2)\mathbf{b}' \\
&+ \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{b} \\
&+ \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{U}R(\psi_3)\mathbf{b}' \\
&+ \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{U}\mathbf{R}_3\mathbf{b} \\
&+ \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{U}\mathbf{R}_3\mathbf{U}R(\psi_4)\mathbf{b}' \, .
\end{aligned}
\tag{A27}
$$

The procedure that will be used to obtain solutions for $(\psi_1, \psi_2, \psi_3, \psi_4)$ is analogous to the procedure that was used to obtain solutions for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

In this section, we obtain a discrete collection of ψ_4 values and assign a value to ψ_4 from this discrete collection. Let

$$
\begin{aligned}
\mathbf{x}^{(3)} = \mathbf{x}^{(2)} &- \mathbf{R}_1\mathbf{b} \\
&- \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{b} \\
&- \mathbf{R}_1\mathbf{U}\mathbf{R}_2\mathbf{U}\mathbf{R}_3\mathbf{b}
\end{aligned}
\tag{A28}
$$

$$
\text{and} \qquad \mathbf{x}^{(4)} = \mathbf{U}^{\dagger}\mathbf{R}_1^{\dagger}\mathbf{x}^{(3)} \, . \tag{A29}
$$

For convenience, we define

$$
\mathbf{G} = \mathbf{R}_2\mathbf{U}\mathbf{R}_3\mathbf{U} \tag{A30}
$$

$$\begin{array}{rrcl} \text{and} & c1 & = & \cos(\psi_1 - \theta_1) \\ & s1 & = & \sin(\psi_1 - \theta_1) \\ & c\psi_j & = & \cos(\psi_j) \\ & s\psi_j & = & \sin(\psi_j) \end{array} \Bigg\} \quad \text{for} \quad j \in \{2,3,4\}\,. \tag{A31}$$

Equation A27 can be expressed in an alternative form as

$$\begin{aligned} \mathbf{x}^{(4)} - \mathbf{G}R(\psi_4)\mathbf{b}' &= \mathbf{U}^\dagger R(\psi_1 - \theta_1)\mathbf{b}' \\ &+ R(\psi_2)\mathbf{b}' \\ &+ \mathbf{R}_2\mathbf{U}R(\psi_3)\mathbf{b}'\,, \end{aligned} \tag{A32}$$

which can be written more explicitly as

$$\begin{aligned} &\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - b_1' \left\{ \begin{pmatrix} c \\ -s \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} c \\ sc\theta_2 \\ ss\theta_2 \end{pmatrix} + \begin{pmatrix} g_{11} \\ g_{21} \\ g_{31} \end{pmatrix} \right\} \\ &\qquad -b_2' \begin{pmatrix} g_{12}c\psi_4 + g_{13}s\psi_4 \\ g_{22}c\psi_4 + g_{23}s\psi_4 \\ g_{32}c\psi_4 + g_{33}s\psi_4 \end{pmatrix} \\ &= b_2' \left\{ \begin{pmatrix} sc1 \\ cc1 \\ s1 \end{pmatrix} + \begin{pmatrix} 0 \\ c\psi_2 \\ s\psi_2 \end{pmatrix} + \begin{pmatrix} -sc\psi_3 \\ cc\theta_2 c\psi_3 - s\theta_2 s\psi_3 \\ cs\theta_2 c\psi_3 + c\theta_2 s\psi_3 \end{pmatrix} \right\}. \end{aligned} \tag{A33}$$

Let

$$\mathbf{x}^{(5)} = \frac{1}{b_2'} \left[\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - b_1' \left\{ \begin{pmatrix} c \\ -s \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} c \\ sc\theta_2 \\ ss\theta_2 \end{pmatrix} + \begin{pmatrix} g_{11} \\ g_{21} \\ g_{31} \end{pmatrix} \right\} \right]\,. \tag{A34}$$

Then

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &- \begin{pmatrix} g_{12}c\psi_4 + g_{13}s\psi_4 \\ g_{22}c\psi_4 + g_{23}s\psi_4 \\ g_{32}c\psi_4 + g_{33}s\psi_4 \end{pmatrix} \\ &= \begin{pmatrix} sc1 \\ cc1 \\ s1 \end{pmatrix} + \begin{pmatrix} 0 \\ c\psi_2 \\ s\psi_2 \end{pmatrix} + \begin{pmatrix} -sc\psi_3 \\ cc\theta_2 c\psi_3 - s\theta_2 s\psi_3 \\ cs\theta_2 c\psi_3 + c\theta_2 s\psi_3 \end{pmatrix}. \end{aligned} \tag{A35}$$

