Published in final edited form as:

J Comput Chem. 2011 August ; 32(11): 2369–2385. doi:10.1002/jcc.21817.

A Versatile Method for Systematic Conformational Searches: Application to CheY

Robert J. Petrella

Department of Chemistry and Chemical Biology, Harvard University

Department of Medicine, Harvard Medical School

Abstract

A novel molecular structure prediction method, the Z Method, is described. It provides a versatile platform for the development and use of systematic, grid-based conformational search protocols, in which statistical information (i.e., rotamers) can also be included. The Z Method generates trial structures by applying many changes of the same type to a single starting structure, thereby sampling the conformation space in an unbiased way. The method, implemented in the CHARMM program as the Z Module, is applied here to an illustrative model problem in which rigid, systematic searches are performed in a 36-dimensional conformational space that describes the relative positions of the ten secondary structural elements of the protein CheY. A polar hydrogen representation with an implicit solvation term (EEF1) is used to evaluate successively larger fragments of the protein generated in a hierarchical build-up procedure. After a final refinement stage, and a total computational time of about two-and-a-half days on AMD Opteron processors, the prediction is within 1.56 Å of the native structure. The errors in the predicted backbone dihedral angles are found to approximately cancel. Monte Carlo and simulated annealing trials on the same or smaller versions of the problem, using the same atomic model and energy terms, are shown to result in less accurate predictions. Although the problem solved here is a limited one, the findings illustrate the utility of systematic searches with atom-based models for macromolecular structure prediction and the importance of unbiased sampling in structure prediction methods.

Keywords

conformational search; molecular conformation; structure prediction; energy function; search algorithm; dihedral angle

2 Introduction

Because the native three-dimensional structure of a protein usually corresponds to the thermodynamically stable state[1], it should be calculable directly from the amino acid sequence, provided a sufficiently accurate representation of the energy surface and enough computational resources. Attempts at “template-free,” “*ab initio*,” or “*de novo*” tertiary structure predictions for whole protein molecules have met with some success, particularly for smaller proteins[2, 3, 4, 5, 6, 7, 8, 9] and the use of all-atom physical potential energy functions for late-stage refinements of protein structures generated with other prediction methods is well-established[10, 11, 12, 13, 14, 15]. To date, however, the most successful approaches to whole-protein fold prediction, particularly for large proteins, have typically relied more heavily on comparative modeling (and statistical database information) than physics[16, 17, 18, 19]. A major obstacle in physics-based prediction for large systems, in which structures are ranked by energy, is that the size of a conformational space grows exponentially with the number of degrees of freedom (DOF), so that for long polymeric sequences such as the polypeptides that fold into native proteins, exhaustive conformational

searches using atom-based models are usually impractical. Two techniques for expanding the range of sampled conformations are reducing the number of DOF by simplifying the protein models[3, 8, 20, 21, 22, 23, 24] and sampling the conformations randomly rather than systematically[3, 8, 23, 25, 26, 27]. In Rosetta template-free predictions, for example, sets of preliminary predictions are often made with simplified side chain representations and knowledge-based (i.e., database-derived) scoring functions[2, 25, 13, 9]. Atomic-level detail is added in subsequent stages, and conformational searches rely on simulated annealing and Monte Carlo (i.e., random) minimization procedures[28, 29].

For a number of limited problems in macromolecular structure prediction, extensive, systematic searches with atom-based energy evaluations of conformations have also been used. In an analysis of prediction accuracy, an atomic potential energy function, CHARMM[30, 31], was used in thorough, systematic grid-based searches of conformational space for individual protein side chains[32]. It was shown that a significant fraction of the errors in side chain prediction methods were not directly related to the (vacuum) potential energy function; instead, they related to factors such as solvent and crystal contacts and incompleteness of the conformational searches. Xiang and Honig[33] demonstrated that accurate side chain prediction could be achieved with systematic, grid-based searches performed serially on single residues, provided the searches were locally extensive (large numbers of conformations per side chain) and the procedure was iterated until converged. Systematic grid-based searches using an atomic energy function were also used to demonstrate that “off-rotamer” orientations, which are statistically unlikely side chain conformations found in high-resolution protein crystal structures, were often not artifacts, but real[34]. In addition, such an approach has been used in hierarchical build-up procedures for the prediction of protein loop structures[35, 36].

This methodology has also been applied to predict tertiary structure in a set of model problems[37]. For the signal transduction protein CheY, which consists of 5 α helices and 5 β strands, and two other proteins (parvalbumin and FNIII₁₀), extensive and systematic searches were carried out with the EEF1 function[38], which uses an explicit atomic representation for the protein and a continuum representation for the aqueous solvent. The internal structure of the secondary structural elements (helices and strands, or SSEs) was assumed to be known, and the positions of the elements were varied relative to one another on multidimensional grids by varying single dihedral angles in the intervening loops; for CheY, since there are 9 loops between the 10 SSEs, the search problem was 9-dimensional. The correct tertiary packing arrangements for the three proteins were predicted to within 0.24 Å RMSD. This suggested that in each case, the global minimum of the EEF1 hypersurface had been found and that it corresponded to the native structure. The results of these studies have shown that at least for some structure prediction problems, systematic searches with atomic potential energy functions can be successfully applied; they also suggest that there are some inherent advantages to this approach.

The main purpose of the current work is to describe the Z Method, which facilitates, extends, and generalizes the published systematic search procedures, and to outline the principles that underlie its design. The Z Method, which is implemented in the CHARMM program[30, 31] as the “Z Module,” is based on the observation in several of the above-mentioned studies that the best prediction results are obtained by generating large numbers of conformations in “parallel”, i.e., by applying the same type of change repeatedly to a common starting structure, and then selecting the resulting structures of the lowest energy. By contrast, the sampling and optimization in “serial-generation” procedures, such as Markov-chain-like procedures[39] or “model-based search”[40], are carried out in such a way that each structure is generated from, and compared to, its immediate predecessor, not a common precursor. The parallel-generation approach results in even sampling—i.e., the

degree of optimization around each conformer is the same; this is important because unequal sampling in structure prediction may introduce systematic errors, as is detailed in a separate paper[41].

The Z Module is built around a systematization of conformational search and, within this framework, provides a versatile platform for carrying out systematic protocols of different kinds. The script-based systematic search procedures carried out previously and described above would correspond to a small fraction of the Z Module's capabilities. The method breaks down the conformation space of the system into component parts, or subspaces; the subspaces can be searched independently or in combinations. Searches can be carried out according to regularly spaced grids, predetermined libraries of conformers (randomly selected or otherwise), or combinations of both. The method can be used for molecular docking and other types of molecular structure prediction, including but not limited to side chain prediction, loop prediction, secondary and tertiary structure prediction, and local structural refinements.

Since the distribution of the search is entirely specified in the Z Method and does not depend on the shape of the energy surface, there is no risk that important regions in the search will be skipped, that the search will be trapped in metastable regions, or that it will revisit the same regions. Because it handles conformational subspaces, the method can be used to carry out searches for hierarchical build-up procedures (see below). It also facilitates the incorporation of statistical information, so that, for example, rotamer libraries for protein side chains can be used in searches.

The Z Module in CHARMM has been used previously to predict the structure of a protein/DNA complex for which an experimental structure was unavailable—namely, that of the human cytomegalovirus processivity factor UL44 complexed with a DNA oligonucleotide[42]. In that study, the Z Module was used for docking the two macromolecules with side chain flexibility, modeling the flexible (crystallographically disordered) loops of the UL44 protein, and carrying out some local structural refinements of the x-ray data. The resulting set of models for the complex was consistent with data from crosslinking and mutational experiments that were run concurrently. In addition, the results of the study suggested both a protein/DNA binding mechanism and a mechanism for the diffusion of the UL44 protein relative to the DNA that involved a spiral ionic track.

Although the Z Module was briefly described in that work, here both the method and an illustrative application are described in detail. The application extends the 9-dimensional search used previously for CheY[37] to 36 dimensions and illustrates the Z Modules's ability to use and generate libraries of conformers, as well as carry out molecular fragment build-up procedures. CheY is an interesting target for this type of prediction study because it has both α and β structure and contacting strands in its β sheet that are non-sequential. It is also monomeric and appropriately sized (128 residues, 10 SSEs). In the present work, each of the 9 sequentially contiguous pairs of helices or strands in the CheY protein is allowed to rotate rigidly about four dihedral angles in each intervening loop, providing a large sampling of the relative positions of the SSEs. As in the previous study, the internal structures of the SSEs are kept fixed and the EEFl function is used for conformer evaluation. However, because of the exponential increase in the size of the conformational space with the number of DOF, the 36-dimensional search problem addressed here is far larger than the previously studied 9-dimensional one, (at a 10° grid spacing, $36^{27} \approx 10^{42}$ times larger.) Hence, the prediction is carried out in a hierarchical fashion, in which successively larger fragments of the protein are generated from combinations of smaller, low-energy ones, much like the "build-up" procedure introduced by Scheraga and coworkers [43, 44] in their 1988 study of whole proteins and later used by Jacobson *et al.* for protein loops[35]. The ZAM method of

Dill *et al.* uses a similar approach, but with fragment conformers obtained from MD simulations[45]. Some *ab initio* protein prediction methods such as conventional Rosetta[46, 47] use fragment assembly (Monte Carlo-based fragment insertion), but they are not hierarchical, while other methods such as LINUS[23] are hierarchical (weighted conformer probabilities), but do not use fragment assembly. The hierarchical methods, including the one used in the present study, have aspects that are similar to “dead-end rotamer elimination” [48, 49, 50] strategies, in that they apply energy criteria to select structures for subsequent stages of the procedure, effectively reducing the size of the overall search.