The right hand side of equation A35 is the sum of three unit vectors, each of which is a function of only one ψ dihedral angle. For convenience, we define

$$\begin{array}{rrcl} & q_i & = & \left((g_{i2})^2 + (g_{i3})^2\right)^{\frac{1}{2}} \\ \text{and} & (\cos\eta_i, \sin\eta_i) & = & \left(\dfrac{g_{i2}}{q_i}, \dfrac{g_{i3}}{q_i}\right) \quad \text{for} \quad i \in \{1,2,3\}\,. \end{array} \tag{A36}$$

The discrete collection from which ψ_4 will be assigned values is reduced by requiring that each component of the left hand side of equation A35 be within the range of possible values for the right hand side. These requirements can be expressed by the equations

$$\begin{aligned}
\cos(\psi_4 - \eta_1) &\in \left[\frac{(x_1 - \delta_1)}{q_1}, \frac{(x_1 + \delta_1)}{q_1}\right] \\
\cos(\psi_4 - \eta_2) &\in \left[\frac{(x_2 - \delta_2)}{q_2}, \frac{(x_2 + \delta_2)}{q_2}\right] \\
\cos(\psi_4 - \eta_3) &\in \left[\frac{(x_3 - \delta_3)}{q_3}, \frac{(x_3 + \delta_3)}{q_3}\right],
\end{aligned} \tag{A37}$$

where

$$\begin{aligned}
\delta_1 &= s + s \\
\delta_2 &= c + 1 + \left((cc\theta_2)^2 + (s\theta_2)^2\right)^{\frac{1}{2}} \\
\delta_3 &= 1 + 1 + \left((cs\theta_2)^2 + (c\theta_2)^2\right)^{\frac{1}{2}}.
\end{aligned} \tag{A38}$$

These three equations determine three subsets of $[-\pi, \pi)$. The discrete collection from which ψ_4 will be assigned values is limited to the intersection of these three subsets. A value from this discrete collection is assigned to ψ_4.

A.7 Calculating (ψ_1, ψ_2, ψ_3)

The values of (ψ_1, ψ_2, ψ_3) are obtained by solving equation A35. Let

$$\mathbf{x}^{(6)} = \mathbf{x}^{(5)} - \begin{pmatrix} g_{12}c\psi_4 + g_{13}s\psi_4 \\ g_{22}c\psi_4 + g_{23}s\psi_4 \\ g_{32}c\psi_4 + g_{33}s\psi_4 \end{pmatrix}. \tag{A39}$$

Then equation A35 can be written in the simplified form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} s\cos(\psi_1 - \theta_1) \\ c\cos(\psi_1 - \theta_1) \\ \sin(\psi_1 - \theta_1) \end{pmatrix} + \begin{pmatrix} 0 \\ c\psi_2 \\ s\psi_2 \end{pmatrix} + \begin{pmatrix} -sc\psi_3 \\ cc\theta_2 c\psi_3 - s\theta_2 s\psi_3 \\ cs\theta_2 c\psi_3 + c\theta_2 s\psi_3 \end{pmatrix}. \tag{A40}$$

The first component of equation A40 can be used to express $(\psi_1 - \theta_1)$ as an explicit function of ψ_3,

$$\begin{aligned}
\cos(\psi_1 - \theta_1) &= \left(\frac{x_1}{s}\right) + c\psi_3 \\
&\equiv \mu_1 + c\psi_3 .
\end{aligned} \tag{A41}$$

For all values of ψ_3 such that $(\mu_1 + c\psi_3) \in (-1, 1)$, there exist two solutions of $(\psi_1 - \theta_1)$,

$$\big(\cos(\psi_1 - \theta_1), \sin(\psi_1 - \theta_1)\big) = \begin{cases} \left(\mu_1 + c\psi_3, \quad \left[1 - (\mu_1 + c\psi_3)^2\right]^{\frac{1}{2}}\right) \\ \left(\mu_1 + c\psi_3, \ -\left[1 - (\mu_1 + c\psi_3)^2\right]^{\frac{1}{2}}\right) \end{cases}. \tag{A42}$$