Here, the searches begin with libraries of single-residue backbone angles that are based on statistically derived libraries [35] (see Methods section), but no other statistical information is used in the searches—after the initial stage the procedure essentially creates its own structural libraries for each successively larger protein fragment. After a final refinement or “reworking” stage, in which the starting structure is taken to be the final, low-energy structure from the build-up procedure, the structure is predicted to within 1.56 Å RMSD of the native structure. The total time for all the calculations is ~ 2.5 CPU days on a 1.6 GHz AMD Opteron machine. In the build-up procedure, the RMSDs of the predicted multiloop fragments from native are found to increase with the size of the fragments, even when the RMSD for the individual loops remains essentially constant; a translation-only least-squares fit RMSD model is developed to explain this type of finding. Analysis of the errors in the backbone dihedral angles of the correctly predicted CheY structures shows them to approximately cancel, extending previous observations[51, 52].

A series of Metropolis Monte Carlo and simulated annealing trials, using the same conformer libraries and energy function as were used in the main set of Z Module calculations, are shown to result in significantly less accurate predictions for the CheY structure, even when the prediction problem is further reduced. Although the total number of DOF in the problem examined here is still limited, the overall study results serve as a proof of principle, that systematic “grid-type” searches using atomic potential energy functions can be combined with statistical information in hierarchical build-up procedures to accurately predict protein tertiary structure. The findings also illustrate the utility of the Z Module and the parallel-generation approach in structure prediction studies.

3 Methods

3.1 Description of the Z Method

The Z Module is a general facility for carrying out systematic conformational searches. It includes 1st-order minimization features (Steepest Descent and Conjugant Gradient), but the fundamental structure of the method is grid-based. The method depends on a partitioning of the conformational space of a molecular system into parts called conformational *subspaces*. Here, a subspace, Ω_i^n , is the space spanned by an n -member subset of the set of N degrees of freedom in the entire conformational space, Ω^N , of the system. An example of a subspace, say $\Omega_1^{n_1}$, would be the conformational space spanned by all of the internal degrees of freedom of a Val side chain. A subspace has associated with it an atomic *subsystem*, Λ_i , which is the set of atoms that are included in the energy calculations whenever the subspace Ω_i^n is included in the search. For a given subspace, the composition of the corresponding subsystem depends on the problem. In the above example, Λ_1 might be composed of all the atoms in the Val residue; it could also include adjacent residue atoms in the sequence, or protein environment atoms. The conformational subspace corresponding to a subsystem does not necessarily include all of the DOF (e.g., internal coordinates or ICs) necessary to specify the subsystem, because some of them can be assigned to other subspaces or, alternatively, can be fixed. For example, Λ_2 could be defined as a particular Lys side chain

and its nearby protein surroundings, but $\Omega_2^{n_2}$ might include only the χ angles of the side chain. Ideally, a subsystem is composed of all the atoms whose positions are entirely specified by the DOF of a subspace(s) involved in the search and the known (fixed) ICs. The inclusion of additional atoms or the exclusion of atoms can lead to errors (see also Discussion).

A *conformer* specifies a point in a conformational subspace. Conformers can also be grouped into subspace *regions*. Specifically, the conformer $\Omega_{i,j,k}^n$ is an n -tuple specifying the k th point in the j th region of the subspace Ω_i^n . A region is a collection of $C_{i,j}$ conformers in a subspace (not a smaller subspace of Ω_i^n). Regions are useful because they increase the flexibility of the search protocols, since a given subspace can switch between two or more regions within the same search (see also below). A search *grid* consists of a combination of regions from the subspaces to be searched, e.g., $\Omega_{1,1}^{10} \times \Omega_{2,1}^8$, as further illustrated below. A *gridpoint* is a particular combination of conformers on a grid—e.g., $(\Omega_{1,1,1}^{10}, \Omega_{2,1,1}^8)$. A *search* consists of the complete set of search combinations or grids. A single CHARMM run can comprise multiple Z Module searches.

Before a Z Module search can be performed, its basic parameters must be specified. First, the degrees of freedom, expressed in ICs—i.e., bond lengths, bond angles, improper or proper dihedral angles—must be defined. The defined DOF must of course include all those to be searched, what are defined here as the *problem degrees of freedom*. Since some of the ICs in the system under study may be taken as fixed in a given problem (as in the applications to CheY illustrated here), the number of problem DOF (36 variable ϕ and ψ angles) may be significantly less than the number of total DOF in the system (e.g., more than 3000 for the polar-H model of CheY).

The problem DOF are grouped into subspaces, and for each subspace, Ω_i^n , the (internal) coordinates for each conformer $\Omega_{i,j,k}^n$ to be evaluated are specified. These two steps are performed simultaneously in the input conformer file (see Table 6, Supplementary Material). If a subspace is to be searched in two regions, the conformers corresponding to both regions are specified. An atomic subsystem is specified for each subspace by the selection of two sets of atoms. The first is the set to be included in the energy calculations (and mean squared deviation calculations). The second is the set of atoms which are to be rebuilt in the search—i.e., those affected by a change in the subspaces.

Once the subspaces and the associated conformers, regions, and subsystems have been specified, they can be searched independently or in combinations. In the case of two regions per subspace (the limit in the current CHARMM implementation) and W total subspaces being searched, the Z Module searches all $W!/(W - W_a - W_b)!W_a!W_b!$ combinations of the W subspaces taken W_a and W_b at a time, where W , W_a , and W_b are non-negative integers specified by the user and $W_a + W_b \leq W$. Each of these combinations is a search grid. For each one, W_a subspaces are searched in their “ a ” regions and W_b subspaces are searched in their “ b ” regions, so that the number of conformers searched on a given grid is

$$\prod_{i=1}^{W_a} C_{i,a} \cdot \prod_{i=1}^{W_b} C_{i,b},$$

where $C_{i,j}$ is as defined above.

For example, if there are 5 total subspaces ($W = 5$), each containing 20 conformers in their a regions ($C_{i,a} = 20$) and 10 conformers in their b regions ($C_{i,b} = 10$), a search that explores 3 subspaces at a time, 1 taken from the a regions of the subspaces and 2 from the b regions ($W_a = 1$; $W_b = 2$), would produce 30 grids ($5!/2!1!2!$), each with $(20)(10)(10) = 2000$ gridpoints or conformer evaluations. See Fig. 1. As also illustrated in Stage 2 of the CheY predictions, the “region” formalism allows for more refined searches without incurring the maximal computational cost. For example, it may be optimal in a given case to explore a number of subspaces in large conformational regions (the “ a ” regions) while exploring the other subspaces in smaller regions (the “ b ” regions), and to do this in all possible combinations of the subspaces. The latter motions essentially allow the system to relax around the larger-amplitude changes, but the total calculation costs far less than searching all of the subspaces over the larger ranges. The idea is similar to that of having motions occur in slow and fast manifolds[53, 54], except that here both the “fast” (small-amplitude) and “slow” (large-amplitude) motions occur within the same search and can switch from one subspace to the other.

The Z Module imposes no *a priori* limits on the number of degrees of freedom that may be defined or used in a search, the number of DOF per subspace, the total number of conformers in a search, the number of conformers per subspace, the total number of subspaces per search (W), the number of subspaces searched at a time (W_a or W_b ; provided $W_a + W_b \leq W$), or the number of constraints defined or imposed (see below). The only practical limits for these parameters are determined by the memory capacity of the hardware and the time considerations of the user. In addition, the Z Module can perform “partial-structure” searches; this means that a single, whole-molecule protein structure file (PSF) can be used for calculations on any part of the structure. This feature facilitates hierarchical build-up procedures, since different PSFs do not need to be generated and used for each molecular fragment.

The results of the overall search consist of the conformers of a subspace, the *product subspace*, that is the union of all the *antecedent* subspaces ($\Omega_i^m = \Omega_1^{n_1} \cup \Omega_2^{n_2} \dots \cup \Omega_W^{n_W}$). Hence, when $W > 1$, the product subspace is always larger than any of the antecedent subspaces. For each grid, the ICs of the subspaces not being searched are held fixed. Analogously to the case for the subspaces, the set of atoms included in the energy evaluations during a search, the *product subsystem*, is the union of the sets of atoms in the subsystems of all of the antecedent subspaces. Hence, the set of atoms included in the calculations is determined at the start of a Z Module search, as is the size of the product subspace. A particular atom can often be a part of more than one subsystem. However, for self-consistency, subspaces involved in the same search should not share common degrees of freedom. (That is, $m = \sum_{i=1}^W n_i$). While the Z Module allows it to be done (the user can assign any DOF to any subspace), including common degrees of freedom in more than one antecedent subspace in a search would overspecify the problem, so that the coordinates of the shared DOF would be overwritten for at least one of the subspaces, depending on the order in which they were taken in the search.

The conformers resulting from the search may be written to a binary .dcd (“trajectory”) file or to an output conformer file. The latter is formatted similarly to the input conformer file, but includes the energies of the product subspace conformers and, if specified, the mean squared deviations of the product subsystems from a reference structure. Hence, an output conformer file from one calculation can be used as an input conformer file for a subsequent calculation. Fixed or descending (e.g., 10 kcal/mol above the running minimum) energy cutoffs can be used to reduce the size of the output. This feature also facilitates the use of a “dead-end elimination”-type of conformation restriction.

For whole-molecule calculations, the Z Module can be used with any energy model. In principle, this includes the use of explicit water molecules to represent bulk solvent. In practice, however, the use of explicit water in grid-style searches is usually impractical, because the displacements of the solute between gridpoints are often large and can result in frequent clashes with the water molecules. Implicit (i.e., continuum) water models are preferred, since water relaxation in such models is effectively instantaneous. In partial-structure calculations, the terms in the energy function must be decomposable. The EEF1 solvation term is pairwise decomposable, and partial-structure EEF1 energies have been implemented in CHARMM. Product subspace conformers may be energy-minimized for a user-specified number of steps, with either the Steepest Descent or Conjugate Gradient methods. The unminimized structure is always restored before the next conformer is evaluated, because unconstrained derivative-based minimization affects all DOF associated with the subsystems, not just the problem DOF. Failure to restore the fixed DOF to their initial values between gridpoints would introduce a bias in favor of gridpoints evaluated later in the grid. Distance and IC constraints may be applied as filters in the searches. The IC constraints may be applied to DOF that are being explicitly searched as well as those that are not. For example, the dihedral and bond angles bridging the two halves of a loop do not exist in either half alone and hence cannot be explicitly searched, but they can be used as constraints when the product subsystem is the whole loop.

For efficiency, the Z Module modifies the CHARMM IC table and rebuilds the structure just once at each gridpoint, rather than rebuilding part of the structure for each corresponding DOF. Hence, use of the facility requires the presence of the (appropriately filled) IC table. The module is designed to be able to be entered and exited easily and often over the course of a single CHARMM script. This allows sections of Z Module commands in CHARMM scripts to be conveniently interspersed with other CHARMM commands and functions.

The Z Method is parallelizable, both at the level of grids within a single search and also at that of parts within a single grid. Efforts to parallelize the Z Module code along these lines are currently underway.

Additional details of the method are provided in Supplementary Material and in the CHARMM documentation.

3.2 Z Module conformational searches of CheY

The CheY protein is a signal transduction protein that has 10 secondary structural elements (SSEs), namely 5 α helices (α_1 through α_5) and 5 β strands (β_1 through β_5), which alternate in the sequence, beginning with β_1 . The SSEs are separated by 9 flexible loops (see Fig. 2B). The coordinates used in the current study were taken from the RCSB protein data bank[55] (2chf, which is a 1.8 Å structure); the structure was not energy-minimized. Hydrogen atoms were added according to a polar-hydrogen model, as previously described[37]. The energy evaluations throughout the CheY calculations were performed using the EEF1 implicit solvation model [38] in the CHARMM program[30, 31]. A single PSF was used for all calculations. The non-bonded cutoff was 10 Å (as per the EEF1 model), and the neighbor list was updated at each step. This high update frequency was used to guarantee that all interactions within the cutoff distance were included at each gridpoint (i.e., step), since large displacements in atomic positions can occur between gridpoints, particularly for large fragments. (For calculations that involve only small fragments, if a long-enough cutoff distance is used, no non-bonded updates are necessary). The BYCC listbuilder[56] was used throughout the studies. It is required in the Z Module because it is currently the only listbuilder in CHARMM that can generate lists of non-bonded atom pairs for partial structures. The CHARMM bonded energy routines were also modified so as to be able to calculate partial-structure bonded energies (the ACTBOND precompiler keyword is

therefore required, in addition to ZEROM, which invokes the Z module). All of the Z Module calculations were carried out on a dual-processor AMD Opteron 242 (1.6 GHz) machine, using the GNU compiler suite. (Intel Fortran compilations gave similar computational efficiency.)

In each of the nine loops, the two smallest adjacent residues were selected for the searches. They are (loop number, 1st residue in pair, 2nd residue in pair): 1) Asp 11, Asp 12; 2) Asn 30, Asn 31; 3) Glu 36, Asp 37; 4) Gly 48, Gly 49; 5) Asn 58, Met 59; 6) Ser 75, Ala 76; 7) Glu 88, Ala 89; 8) Gly 101, Ala 102; 9) Phe 110, Thr 111. The effect of the side chains of the loop residues have been shown to be very small in this type of study[37]; nonetheless, the smallest residues in the loops were chosen here to minimize the effect. The ϕ and ψ angles for both residues in each loop were included in the problem DOF set, so that there were four variable angles in each of the nine loops, for a total of 36 problem DOF. The internal structures of the helices and β strands were fixed (i.e., they correspond to the fixed DOF) at the native positions, so that the problem involves finding the ϕ and ψ backbone dihedral angle values that result in the lowest- energy arrangements of the 10 SSEs.

The prediction was carried out in two stages. Stage 1 was a build-up procedure in which fragments of the structure, or more precisely, low-energy conformers of small subspaces of the molecule, were generated and combined to form larger fragments or subspace conformers. In the first step, single-loop searches were performed. This means that the ϕ and ψ angles of the two variable residues (for a total of 4 dihedral angles) in each loop were searched, one loop at a time. Throughout this work, references to searches of a “loop” imply that the SSEs flanking the loop are also included in the energy evaluations (i.e., the corresponding subsystems). Hence, the calculations are not loop predictions in the conventional sense. For example, the subspace searched in loop 2 is composed of the (ϕ , ψ) angles of residues 30 and 31, but the corresponding subsystem includes all residues whose positions are specified exactly by this subspace and the known (fixed) internal coordinates of the structure— in this case residues 13–35 (α_1 , loop₁, β_2 ; residue 13 is actually part of the previous loop, but in this problem its structure is also entirely specified in a loop 2 search, as is part of residue 12).

For each loop, all pairwise combinations of the single-residue conformers were evaluated. The latter were derived from a modified version of the (ϕ , ψ) protein backbone libraries developed by Jacobson *et al.* [35], which are based on the backbone conformations occurring in the RCSB protein data bank[55] and are discretized on a 5° grid. Here, the original libraries for Gly and non-Gly, non-Pro residues were both extended slightly by ensuring that conformers within a range of $\pm 10^\circ$ in the ϕ and ψ directions around any of the original conformers was present. This was done because some of the dihedral angle values in the native CheY structure were very near the boundaries of the original libraries. The libraries were then discretized to a coarser grid of 10°, to reduce computational time for the build-up procedure. The total number of conformers for each Gly residue was 508 and for each non-Gly residue it was 402, so that, for example, the first loop calculation (Asp 11, Asp 12) had $402^2 = 161,604$ gridpoints or conformer evaluations.

In the second step, low-energy conformers for the single-loop calculations were combined in double-loop (three SSE) calculations. The “low-energy” single-loop conformers, both in this step and throughout the CheY calculations, were selected in a systematic way on the basis of the energy distributions, the number of conformers desired, and an energy cutoff of 10 kcal/mol. Specifically, if the number of conformers to be taken from the results of a particular search was C , then if there were greater than C conformers within 10 kcal/mol of the minimum, C conformers were selected randomly from that 10 kcal/mol range. If there were fewer than C conformers within the range, then the C lowest-energy conformers overall

were selected. The selection of 10 kcal/mol as the cutoff criterion was based on the observation that this type of search in a protein—i.e., a systematic, flexible loop search with fixed SSEs—tends to generate significant numbers of near-native structures within a roughly 10 or 15 kcal/mol band above the minimum-energy conformer. In the selection of the single-loop conformers for use in the double-loop searches (as well as for selection of double-loop conformers for the 4- and 5-loop searches), $C = 1000$. Hence, for each double-loop search, there were $(1000)(1000) = 10^6$ gridpoints. The pairs of loops are adjacent in the sequence: e.g., loop 1 paired with loop 2, loop 3 paired with loop 4, etc. As mentioned, the Z Module takes the subsystem of the product subspace to be the union of the individual antecedent subsystems, so that for the double-loop searches, the atoms included in the calculation are those contained by the three SSEs connected by the two loops (e.g., β_1 , α_1 , and β_2 for loops 1 and 2), along with the atoms in the loops themselves. Since the total number of loops in the protein does not equal 2^I , where I is an integer, in an additional part of this second step in Stage 1, 1000 conformers from loop 9 were combined with 1000 conformers from the combined loop 7–loop 8 calculation. The details of both stages of the protocol are given in Table 1. In the last step of Stage 1, the low-energy conformers for the entire conformational space (36 dihedral angles) were generated by combining the results from the loop 1–4 (5-SSE) calculation and the loop 5–9 (6-SSE) calculation. The lowest energy structure from this final Stage 1 search was selected, and this was used as the starting structure for Stage 2.