The second component of equation A40 can be used to express ψ_2 as an explicit function of ψ_3,

$$\begin{aligned}
c\psi_2 &= (x_2 - c\mu_1) - c(1 + c\theta_2)c\psi_3 + s\theta_2 s\psi_3 \\
&\equiv \mu_2 + h\cos(\psi_3 - \beta) .
\end{aligned} \tag{A43}$$

For all values of ψ_3 such that $(\mu_2 + h\cos(\psi_3 - \beta)) \in (-1, 1)$, there exist two solutions of ψ_2,

$$\big(c\psi_2, s\psi_2\big) = \begin{cases} \left(\mu_2 + h\cos(\psi_3 - \beta), \quad \left[1 - \big(\mu_2 + h\cos(\psi_3 - \beta)\big)^2\right]^{\frac{1}{2}}\right) \\ \left(\mu_2 + h\cos(\psi_3 - \beta), \ -\left[1 - \big(\mu_2 + h\cos(\psi_3 - \beta)\big)^2\right]^{\frac{1}{2}}\right) \end{cases}. \tag{A44}$$

These expressions for $(\psi_1 - \theta_1)$ and ψ_2 in terms of ψ_3 can then be substituted into the third component of equation A40 to give the four equations

$$\begin{aligned}(sgn1)\Big[1-(\mu_1+c\psi_3)^2\Big]^{\frac{1}{2}} \quad &+ \quad (sgn2)\Big[1-\big(\mu_2+h\cos(\psi_3-\beta)\big)^2\Big]^{\frac{1}{2}} \\ &+ \quad z\cos(\psi_3-\gamma)-x_3=0\,,\end{aligned} \tag{A45}$$

where (sgn1) and (sgn2) are equal to $+1$ or -1,

$$\begin{aligned} z &= \Big((cs\theta_2)^2+(c\theta_2)^2\Big)^{\frac{1}{2}} \\ \text{and} \qquad (\cos\gamma,\sin\gamma) &= \left(\frac{cs\theta_2}{z},\frac{c\theta_2}{z}\right)\,. \end{aligned} \tag{A46}$$

The set of ψ_3 values for which these functions are defined is the intersection of the subsets of $[-\pi,\pi)$ for which $(\psi_1 - \theta_1)$ and ψ_2 are defined. The zeros of these four equations are obtained using numerical methods.

A.8 Discussion

Our procedure for generating approximate loop deformations can easily be extended to allow generation of alternative values for seven or more (ψ,ϕ) pairs. When six (ψ,ϕ) pairs are used, the number of deformations that is generated sometimes exceeds 2,500. When seven (ψ,ϕ) pairs are used, the number of deformations that is generated is greater than the number of energy minimizations that is currently feasible. Therefore, the number of (ψ,ϕ) pairs for which alternative values are obtained has been limited to six in our global search procedure.

B Removing Overlaps

In this Appendix, we describe our procedure for removing overlaps. We define an overlap to be a 12-2e interaction with energy greater than 3.5 kcal/mol. Overlaps are eliminated (if possible) by minimizing the 12-2e interaction energy using a procedure that achieves efficiency by exploiting properties that are specific to the function being minimized.

Let X be a subset of degrees of freedom that controls the motion of a surface loop, let A be the set of protein atoms, and let $E \subset A\times A$ be the set of atom pairs for which interactions are calculated in local minimization of the total energy with respect to X (using our procedure for local minimization). The number of degrees of freedom in X, the number of atoms in A, and the number of atom pairs in E will be represented by $\#X$, $\#A$, and $\#E$, respectively. The values of the dihedral angles in X will be represented by the vector χ. Using this notation, the 12-2e interaction energy can be written explicitly as

$$f(\chi) = \sum_{(a,b)\in E} f_{(a,b)}(\chi) \tag{B1}$$

where

$$f_{(a,b)}(\chi) = \epsilon_{(T_a,T_b)}\left\{n\left(\frac{\rho_{(T_a,T_b)}}{r_{(a,b)}(\chi)}\right)^{12} - 2\left(\frac{\rho_{(T_a,T_b)}}{r_{(a,b)}(\chi)}\right)^{2e}\right\}\,. \tag{B2}$$