Stage 2 was a *reworking stage*, because it attempted to improve the energetics of a starting structure for the entire molecule. This is to be contrasted with an *initiation stage*, such as Stage 1 (here the build-up stage), in which there is no starting structure and the values of all of the problem degrees of freedom are unknown. The design of a reworking stage is thus based on the assumption that part or all of the structure is approximately correct, and that more localized or “regional” searches in conformational space can be done to locate the global minimum. In Stage 2, conformations were first generated on a 5° grid (the original grid spacing in the backbone conformer libraries) for each individual loop, analogously to the 10° case described above for the build-up procedure. From these results, two sets of conformers were then created for each loop, one in a wide range, and one in a narrow range. For the wide-range regions, 1000 conformers were selected in an energy band of 10 kcal/mol above the minimum for each of the single-loop searches, with no angular restrictions. For the narrow-range regions, 200 conformers were selected in the same energy bands, with the added criterion that all of the dihedral angle positions in the chosen conformer had to be within $\pm 60^\circ$ of those of the starting structure. The $\pm 60^\circ$ cutoff criterion was chosen because it significantly reduces the computational time, (in 2-dimensional grids, by a factor of ≈ 9), while still allowing for local sampling to occur. Once these regional loop conformers were selected, they were used to generate conformers around the starting structure for each half of the molecule in two separate searches. In one search, low-energy conformers were generated for loops 1–5, and in the other search for loops 6–9. In both, the searches were performed by varying 2 of the loops of the structure at a time: one over its wide range and one over its narrow range. In a final calculation, 2000 low-energy structures from each half were combined (4×10^6 total gridpoints) to generate the final predicted structure.

3.3 Monte Carlo and simulated annealing calculations

Monte Carlo calculations were carried out for variously sized fragments of the CheY structure that were found to be stable (i.e., well-predicted) in the Z Module build-up procedure. The four fragments studied were A) loop 1; B) loops 1 and 2; C) loops 5–9, and D) all nine loops (whole protein). In each case, the variable residues were chosen to be the same as in the main set of CheY calculations; the backbone libraries and the energy function used were also the same. The only difference between these trials and the main (Z Module) set was the method of search. Trials were conducted at 300K and 10000K for 1 million to 5

million steps, using a standard Metropolis acceptance criterion. For the largest fragment, a trial was also carried out at 100000K. At each step, the residue(s) to be varied were chosen randomly from the variable set, and the for those residues, dihedral angle values were randomly chosen from the corresponding backbone libraries. In addition, a number of simulated annealing trials were carried out in which the final structure from the 10000K calculations was used as a starting structure. An exponential temperature schedule was used, in which $T_{k+1} = \alpha T_k$, where k is the k th temperature in the protocol, and where $\alpha = (T_f/T_0)^{N_c/(N_f-N_c)}$, with T_f the final temperature (300K here), T_0 the initial temperature (10000K), N_c the number of steps to be carried out at a given temperature, and N_f the total number of steps. Most of these calculations were carried out on a cluster of Intel Xeon CPUs (GNU compiler suite); some were carried out on the AMD Opteron machine mentioned previously.

3.4 Dihedral angle compensation

A series of exploratory calculations was carried out to investigate dihedral angle compensation (the tendency of changes in dihedral angles to cancel) in loops with fixed ends and otherwise fixed internal geometries. Analytic ring closure software[57] was used to generate alternative dihedral angle combinations for amino acid trimers in the backbones of three proteins: turkey egg lysozyme (135L, 129 residues), adipocyte lipid-binding protein (1l1f, 131 residues), and CheY. At each trimer position, or register, along the backbone, the program was used to generate alternative sets of (ϕ , ψ) angles for the trimer, given fixed ends, and fixed bond angles, bond lengths, and ω dihedral angles. There were sometimes multiple alternative solutions (i.e., multiple alternative sets of angles satisfying the constraints) in a given register.

4 Results

4.1 Tertiary structure prediction of CheY

In this reduced problem the Z Module was used in a series of systematic, combinatorial searches to find the native structure of the signal transduction protein CheY. The relative positions of the SSEs of the protein were explored by varying the backbone dihedral angles of the smallest 2 adjacent residues in each loop of the protein, so that there were 4 variable angles (2 ϕ angles and 2 ψ angles) in each of the 9 loops, for a total of 36 problem degrees of freedom. An exhaustive search of such a large conformational space is impractical: at a grid size of 5° , there are 72^{36} unique structures, and assuming an evaluation speed of 10^9 gridpoints/day on a single CPU (faster than the 5×10^6 gridpoints/day obtained on the AMD Opteron CPUs in the study), the total computational time required would be $\approx 7.3 \times 10^{57}$ CPU years. In a previous study[37], which examined the same type of problem in 3 proteins, the CheY search involved nine degrees of freedom, one per loop. Throughout that study, extensive, systematic (though not exhaustive) searches were performed on the entire structure, and in an initial stage included all nine degrees of freedom. The current study also uses extensive, systematic searches, but for the much larger problem addressed here, the approach has had to be modified. The prediction is carried out in a two-stage protocol. Stage 1 involves a build-up procedure in which small fragments of the structure (conformers of small subspaces) are generated and combined into larger fragments, in a hierarchical fashion. In the last step of this build-up stage, low-energy conformers of the entire molecule are generated. The lowest-energy structure is then selected, and this is used as the starting structure for Stage 2. Stage 2 is a "reworking" stage, in which parts of the structure are varied around their starting conformations (see Methods section for details).

The plots of 4 energy distributions from the first step in the protocol, which involve single loops (4 dihedral angles; 2 sequential SSEs), are shown in Fig. 3. The four panels (A–D) correspond to loops 1 through 4. The distributions tend to be much denser and broader at

higher energies, sparser and narrower at lower energies. For loop 1, (panel A), there is a rather deep minimum in the region of the native structure, with a gap of about 5 kcal/mol between the two lowest-energy structures, which are less than 0.5 Å from the native structure, and a cluster of structures which are next-highest in energy. However, in loops 2 and 4 (Panels B and D), there are non-native minima that are more than 3.4 Å away from the native position and that are lower in energy than the near-native conformations. In fact, all of the near-native conformers sampled in both of these loops are greater than 5 kcal/mol higher in energy than the corresponding minimum-energy conformers. In loop 3, the distribution is split into two parts. The first is around the native structure, and has a minimum-energy conformer 0.91 Å from the native. The second is funnel-shaped and roughly 4–7 Å from the native structure, and it has a minimum that is 5.32 Å from the native. The lowest-energy conformers from the two parts of the distribution are essentially degenerate ($\Delta E = 0.0033$ kcal/mol) and there are 5 non-native conformers within 1.0 kcal/mol of the minimum. The RMSDs from native for the lowest-energy conformers found in each of the single-loop calculations is shown in Table 2. Of the 9 loops, 6 (all but loops 1, 6, and 9) have minima greater than 1.6 Å in RMSD away from the native structure. (The energy distributions and minima for the 5° single-loop searches, which were used for Stage 2, are similar to those for the 10° searches and are not shown).

An energy comparison of the predicted and native loop (and flanking SSE) structures demonstrates that 7 of the 10 predicted structures (there are 2 for loop 3) have lower energies than native. Minimization of the native structure by 10 steps SD and assignment of the predicted dihedral angles to this minimized native structure demonstrate large drops (> 15 kcal/mol) in the van der Waals and total energies of the minimized native structure relative to the predictions for 3 of the 7 cases. This indicates that in the native structure, whose non-hydrogen atoms were taken directly from the X-ray crystal structure, without optimization on the CHARMM (here, EEF1) energy hypersurface, there were some “bad” or “suboptimal” contacts. The other 4 predicted loop structures that were lower in energy than native remained lower after the 10-step SD optimization of the native structure; in these structures, there are significant stabilizing contributions from electrostatic or solvation terms. Three of the structures are greater than 2.3 Å from native, indicating that in the absence of the rest of the protein, there are stable non-native conformations for some of the single-loop fragments. The complete energy component analysis is provided in Table 7 (Supplementary Material).

Since tertiary contacts between non-adjacent SSEs are excluded in these individual loop predictions, the findings here suggest that while interactions between secondary structural elements that are adjacent in the sequence do tend to favor native-like structures, they alone do not determine the fold, even in the presence of the native secondary structures. This is further supported by the fact that if the predictions for these individual loops are used to assemble the entire structure (with the near-native minimum used for loop 3), the result has an RMSD from native of more than 15 Å.

The 5 β strands (numbered $\beta_1 \dots \beta_5$) in CheY form a sheet, but the order of the strands in the structure is not sequential: $\beta_2, \beta_1, \beta_3, \beta_4, \beta_5$. In the prior study, [37] it was observed that mainly for steric reasons, this spatial order affected the order of folding. The pathways in which β_2 and β_1 associate first and then associate with β_3 were found to be favored over those in which β_2 and β_3 assume their near-native orientations first and β_1 then moves between them. In the 2-loop step of the build-up procedure here (results not shown), only the second of the four fragments, which contained SSEs β_2, α_2 , and β_3 , had a minimum energy conformation that was very far (≈ 18 Å) from native, which is consistent with the results of the prior study—i.e., since β_1 is “missing,” this fragment is unlikely to adopt a native-like conformation.

As shown in Fig. 2A, the lowest-energy structure generated at the end of Stage 1 is an “open” structure, with an RMSD from the native of 20.46 Å. The energy distribution for the last step of this stage is shown in Fig. 4A. The distribution is centered approximately around the non-native minimum, and there are no low-energy structures sampled around the native state. Out of 1,000,000 sampled structures, 427,736 structures have energies that are within 20 kcal/mol of the minimum; only 5 of those are closer than 6.7 Å from the native structure (they are between 6.5 and 6.7 Å in RMSD away). However, as shown in Table 2 the RMSDs for most of the individual loops are less than 1 Å from native. The large RMSD for the overall structure is mainly attributable to the large error in loop 6. Remarkably, loop 6 was one of the 3 loops whose initial single-loop energy surfaces had a near-native structure as the minimum (RMSD 0.417 Å). Hence this fragment has “opened up” to a non-native conformation, which represents a metastable “intermediate.” Consistent with these results, most of the dihedral angles in the Stage 1 structure are close to the native values—28 of 36 are within $\pm 20^\circ$ of the native. The results are shown in Table 3. Of the nine loops, six have all four dihedral angles within 26° of the native. Loops 3 and 4, which involve the β strands that associate non-sequentially in the native structure, have RMSDs from native of less than 1 Å. The final Stage 1 structure has total and van der Waals energies that are 141.9 and 156.1 kcal/mol higher than native, respectively, as shown in Table 7 (Supplementary Material); there are large, offsetting differences in electrostatic (136.1 kcal/mol) and solvation (-151.6 kcal/mol) terms, related to the greater solvent exposure of the predicted structure, which is far less compact than the native. These offsetting differences are consistent with results of previous whole-protein calculations using the EEF1 solvation function[37]. The current findings indicate that the Stage 1 search was extensive enough to generate the correct fold for parts of the structure, but not extensive enough to locate the correct global minimum for the entire molecule.

In many cases, the errors in the dihedral angles approximately cancel. In 3 of the 9 loops (loops 1, 2 and 9), errors of $7\text{--}45^\circ$ in the ψ angles of the first variable residue in the loop are compensated by errors in the ϕ angle of the second variable residue in the loop, such that $|\Delta\psi_i + \Delta\phi_{i+1}| < 3^\circ$, where i is the residue number. It has been noted that this pattern of compensation can occur in *trans* peptides because the torsional axis for the ψ angle of residue i (i.e., that along the $\text{C}\alpha\text{--C}$ bond) is roughly parallel to (although not colinear with) the torsional axis for the ϕ angle of residue $i + 1$ (i.e., that along the $\text{N--C}\alpha$ bond)[51, 52]. For large compensatory changes ($|\Delta\psi_i| + |\Delta\phi_{i+1}| > 200^\circ$), such a pattern has been termed a “peptide plane flip” and has been observed to occur between different crystal structures of the same protein[51, 52] and in molecular dynamics simulations[58]. These types of dihedral angle compensation patterns are particularly remarkable in the current work because the entire structure, aside from the 36 variable loop dihedral angles, is fixed, so that there is no additional opportunity for local relaxation of the structure around the regions (i.e., loops) that display compensatory changes. The findings suggest that, like a peptide-plane flip, the local compensatory changes in ψ_i and ϕ_{i+1} seen here can occur without a large energetic penalty. Moreover, there is a rough correlation between the compensation of the errors in the dihedral angles for the individual loops and their RMSDs from native. The 3 most poorly predicted loops in terms of RMSD (3, 5, and 6) have the least dihedral angle compensation, particularly loop 6 (sum of dihedral angle errors $\approx 124^\circ$).

In the final Stage 2 structure, all of the individual-loop RMSDs are ≤ 0.71 Å from the native structures, as shown in Table 2. The RMSD for the whole structure, shown in Fig. 2B, is 1.56 Å from the native. The results indicate that the searches were adequate for locating the correct global protein fold. The fact that the global RMSD is significantly greater than the average of the partial-structure RMSDs (0.468 Å), in this stage and in the previous stages, arises naturally from the worse fitting achievable in larger structures, which is considered further in the Discussion and Appendix. As can be seen in the figure, the native structure is

slightly more compact than the prediction. The radii of gyration are 13.52 Å and 14.24 Å respectively; this is partly a consequence of the coarseness of the final grid spacing (5°). The total and van der Waals energies of the final prediction are 84.2 and 74.5 kcal/mol higher, respectively, than the (unoptimized) native structure, as shown in Table 7 (Supplementary Material). Achieving a closer superposition would require a refinement or reworking stage with a small grid spacing (e.g., 1 or 2°[37]) and an assumption that the structure was quite close to the global minimum. Because of the steepness of the repulsive component of the van der Waals energy and the size and compactness of the subsystem here (whole molecule), the energy landscape near the global minimum is steep as a function of RMSD from native (see below), and also varies rapidly as a function of the grid spacing.

The energy distribution for the final step of Stage 2 (in which conformers for both halves of the molecule were combined) is shown in Fig. 4B. There is a deep, narrow branch of the distribution around the minimum, which extends from the minimum to more than 20 kcal/mol above it. There is also a cluster of structures at ≈ 20 Å from the native structure, which is greater than 18 kcal/mol higher than the minimum. Table 3 demonstrates that, as in the Stage 1 structure, 28 of the 36 angles are within $\pm 20^\circ$, although not all of the incorrect angles are the same as those from Stage 1. Four of the loops (3, 5, 6, and 9) differ between the Stage 1 and Stage 2 structures. This is expected, because two loops were allowed to vary at a time (in all combinations) in each of the two halves of the molecule in Stage 2. In six of the nine loops of the final structure, there is (ψ_i, ϕ_{i+1}) compensation to within $\Delta\psi_i + \Delta\phi_{i+1} \leq 2.85^\circ$ between the middle two residues in the loop.

The dihedral angle error compensation is marked in the Stage 2 structure: although there are individual errors of up to 51.42° , the sum of the errors for whole loops have absolute values of no more than 12.83° , and in 7 of the 9 loops they are no more than 3.84° . Also of note is that in both the Stage 1 and Stage 2 structures, the sum of the errors in loop 3 (which is predicted to within 0.80 Å in both structures) is exactly -12.83° , although the errors in 3 of the 4 individual angles are significantly different. Finally, the sum of the dihedral angle errors for the entire Stage 2 structure, which is approximately correct, is only 3.96° , while that for the entire Stage 1 structure, which differs significantly from the native, is 133.96° . Taken together, the current results indicate that, as in the case of the compensatory changes in ψ_i and ϕ_{i+1} , errors in all of the variable ϕ and ψ angles of the correctly predicted CheY structure, as well as in the correctly predicted fragments, are compensating in an approximately additive way.

4.2 Monte Carlo calculations

A number of Monte Carlo and simulated annealing conformational sampling calculations were carried out on CheY using the same energy function (EEF1 solvent model in CHARMM) and the same variable angles and angular libraries as used in the Z Module calculations. The calculations were performed on 1-loop (containing 2 SSEs), 2-loop (3-SSE), and 5-loop (6-SSE) portions of the protein, as well as on the entire protein. The results are shown in Table 4. The best results (RMSD from native of 0.52 or less) were obtained for small parts of the molecule (one or two loops), varying 1 or 2 residues at a time, at either 300K or in simulated annealing runs having a final temperature of 300K. For these small parts of the molecule, trials at a higher temperature or varying a larger number of residues resulted in final structures that deviated significantly from the native structure (RMSD 3.75 to 12.79 Å). For the calculations involving either the 5-loop part of the molecule or the entire structure, all the trials resulted in structures with an RMSD from native of more than 10 Å. In these cases, the simulated annealing trials and MC trials run at 300K tended to do achieve slightly better predictions (RMSD 10–12 Å) than the higher temperature trials (RMSD 17.12 to 39.09 Å). For these larger portions of the molecule, the final structures generally have less than half of the dihedral angles correct (less than 40° from the native

values) and there is no dihedral angle compensation within or between loops. In addition, many of the 300K trials became trapped well before the end of the simulation (6/11 became trapped before 30% of the search was complete, and 9/11 before 90% was completed.) None of the 10000K runs became trapped, except in the largest (9-loop) structures, but the prediction results tended to be poor. The 9-loop, 10^5 -step trial at 100000K also resulted in a poor prediction (≈ 30 Å).