B.1 Extremely Small Cutoff Distance

For a conformation χ', let $E'(\chi')$ be the set of atom pairs $(a,b)\in E$ such that $r^2_{(a,b)}(\chi') < \rho^2_{(T_a,T_b)}$, and let $(\#E')$ be the number of atom pairs in E'. If the conformation χ' contains one or more overlaps, then

$$\sum_{(a,b)\in E} f_{(a,b)}(\chi) \approx \sum_{(a,b)\in E'(\chi')} f_{(a,b)}(\chi) \tag{B3}$$

holds for all conformations χ in a region of space that surrounds χ'. Each step of our procedure for eliminating overlaps consists of calculating $E'(\chi')$ and minimizing

$$\sum_{(a,b)\in E'(\chi')} f_{(a,b)}(\chi) \tag{B4}$$

plus a sum of harmonic distance constraints, within a region for which equation B3 is expected to hold, using Newton's method. This basic unit of computation, which will be referred to as a macrostep, is iterated either until all overlaps are eliminated or until the value of the function ceases to change. Typically, between 10 and 20 macrosteps are required to eliminate overlaps from the backbone deformations that are generated by our search procedure. At each macrostep, minimization of B4 plus a sum of harmonic distance constraints within a region for which equation B3 is expected to hold typically requires between 10 and 20 steps using Newton's method. These minimization steps will be referred to as microsteps. The extremely small cutoff distance $\rho_{(T_a,T_b)}$, which allows minimization steps to be calculated using extremely small collections of interactions, is the primary means by which the efficiency of minimization is increased.

B.2 $\frac{r^2}{2}$ Expansion

To increase further the efficiency of minimization, we approximate the sum B4 by a nearly equivalent sum for which derivatives can be obtained with less computation. For each atom pair $(a,b) \in E'(\chi')$, we replace $f_{(a,b)}$ by a nearly equivalent function in which $\frac{r^2_{(a,b)}}{2}$ is replaced by a second order Taylor expansion about χ'. For all $(a,b) \in E'(\chi')$, let $S_{(a,b)} \subset X$ be the set of $j \in X$ such that $r_{(a,b)}$ is dependent on the value of χ_j, and let $\#S_{(a,b)}$ be the number of dihedral angles in $S_{(a,b)}$. For all $j,k \in S_{(a,b)}$, let

$$\begin{aligned} s^{(a,b)}_j &= \frac{\partial}{\partial \chi_j}\left(\frac{r^2_{(a,b)}}{2}\right) \\ s^{(a,b)}_{jk} &= \frac{\partial^2}{\partial \chi_j \partial \chi_k}\left(\frac{r^2_{(a,b)}}{2}\right) . \end{aligned} \tag{B5}$$

For each $(a,b) \in E'(\chi')$, we replace $\frac{r^2_{(a,b)}}{2}$ by

$$\begin{aligned} q_{(a,b)}(\chi'+\delta\chi) &= \frac{r^2_{(a,b)}(\chi')}{2} \\ &+ \sum_{j\in S_{(a,b)}} s^{(a,b)}_j(\chi')\ \delta\chi_j \\ &+ \frac{1}{2}\sum_{j\in S_{(a,b)}}\sum_{k\in S_{(a,b)}} s^{(a,b)}_{jk}(\chi')\ \delta\chi_j\delta\chi_k , \end{aligned} \tag{B6}$$

we replace $f_{(a,b)}$ by

$$g_{(a,b)}(\chi) = \epsilon_{(T_a,T_b)}\left\{ n\left(\frac{\rho^2_{(T_a,T_b)}}{2q_{(a,b)}(\chi)}\right)^6 - 2\left(\frac{\rho^2_{(T_a,T_b)}}{2q_{(a,b)}(\chi)}\right)^e \right\} , \tag{B7}$$

and we replace the sum B4 by

$$\sum_{(a,b)\in E'(\chi')} g_{(a,b)}(\chi) . \tag{B8}$$

B.3 Algorithm

For all $j \in X$, let $\mathrm{BSE}_j \in A$ be the second atom of the bond corresponding to dihedral angle j. For all $(a,b) \in E'(\chi')$, let

$$\begin{aligned} A_{(a,b)} &= n\epsilon_{(T_a,T_b)}\left(\frac{\rho^2_{(T_a,T_b)}}{2}\right)^6 \\ B_{(a,b)} &= -2\epsilon_{(T_a,T_b)}\left(\frac{\rho^2_{(T_a,T_b)}}{2}\right)^e . \end{aligned} \tag{B9}$$