4.3 Dihedral Angle Compensation

The approximate cancellation of all dihedral angle errors in the correctly predicted subsystems of CheY and the absence of such cancellation in the incorrectly predicted subsystems raises the question of whether the result is a coincidence or the consequence of a general rule. As described in Methods, exploratory calculations were carried out with analytical loop closure software. Figure 5 shows the results for alternative conformations generated for successive trimers along the backbones of native lysozyme, CheY, and

adipocyte binding protein. The quantity $\Sigma_2 = \sum_{i=k}^{k+3} |\Delta\phi_i| + |\Delta\psi_i|$ is plotted on the vertical axis and the quantity $\Sigma_1 = \sum_{i=k}^{k+3} \Delta\phi_i + \Delta\psi_i$ is plotted on the horizontal axis, where k is the residue number, and the Δ 's refer to the deviations from the values in the native structures. Hence, the plot compares the absolute magnitudes of the angular deviations to the degree of their cancellation, where a 0 value on the horizontal axis indicates exact cancellation. A total of 880 alternative trimer conformations were generated for the three proteins. Of the 229 trimer structures with $\Sigma_2 < 150^\circ$, 192 or 83.8% have Σ_1 within 5° of 0; for a cutoff of $\Sigma_2 < 90^\circ$, the fraction is 150/160 or 93.8%. There is a wider spread of Σ_1 values at higher values of Σ_2 (not shown); still, for all 880 solutions, which range up to $\Sigma_2 \leq 712.9^\circ$, 739 or 83.9% have Σ_1 within 30° of 0. Further analysis indicates that a significant portion, but not all, of the overall cancellation is due to ψ_i/ϕ_{i+1} -type compensatory changes (results not shown). There are only two ψ_i/ϕ_{i+1} pairs in a trimer; thus, perfect cancellation involves other effects. A repeat calculation using a starting lysozyme structure whose ϕ and ψ angles were randomized and selected for stability showed similar results. The results indicate that for two alternative conformations of a protein loop with fixed ends and all ICs fixed except for the ϕ and ψ angles, the sum over the differences in the dihedral angles between the two structures tends towards zero. The tendency is stronger when the changes in the dihedral angles are small.

5 Discussion

Atom-based potential energy functions are being employed with increasing success in the field of biomolecular structure prediction[3, 4, 9, 59, 45]. The current study describes a general method, implemented in the CHARMM program as the Z Module, for carrying out extensive, systematic energy-based conformational searches that can be used to predict molecular structure with an atom-based model. Hallmarks of the method are that (1) it generates sampled structures in a “parallel” fashion, by repeatedly applying the same type of change to a single starting structure; (2) it does this systematically by breaking down the conformational space and searching the subspaces in combinations; and (3) it can be used to build up libraries of successively larger fragments of the predicted structure from libraries of smaller fragments. The module has been used previously to solve “real” structure prediction problems. In a 2008 study, the Z Module was used to predict the structure of the CMV UL44 processivity factor complexed with a DNA oligomer. The resulting set of models was found to be consistent with data from biochemical and mutational experiments[42] and to suggest mechanisms for DNA/protein binding and processivity. The main intent of the current paper has been to describe the Z Method, its implementation in CHARMM as the Z Module, and some of its motivating principles. In addition, another application of the method has been demonstrated, whereby the tertiary structure of the protein CheY is predicted in a model

problem. This was done by performing a series of systematic searches in a 36-dimensional conformational space describing the relative positions of the secondary structural elements (SSEs, i.e., β sheets and α helices) of the protein. A hierarchical fragment build-up procedure, followed by a reworking or refinement stage, resulted in a prediction that was within 1.56 Å of the native structure (all-atom RMSD) in a total time of approximately 2 1/2 CPU days on 1.6 GHz AMD Opteron processors.

Over the course of the build-up stage, the fraction of low-energy CheY fragments that had conformations within 2.0 Å of native increased from about 1/2 to 7/9, suggesting that interactions between sequential parts of the structure favor but do not determine the native fold, even when the SSEs, themselves, are in their native conformations. The energy vs. RMSD distributions tended to be funnel-shaped, with large numbers of structures at higher energies and fewer structures, usually closer to native, at lower energies, similar to previous observations for protein folding landscapes [37, 60, 61, 62, 63, 64]. In the reworking stage, the native structure was found to reside in a deep, narrow minimum on the energy surface, a finding consistent with prior studies of CheY[37]. A series of Markov-chain-type Monte Carlo and simulated annealing trials using the same energy surface as in the Z Module calculations generally resulted in poorer predictions of the native CheY structure, and the MC trials at room temperature often became trapped.

The accurate CheY prediction achieved here represents the solution of a much smaller problem than *de novo* whole-protein prediction. Nonetheless, together with the results of previous studies[32, 33, 34, 35, 36, 37], the current findings suggest that Z Method searches – i.e., systematic, “parallel-generation” searches using atom-based potential energy functions—can be useful at least for a certain range of problems, and that this approach to structure prediction has particular advantages. One obvious advantage is that the method avoids becoming trapped in metastable states. Another, which is less obvious, is that it avoids discrepancies in the degree of local sampling around the conformers being evaluated; such discrepancies can affect prediction accuracy, as described in a separate work[41]. By contrast, methods that produce and evaluate structures serially, such as classical Monte Carlo or other chain methods, do not guarantee even sampling unless the number of steps is large relative to the number of minima. The practical limitations of standard room-temperature Metropolis MC algorithms for searching the phase (or conformational) space of systems with larger volumes and multiple minima are well-known[65, 66, 67, 68]. In large, heterogeneous systems like proteins, a single classical Monte-Carlo-type serial search procedure is not expected to result in global optimization[69]; multiple MC or simulated annealing runs are usually carried out from different starting configurations[2, 27, 70, 71] so as to extend the range of the search. *De novo* Rosetta protein structure prediction protocols often use several thousand MC fragment replacement trials per molecule[28]. This is not to overlook the utility of MC and other chain-type methods, but to suggest that at least some of their limitations are related to the fact that the structures they generate have search histories of varying lengths.

The limitations of parallel-generation methods involve their computational cost. To make the Z Method practical for complete template-free predictions of protein structure, more external statistical information may have to be included, if generating libraries of local (e.g., secondary) structure by the procedure internally, as is done in the current study for a limited number of loop backbone angles, proves insufficient. Adherence to “unbiased” sampling does not preclude the use of statistical information in energy-based structure prediction, although it does require that, for a given conformational subspace, the statistically derived conformers composing the library have the same degree of local optimization.

The use of fragment-based hierarchical build-up procedures, as in this work, is another technique for making Z Module calculations tractable. Since an exhaustive search of a 36-dihedral-angle space at 5- or 10° step sizes would involve $\approx 10^{56}$ – 10^{66} unique structures, the current work, which required the evaluation of a total of less than 10^8 structures, serves as a small-scale illustration of how a large reduction in the size of the search problem may be effectively achieved with this type of approach. The success of this and other structure prediction methods based on hierarchical techniques [23, 35, 43, 44, 45] is consistent with experimental[72, 73, 74] and theoretical[75, 76, 77, 78, 79, 80, 81, 82] observations suggesting that protein folding often involves hierarchical mechanisms[83]—i.e., local folding followed by larger-scale folding.

A particular advantage of the fragment-based hierarchical approach is that searches on parts of the molecule can be carried out independently of the rest. When partial-structure searches are performed and the whole molecule is included in the energy evaluation, there are by definition “bystander” parts of the molecule that need to be assigned (fixed) conformations. No matter the choice—e.g., random coil, conformations based on homology modeling or the results of previous searches, *etc.*—the presence of these parts of the structure introduces a bias into the regions of the potential energy hypersurface that are accessible to the subspaces being searched. As mentioned in Methods, the atomic subsystem corresponding to a set of conformational subspaces ideally includes atoms whose positions are entirely specified by the subspaces being searched and the known ICs. For this type of reason, in the present study it is only in the reworking stage, when the structure is presumably close enough to native—i.e., when sufficiently few rearrangements are required— that the entire structure is included in the calculations. On the other hand, because tertiary interactions appear to have a significant role in stabilizing some local protein structure—e.g., the best predictors of secondary structure from sequence information have a maximal accuracy of $\approx 80\%$ [84, 85]—such late-stage optimizations involving all or most of the structure appear to be a likely requirement. Moreover, some SSEs that are adjacent in sequence are not always spatially adjacent in the native structure[86, 87], so that searches involving these fragments in isolation are unlikely to result in native-like conformations, as is illustrated in the current results for the β_2 and β_3 strands of CheY. This raises the question of how to determine when the prediction is close enough to the native structure to proceed from the build-up to the reworking or whole-protein stage. Here, the switch was made naturally when the whole-protein structures were generated from the union of the 4 and 5-subspace fragments, but there are more systematic methods to determine native-like structures which may help determine whether more extensive fragment-based searching needs to be done[88, 89, 90, 91].