For all $j \in S_{(a,b)}$, let $c_j^{(a,b)}$ be a unit vector along the bond corresponding to dihedral angle j in the direction of the chain of bonds connecting atoms a and b, and let

$$\begin{aligned} \tau_j^{(a,b)} &= c_j^{(a,b)} \times (x_a - x_{\mathrm{BSE}_j}) \\ t_j^{(a,b)} &= c_j^{(a,b)} \times (x_b - x_{\mathrm{BSE}_j}) . \end{aligned} \tag{B10}$$

Using this notation, the coefficients of the Taylor expansion of $\frac{r^2_{(a,b)}}{2}$ about χ' can be written explicitly as

$$\begin{aligned} s_j^{(a,b)}(\chi') &= \big(x_b(\chi') - x_a(\chi')\big) \cdot t_j^{(a,b)}(\chi') \\ s_{jk}^{(a,b)}(\chi') &= \tau_j^{(a,b)}(\chi') \cdot t_k^{(a,b)}(\chi') \end{aligned} \tag{B11}$$

where j precedes k in the chain of bonds connecting atoms a and b.

Each macrostep consists of calculating $E'(\chi')$; calculating $\frac{r^2_{(a,b)}(\chi')}{2}$, $A_{(a,b)}$, $B_{(a,b)}$, and $S_{(a,b)}$ for each atom pair $(a,b) \in E'(\chi')$; calculating $s_j^{(a,b)}(\chi')$, $\tau_j^{(a,b)}(\chi')$, and $t_j^{(a,b)}(\chi')$ for each dihedral angle $j \in S_{(a,b)}$ for each $(a,b) \in E'(\chi')$; and minimizing B8 plus a sum of harmonic distance constraints using a sequence of microsteps to locate the minimum within a region of space for which there is high probability that

$$\sum_{(a,b)\in E} f_{(a,b)}(\chi) \approx \sum_{(a,b)\in E'(\chi')} g_{(a,b)}(\chi) . \tag{B12}$$

This region will be referred to as the trust region. The value of the radius of the trust region is assigned at the start of a minimization and is modified after each macrostep by an amount that depends on the difference between the predicted energy change and the actual energy change. To determine $E'(\chi')$, we calculate $r^2_{(a,b)}(\chi')$ for each atom pair $(a,b) \in E$.

Each microstep consists of calculating the function, first derivative, and second derivative of B8 plus a sum of harmonic distance constraints and stepping to the minimum of the resulting second order Taylor expansion within a microstep trust region. The function, first derivative, and second derivative of the sum B8 are calculated using

$$\begin{aligned} p &= \frac{A_{(a,b)}}{\big(q_{(a,b)}\big)^6} + \frac{B_{(a,b)}}{\big(q_{(a,b)}\big)^e} \equiv p_{12} + p_{2e} \\ p' &= \left(\frac{-6}{q_{(a,b)}}\right) p_{12} + \left(\frac{-e}{q_{(a,b)}}\right) p_{2e} \equiv p'_{12} + p'_{2e} \\ p'' &= \left(\frac{-7}{q_{(a,b)}}\right) p'_{12} + \left(\frac{-(e+1)}{q_{(a,b)}}\right) p'_{2e} \end{aligned} \tag{B13}$$

and

$$
\begin{aligned}
g_{(a,b)}(\chi' + \delta\chi) &= p \\
\frac{\partial g_{(a,b)}(\chi' + \delta\chi)}{\partial \chi_j} &= p' \left(s_j^{(a,b)} + \sum_{l \in S_{(a,b)}} s_{jl}^{(a,b)} \delta\chi_l \right) \\
\frac{\partial^2 g_{(a,b)}(\chi' + \delta\chi)}{\partial \chi_j \partial \chi_k} &= p'' \left(s_j^{(a,b)} + \sum_{l \in S_{(a,b)}} s_{jl}^{(a,b)} \delta\chi_l \right) \left(s_k^{(a,b)} + \sum_{l \in S_{(a,b)}} s_{kl}^{(a,b)} \delta\chi_l \right) \\
&+ p' s_{jk}^{(a,b)} \, .
\end{aligned}
\tag{B14}
$$

Because of the structure of $s_{jk}^{(a,b)}$ given by equation B11, the calculation of $q_{(a,b)}(\chi' + \delta\chi)$ using equation B6 requires at most time proportional to $\#X$. The function, first derivative, and second derivative for the sum of harmonic distance constraints are calculated using no approximations.

B.4 Results and Conclusions

In order to relate the ideas of our procedure to the efficiency of minimization, we consider results that were obtained for a nine-residue surface loop of trypsin. In the case that is considered, the 12-2e interaction energy was minimized with respect to the 27 backbone dihedral angles of the loop residues, and interactions involving side chains were not included in the calculations.

For trypsin, $\#A \approx 4{,}000$. Using a cutoff distance of 18Å, $\#E \approx 115{,}000$. $\#E' \approx 160$, which is approximately three times the number of backbone atoms in the surface loop. In fact, many of these atom pairs correspond to 1–4 interactions, which contribute little to the computer time that is required to calculate the derivatives at each microstep. The time required to calculate the first derivative at each microstep is proportional to

$$
\sum_{(a,b) \in E'} \#S_{(a,b)} \, , \tag{B15}
$$

and the time required to calculate the second derivative at each microstep is proportional to

$$
\sum_{(a,b) \in E'} \frac{\left(\#S_{(a,b)} \right)^2}{2} \, . \tag{B16}
$$

For the case that is being considered, these quantities have values of ≈ 600 and $\approx 4{,}500$ respectively.

The time required to determine E' represents about $1/2$ of the time required to complete a macrostep. The time required to calculate $\frac{r^2_{(a,b)}}{2}$, $A_{(a,b)}$, $B_{(a,b)}$, $S_{(a,b)}$, $s_j^{(a,b)}$, $\tau_j^{(a,b)}$, and $t_j^{(a,b)}$ for each atom pair $(a,b) \in E'$ and for each dihedral angle $j \in S_{(a,b)}$ is negligible. The time required to calculate the function, first derivative, and second derivative of B8 for a sequence of microsteps represents about $1/6$ of the time required to complete a macrostep. The time required to calculate the function, first derivative, and second derivative of the harmonic distance constraints for a sequence of microsteps is negligible. The time required to calculate a sequence of microsteps using Newton's method represents about $1/3$ of the time required to complete a macrostep.

For small subsets of degrees of freedom such as those used in the global search procedure, the $\frac{r^2}{2}$ expansion is at least moderately effective. The procedure that we use to calculate the second derivative of the total energy requires time proportional to $\frac{(\#X)^3}{6}$. When minimizing the total energy, the calculations requiring time proportional to $\frac{(\#X)^3}{6}$ are insignificant in comparison to the calculations requiring time proportional to $\#E$. When minimizing B4 using the same procedure, calculation of the second derivative is a computational bottleneck. The time required to calculate the function, first derivative, and second derivative at each microstep is reduced by a factor of about eight when B4 is replaced by B8. However, the $\frac{r^2}{2}$ expansion reduces the time required to complete a macrostep by only a factor of two because much of the required computation is not affected by this approximation. The $\frac{r^2}{2}$ expansion might be

less effective if the procedure used to calculate the second derivative of the total energy required time proportional to $\frac{(\#X)^2}{2}$.[36]

B.5 Discussion

The Taylor series expansion of $f_{(a,b)}$ about χ' can be written as

$$\begin{aligned} f_{(a,b)}(\chi' + \delta\chi) &= f_{(a,b)}(\chi') \\ &+ \sum_{j \in S_{(a,b)}} \frac{\partial f_{(a,b)}(\chi')}{\partial \chi_j} \delta\chi_j \\ &+ \frac{1}{2} \sum_{j \in S_{(a,b)}} \sum_{k \in S_{(a,b)}} \frac{\partial^2 f_{(a,b)}(\chi')}{\partial \chi_j \partial \chi_k} \delta\chi_j \delta\chi_k \\ &+ \frac{1}{6} \sum_{j \in S_{(a,b)}} \sum_{k \in S_{(a,b)}} \sum_{l \in S_{(a,b)}} \frac{\partial^3 f_{(a,b)}(\chi')}{\partial \chi_j \partial \chi_k \partial \chi_l} \delta\chi_j \delta\chi_k \delta\chi_l \\ &+ \cdots . \end{aligned} \tag{B17}$$

For all $(a,b) \in E'(\chi')$, the expansion of $f_{(a,b)}$ to second order in $\delta\chi$ is accurate over a relatively small region of space. The derivatives of $f_{(a,b)}$ increase in magnitude as the order increases. Therefore, contributions from terms of third order and higher become important for relatively small values of $\delta\chi$.