The Z Module calculations generated structures that in many cases were “correct” in terms of RMSD from native, but that differed significantly from the native in their backbone dihedral angle values. In the correct structures, these errors approximately cancel; in the incorrectly predicted structures, they do not. Most of the cancellation is due to approximately equal and opposite changes in ψ_i and ϕ_{i+1} , similar to, but smaller than, the “peptide-plane flip” changes previously described[51, 52]. However, because there is only one variable (ψ_i, ϕ_{i+1}) pair per loop in this problem, there are other types of ϕ/ψ contributions to the cancellation. Since the internal geometry of the structure is fixed here except for the variable backbone dihedral angles, these findings suggest that, for protein backbone chains with fixed bond angles and lengths, and whose ends are fixed, the sum of the ϕ and ψ angles tends towards a constant for alternative chain conformations. Similar observations have also been made by crystallographers (James Hogle, personal communication). Backbone dihedral angles in α -helices tend to fall near the line $\phi + \psi \approx -110^\circ$ [92, 93], and deviations in either angle tend to offset those in the other within a given helix. (In the current study of CheY, few of the variable residues have native structures in

the this region of the Ramachandran map.) In one theoretical study, the distribution of all the backbone torsional angle deviations for a protein undergoing local minimization was found consistently to have a mean near zero (and standard deviations of 2–38°)[94]. Here, a series of exploratory calculations using analytic ring closure algorithms to vary the ϕ and ψ angles in trimeric portions of protein backbones provided further evidence for this type of angular compensation, particularly for small dihedral angle changes (less than an average of $\approx 20^\circ$ per angle). Interestingly, preliminary calculations varying only ϕ backbone angles have demonstrated markedly less compensation. Overall, these findings indicate that the cancellation of errors in the main set of CheY predictions was not coincidental. In addition, they suggest ways to reduce the search space for certain protein loop prediction problems. Further investigation will be required.

In the hierarchical build-up procedure of the CheY protein, the global RMSD generally increased with the number of SSEs included in the comparisons, even when the quality of the predictions for the individual parts of the structure was essentially the same. It is generally recognized that RMSD tends to increase with the number of atoms in the system[95], and Carugo and Pondor[96] have described this statistical trend in protein structures and developed a formula for size-normalized RMSD. Partly for this reason[95], CASP competitions have introduced other measures of prediction accuracy, such as GDT-TS[97, 98], which have proven very useful. However, for the same fit, the RMSD,

$(\sum_{i=1}^N |\mathbf{r}_i - \mathbf{r}_{i,0}|^2 / N)^{1/2}$, where $\mathbf{r}_i$ and $\mathbf{r}_{i,0}$ are the position vectors for atom i in the test and reference structures, is formally constant in the number of atoms (N). As illustrated in a simple translation-only least-squares fit model in the Appendix, the tendency of the RMSD to increase with system size is purely a function of the global fit, itself, not of the RMSD as a measure of the fit. Other global, homogeneous measures of fit such as mean square deviation or those having simple linear terms, such as the sum of the absolute values of the differences in each coordinate, also tend to increase with system size. Since there are more atomic positions which must be simultaneously optimized, the fit tends to become worse for larger structures, even when the local RMSD—i.e., that between smaller parts of fixed size—remains, on average, constant. This is an inherent property of the fitting problem and does not reflect limitations in the fitting algorithms or methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks Matt Jacobson and the Friesner Group for providing the single-residue backbone libraries, and Felix Koziol, Robert Yelle, Andrei Golosov, Martin Spichty, and Milan Hodoscek for technical help. He specially thanks Martin Karplus of the Department of Chemistry and Chemical Biology at Harvard for many insightful discussions, his thoughtful reading of the manuscript, and his support of the work. Victor Ovchinnikov also provided useful suggestions for the manuscript. The ring closure software, by C. Seok, was provided by the Dill Group at the University of California, San Francisco. Evangelos Coutsias assisted with use of the software and generated the structures for the all- ϕ portion of the supplemental calculations with a modified version. The author also acknowledges James Hogle of Harvard Medical School for his insights. The Linux cluster of the Harvard Division of Science Research Computing was used for part of the MC/SA calculations (Intel Xeon). This work was supported in part by grants from the Eppley Foundation, Novo Nordisk, and the National Institutes of Health, including grant R01 RR023920. The latter was administered through the group of Charles L. Brooks, III, Dept of Chemistry, University of Michigan.

References

1. Anfinsen CB, Scheraga HA. Principles that govern the folding of protein chains. *Science*. 1973; 181:223–230. [PubMed: 4124164]

2. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* 1997; 268:209–225. [PubMed: 9149153]
3. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA.* 1999; 96:5482–5485. [PubMed: 10318909]
4. Simmerling C, Strockbine B, Roitberg AE. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* 2002; 124:11258–11259. [PubMed: 12236726]
5. Pitera JW, Swope W. Understanding folding and design: replica-exchange simulations of "Trp-cage" fly miniproteins. *Proc. Natl. Acad. Sci. USA.* 2003; 100:7587–7592. [PubMed: 12808142]
6. Lei H, Duan Y. Ab initio folding of albumin binding domain from all-atom molecular dynamics simulation. *J. Phys. Chem. B.* 2007; 111:5458–5463. [PubMed: 17458992]
7. Chowdhury S, Lee MC, Xiong G, Duan Y. Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* 2003; 327:711–717. [PubMed: 12634063]
8. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* 2003; 85:1145–1164. [PubMed: 12885659]
9. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science.* 2005; 309:1868–1871. [PubMed: 16166519]
10. Vieth M, Kolinski A, Brooks CL, Skolnick J. Predictions of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* 1994; 313:361–367. [PubMed: 8151697]
11. Lee MR, Tsai J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* 2001; 313:417–430. [PubMed: 11800566]
12. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Prot. Sci.* 2004; 13:211–220.
13. Misura KMS, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with Rosetta can be more accurate than their templates. *Proc. Natl. Acad. Sci.* 2005; 103:5361–5366. [PubMed: 16567638]
14. Chen J, Brooks CL III. Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins.* 2007; 67:922–930. [PubMed: 17373704]
15. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, Di-Maio F, Lange O, Kinch L, Sheffler W, Kim B, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins.* 2009; 77S9:89–99. [PubMed: 19701941]
16. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. *J Bioinform. Comput. Biol.* 2003; 1:95–117. [PubMed: 15290783]
17. Zhou H, Skolnick J. Protein structure prediction by pro-sp3-tasser. *Bioph. J.* 2009; 96:2119–2127.
18. Xu J, Peng J, Zhao F. Template-based and free modeling by RAPTOR++ in CASP8. *Proteins.* 2009; 77S9:133–137.
19. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. *Proteins.* 2009; 77S9:128–132. [PubMed: 19626712]
20. Dill K. Theory for the folding and stability of globular proteins. *Biochem.* 1985; 24:1501–1509. [PubMed: 3986190]
21. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA.* 1992; 89:2536–2540. [PubMed: 1557356]
22. Bromberg S, Dill KA. Side-chain entropy and packing in proteins. *Prot. Sci.* 1994; 3:997–1009.
23. Srinivasan R, Rose GD. Ab initio prediction of protein structure using LINUS. *Proteins.* 2002; 47:489–495. [PubMed: 12001227]
24. Mann M, Smith C, Rabbath M, Edwards M, Will S, Backofen R. CPSP-web-tools: a server for 3d lattice protein studies. *Bioinformatics.* 2009; 25:676–677. [PubMed: 19151096]
25. Rohl C, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. *Methods Enz.* 2004; 383:66–93.
26. Brunette TJ, Brock O. Improving protein structure prediction with model-based search. *Bioinformatics.* 2005; 21S1:i66–i74. [PubMed: 15961500]

27. Mamonov AB, Bhatt D, Cashman DJ, Ding Y, Zuckerman DM. General library-based Monte Carlo technique enables equilibrium sampling of semi-atomistic protein models. *J. Phys. Chem.* 2009; 113:10891–10904.
28. Bradley P, Baker D. Improved beta-protein structure prediction by multilevel optimization of non-local strand pairings and local backbone conformation. *Proteins.* 2006; 65:922–929. [PubMed: 17034045]
29. Das R, Baker D. Macromolecular modeling with Rosetta. *Ann. Rev. Biochem.* 2008; 77:363–382. [PubMed: 18410248]
30. Brooks BR, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* 1983; 4:187–217.
31. Brooks BR III, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* 2009; 30:1545–1614. [PubMed: 19444816]
32. Petrella RJ, Lazaridis T, Karplus M. Protein sidechain conformer prediction: a test of the energy function. *Folding Des.* 1998; 3:353–377.
33. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* 2001; 311:421–430. [PubMed: 11478870]
34. Petrella RJ, Karplus M. The energetics of off-rotamer protein sidechain conformations. *J. Mol. Biol.* 2001; 312:1161–1175. [PubMed: 11580256]
35. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. *Proteins.* 2004; 55:351–367. [PubMed: 15048827]
36. Marshall GR, Galaktionov S, Nikiforovich GV. Ab initio modeling of small, medium, and large loops in proteins. *Biopolymers.* 2001; 60:153–168. [PubMed: 11455548]
37. Petrella RJ, Karplus M. A limiting-case study of protein structure prediction: Energy-based searches of reduced conformational space. *J. Phys. Chem. B.* 2000; 104:11370–11378.
38. Lazaridis T, Karplus M. Effective energy function for proteins in solution. *Proteins.* 1999; 35:133–152. [PubMed: 10223287]
39. Markov, MM. *Dynamic Probabilistic Systems, volume 1: Markov Chains.* New York: John Wiley and Sons; 1971. reprinted in Appendix B
40. Brunette TJ, Brock O. Improving protein structure prediction with model-based search. *Bioinformatics.* 2005; 21 Suppl 1:i66–i74. [PubMed: 15961500]
41. Petrella RJ. (Manuscript in preparation).
42. Komazin-Meredith G, Petrella RJ, Santos WL, Filman DJ, Hogle JM, Verdine GL, Karplus M, Coen DM. The human cytomegalovirus UL44 C clamp wraps around DNA. *Structure.* 2009; 16:1214–1225. [PubMed: 18682223]
43. Vasquez M, Scheraga HA. Calculation of protein conformation by the build-up procedure. application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J. Biomol. Struc. Dyn.* 1988; 5:705–755.
44. Scheraga HA. Some approaches to the multiple-minima problem in the calculation of polypeptide and protein structures. *Int. J. Quantum Chem.* 1992; 42:1529–1536.
45. Shell MS, Ozkan SB, Voelz V, Wu GA, Dill KA. Blind test of the physics-based prediction of protein structures. *Biophys. J.* 2009; 96:917–924. [PubMed: 19186130]
46. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins.* 1999; S3:171–176. [PubMed: 10526365]
47. Rohl CA. Protein structure estimation from minimal restraints using ROSETTA. *Meth. Enzymol.* 2005; 394:244–260. [PubMed: 15808223]
48. Desmet J, DeMaeyer M, Hayes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature.* 1992; 356:539–542. [PubMed: 21488406]
49. Lasters I, DeMaeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side-chains. *Protein Eng.* 1995; 8:815–822. [PubMed: 8637851]