Expansion of $f_{(a,b)}$ to higher order in $\delta\chi$ would not be computationally efficient even if the time required for microsteps on the resulting surface was negligible. We will refer to the region of space for which a Taylor expansion to order n accurately approximates the function that is being expanded as the nth order trust region. The computation required to calculate third and fourth derivatives is on the order of $\frac{(\#X)^2}{12}$ times larger than the computation required to calculate first and second derivatives. Expansion of $f_{(a,b)}$ to fourth order is equivalent to expansion of $f_{(a,b)}$ to second order and expansion of each component of the second derivative of $f_{(a,b)}$ to second order. The second-order trust region for the second derivative of $f_{(a,b)}$ is approximately equal to the second-order trust region for $f_{(a,b)}$. If the second derivative of $f_{(a,b)}$ is accurate at the radius of the second order trust region for $f_{(a,b)}$, then $f_{(a,b)}$ will be accurate to approximately twice this radius. In other words, the fourth order trust region for $f_{(a,b)}$ has approximately twice the radius of the second order trust region for $f_{(a,b)}$. Therefore, the volume of the fourth order trust region is $\approx 2^{(\#X)}$ times larger than the volume of the second order trust region. Unfortunately, only a small fraction of this information is useful for local minimization. Two minimization steps based on expansions to second order result in movement of approximately the same distance using information about $\approx \frac{2}{2^{(\#X)}}$ as much volume.

The expansion of $\frac{r^2}{2}$ to second order in $\delta\chi$ is accurate over a region that is larger than the region for which equation B3 is valid. In other words, at each macrostep, the number of possible microsteps is limited by the requirement that E' be updated rather than by the requirement that the $\frac{r^2}{2}$ expansion coefficients be updated. The derivatives of $\frac{r^2}{2}$ decrease in magnitude as the order increases. Therefore, contributions from terms of third order and higher do not become important until $\delta\chi$ becomes relatively large. In fact, expansion of $\frac{r^2}{2}$ to first order is also accurate over a region larger than the region for which equation B3 is valid. A second order $\frac{r^2}{2}$ expansion requires approximately three times more computation than a first order $\frac{r^2}{2}$ expansion. These expansions give virtually indistinguishable results for all macrosteps with the exception of the final macrostep in cases where one or a few atom pairs have interaction energies close to 3.5 kcal/mol at the local minimum being approached. In these cases, the extra accuracy of the second order $\frac{r^2}{2}$ expansion sometimes allows the value of the function to be lowered slightly farther. We use a second order $\frac{r^2}{2}$ expansion because the resulting increase in accuracy requires an acceptable amount of additional computation.

Including the $\frac{1}{r^{2e}}$ term in the minimization allows the same definition of an overlap to be used for both

hydrogen bonding and non-hydrogen bonding atom pairs. Let ρ_{min} and ϵ_{min} be the position and depth of the 12-10 potential for a hydrogen bonding atom pair (a, b). These quantities are related to $\rho_{(T_a,T_b)}$ and $\epsilon_{(T_a,T_b)}$ by the equations

$$\left(\frac{\rho_{(T_a,T_b)}}{\rho_{min}}\right)^2 = \frac{5}{3} \tag{B18}$$

$$\epsilon_{min} = -\epsilon_{(T_a,T_b)} \left(\frac{1}{3}\right) \left(\frac{5}{3}\right)^5 . \tag{B19}$$

If equation B2 was replaced by

$$f_{(a,b)}(\chi) = \epsilon_{(T_a,T_b)} n \left(\frac{\rho_{(T_a,T_b)}}{r(\chi)}\right)^{12}, \tag{B20}$$

then the value of $f_{(a,b)}$ for a hydrogen bonding atom pair (a, b) separated by a distance $r_{(a,b)} = \rho_{min}$ would be $-5\epsilon_{min}$. Therefore, a different definition of an overlap would be needed for hydrogen bonding and non-hydrogen bonding atom pairs.

Our procedure for removing overlaps requires about four times less computation than our procedure for minimizing the total energy of the resulting overlap free structure. Therefore, further reduction of the required computation is not a high priority.

Acknowledgment

We thank M. Vásquez, M. H. Lambert, and E. O. Purisima for useful discussions throughout the project. We also thank E. O. Purisima and M. H. Lambert for graphics software support. This work was supported by a research grant from the National Institute of Heart, Lung, and Blood Diseases (HL-30616) of the National Institutes of Health.

References

[1] M. J. Dudek, and H. A. Scheraga, *work in progress*, Paper II in this series.

[2] M. J. Dudek, and H. A. Scheraga, *work in progress*, Paper III in this series.

[3] K. D. Gibson, and H. A. Scheraga, in *Structure and Expression: Vol. 1: From Proteins to Ribosomes*, eds. M. H. Sarma, and R. H. Sarma, Adenine Press, Guilderland, N. Y. 67–94 (1988).

[4] R. E. Bruccoleri, and M. Karplus, *Biopolymers*, **26**, 137–168 (1987).

[5] R. E. Bruccoleri, E. Haber, and J. Novotný, *Nature*, **335**, 564–568 (1988).

[6] R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal, *Proteins*, **1**, 342–362 (1986).

[7] P. S. Shenkin, D. L. Yarmush, R. M. Fine, H. Wang, and C. Levinthal, *Biopolymers*, **26**, 2053–2085 (1987).

[8] J. Moult, and M. N. G. James, *Proteins*, **1**, 146–163 (1986).

[9] T. A. Jones, and S. Thirup, *EMBO J.*, **5**, 819–822 (1986).

[10] T. Blundell, D. Carney, S. Gardner, F. Hayes, B. Howlin, T. Hubbard, J. Overington, D. A. Singh, B. L. Sibanda, and M. Sutcliffe, *Eur. J. Biochem.*, **172**, 513–520 (1988).

[11] N. Gō, and H. A. Scheraga, *Macromolecules*, **3**, 178–187 (1970).

[12] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361–2381 (1975).

[13] G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883–1887 (1983).

[14] M. J. Dudek, Ph.D. Thesis, Cornell University, Ithaca, NY, (1989).

[15] J. S. Richardson, *Adv. Protein Chem.*, **34**, 167–339 (1981).

[16] W. F. van Gunsteren, and H. J. C. Berendsen, *J. Computer-Aided Molecular Design*, **1**, 171–176 (1987).

[17] M. L. Connolly, *J. Appl. Cryst.*, **16**, 548–558 (1983).

[18] Z. I. Hodes, G. Némethy, and H. A. Scheraga, *Biopolymers*, **18**, 1565–1610 (1979).

[19] Y. K. Kang, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **91**, 4105–4109 (1987).

[20] M. Vásquez, and H. A. Scheraga, *Biopolymers*, **24**, 1437–1447 (1985).

[21] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, **112**, 535–542 (1977).

[22] S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga, *Macromolecules* **10**, 1–9 (1977).

[23] A. Wlodawer, J. Walter, R. Huber, and L. Sjolin, *J. Mol. Biol.* **180**, 301–329 (1984).

[24] J. Walter, W. Steigemann, T. P. Singh, H. Bartunik, W. Bode, and R. Huber, *Acta Cryst.*, **B 38**, 1462–1472 (1982).

[25] P. K. Warme, F. A. Momany, S. V. Rumball, R. W. Tuttle, and H. A. Scheraga, *Biochemistry*, **13**, 768–782 (1974).

[26] J. Greer, *J. Mol. Biol.*, **153**, 1027–1042 (1981).

[27] B. Furie, D. H. Bing, R. J. Feldmann, D. J. Robison, J. P. Burnier, and B. C. Furie, *J. Biol. Chem.*, **257**, 3875–3882 (1982).

[28] W. Braun, and N. Gō, *J. Mol. Biol.*, **186**, 611–626 (1985).

[29] T. F. Havel, and K. Wuthrich, *Bull. Math. Biol.*, **46**, 673–698 (1984).

[30] D. Hall, and N. Pavitt, *J. Comp. Chem.*, **5**, 441–450 (1984).

[31] M. Whitlow, and M. M. Teeter, *J. Am. Chem. Soc.* **108**, 7163–7172 (1986).

[32] W. L. Jorgensen, and J. Tirado-Rives, *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).

[33] F. E. Cohen, P. A. Kosen, I. D. Kuntz, L. B. Epstein, T. L. Ciardelli, and K. A. Smith, *Science*, **234**, 349–352 (1986).

[34] F. E. Cohen, and I. D. Kuntz, *Proteins*, **2**, 162–166 (1987).

[35] M. H. Lambert, and H. A. Scheraga, *J. Comp. Chem.*, **10**, 770, 798, 817 (1989).

[36] H. Abe, W. Braun, T. Noguti, and N. Gō, *Computers & Chemistry*, **8**, 239–247 (1984).