50. Adcock SA. Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield. *J. Comput. Chem.* 2004; 25:16–27. [PubMed: 14634990]
51. Hayward S. Peptide-plane flipping in proteins. *Protein Sci.* 2001; 10:2219–2227. [PubMed: 11604529]
52. Milner-White JE, Watson JD, Qi G, Hayward S. Amyloid formation may involve α to β sheet interconversion via peptide plane flipping. *Structure.* 2006; 14:1369–1376. [PubMed: 16962968]
53. Wu X, Brooks BR. Self-guided langevin dynamics simulation method. *Chem Phys Lett.* 2003; 381:512–518.
54. Andricioaei I, Dinner AR, Karplus M. Self-guided enhanced sampling methods for thermodynamic averages. *J. Chem. Phys.* 2003; 118:1074–1084.
55. Berman HM, Henrick K, Nakamura H. Announcing the worldwide protein data bank. *Nature Struc Biol.* 2003; 10:980.
56. Petrella RJ, Andricioaei I, Brooks BR, Karplus M. An improved method for nonbonded list generation: Rapid determination of near-neighbor pairs. *J. Comput. Chem.* 2003; 24:222–231. [PubMed: 12497601]
57. Coutsias EA, Seok C, Jacobson MP, Dill KA. A kinematic view of loop closure. *J. Comput. Chem.* 2004; 25:510–528. [PubMed: 14735570]
58. Yang M, Lei M, Yordanov B, Huo S. Peptide plane flip can flip in two opposite directions: implication in amyloid formation of transthyretin. *J. Phys. Chem.* 2006; 110:5829–5833.
59. Nicosia G, Stracquadanio G. Generalized pattern search algorithm for peptide structure prediction. *Biophys. J.* 2008; 95:4988–4999. [PubMed: 18487293]
60. Schlierf M, Rief M. Single-molecule unfolding force distributions reveal a funnel-shaped energy landscape. *Biophys. J.* 2006; 90:L33–L35. [PubMed: 16361331]
61. Simler BR, Levy Y, Onuchic JN, Matthews CR. The folding energy landscape of the dimerization domain of Escherichia coli trp repressor: A joint experimental and theoretical investigation. *J. Mol. Biol.* 2006; 363:262–278. [PubMed: 16956620]
62. Akturk E, Arkin H. The structure of the free energy surface of coarse-grained off-lattice protein models. *Internat. J. Mod. Phys. C.* 2007; 18:99–106.
63. Kondov I, Verma A, Wenzel W. Folding path and funnel scenarios for two small disulfide-bridged proteins. *Biochem.* 2009; 48:8195–8205. [PubMed: 19610617]
64. Hori N, Chikenji G, Berry RS, Takada S. Folding energy landscape and network dynamics of small globular proteins. *Proc. Natl. Acad. Sci. USA.* 2009; 106:73–78. [PubMed: 19114654]
65. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science.* 1983; 220:671–680. [PubMed: 17813860]
66. Berg BA, Neuhaus T. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B.* 1991; 267:249–253.
67. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* 1997; 281:140–150.
68. Andricioaei I, Straub J. On Monte Carlo and molecular dynamics methods inspired by Tsallis statistics: Methodology, optimization, and application to atomic clusters. *J. Chem. Phys.* 1997; 907:9117–9124.
69. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA.* 2000; 97:10383–10388. [PubMed: 10984534]
70. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* 2001; 306:1191–1199. [PubMed: 11237627]
71. Bonneau R, Strauss CEM, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* 2002; 322:65–78. [PubMed: 12215415]
72. Munoz V, Blanco FJ, Serrano L. The hydrophobic-staple motif and a role for loop-residue in α -helix stability and protein folding. *Nat. Struct. Biol.* 1995; 2:380–385. [PubMed: 7664095]
73. Blanco FJ, Rivas G, Serrano L. A short linear peptide that folds into a native stable β -hairpin in aqueous solution. *Nat. Struct. Biol.* 1994; 1:584–590. [PubMed: 7634098]

74. Viguera AR, Serrano L. Experimental analysis of the schellman motif. *J. Mol. Biol.* 1995; 251:150–160. [PubMed: 7643384]
75. Bystroff C, Garde S. Helix propensities of short peptides: molecular dynamics versus bioinformatics. *Proteins.* 2003; 50:552–562. [PubMed: 12577261]
76. Ho BK, Dill KA. Folding very short peptides using molecular dynamics. *PLoS Comput. Biol.* 2006; 2:e27. [PubMed: 16617376]
77. Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* 2001; 8:552–558. [PubMed: 11373626]
78. Lesk AM, Rose GD. Folding units in globular proteins. *Proc. Natl. Acad. Sci. USA.* 1981; 78:4304–4308. [PubMed: 6945585]
79. Krishna MMG, Maity H, Rumbley JN, Lin Y, Englander SW. Order of steps in the cytochrome c folding pathway: evidence for a sequential stabilization mechanism. *J. Mol. Biol.* 2006; 359:1411–1420.
80. Callender RH, Dyer RB, Gilmanishin R, Woodruff WH. Fast events in protein folding: the time evolution of primary processes. *Ann. Rev. Phys. Chem.* 1998; 49:173–202. [PubMed: 9933907]
81. Debe DA, Carlson MJ, Goddard WA. The topomer-sampling model of protein folding. *Proc. Natl. Acad. Sci. USA.* 1999; 96:2596–2601. [PubMed: 10077555]
82. Makarov DE, Keller CA, Plaxco KW, Metiu H. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci. USA.* 2002; 99:3535–3539. [PubMed: 11904417]
83. Baldwin RL, Rose G. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends in Biochemical Sciences.* 1999; 1:26–33. [PubMed: 10087919]
84. Karypis G. Yasspp: Better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins.* 2006; 64:575–586. [PubMed: 16763996]
85. Kountouris P, Hirst JD. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinformatics.* 2009; 10:437. [PubMed: 20025785]
86. Shih ESC, Hwang M-J. Alternative alignments from comparison of protein structures. *Proteins.* 2004; 56:519527.
87. Kolbeck B, May P, Schmidt-Goenner T, Steinke T, Knapp E-W. Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinform.* 2006; 7:510.
88. Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins.* 1999; 34:82–95. [PubMed: 10336385]
89. Tosatto SCE. The Victor/FRST function for model quality estimation. *J. Comput. Biol.* 2005; 10:1316–1327. [PubMed: 16379537]
90. Bahadur RP, Chakrabarti P. Discriminating the native structure from decoys using scoring functions based on the residue packing in globular proteins. *BMC Struc. Biol.* 2009; 9:76.
91. Benkert P, Tosatto SCE, Schwede T. Global and local model quality estimation at CASP8 using the scoring functions qmean and qmeanclust. *Proteins.* 2009; 77S9:173–180. [PubMed: 19705484]
92. Barlow DJ, Thornton JM. Helix geometry in proteins. *J. Mol. Biol.* 1988; 201:601–619. [PubMed: 3418712]
93. Hovmoller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. *Acta Crystallographica.* 2002; D58:768–776.
94. Bruccoleri RE, Karplus M. Spatially constrained minimization of macromolecules. *J Comput Chem.* 1986; 7:165–175.
95. Cozzetto D, Kryshstafovich A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. *Proteins.* 2009; 77S9:18–28. [PubMed: 19731382]
96. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* 2001; 10:1470–1473. [PubMed: 11420449]
97. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins.* 1999; 3:22–29. [PubMed: 10526349]

98. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. *Proteins*. 2001; 5:13–21. [PubMed: 11835478]

Appendix. Translation-only Least-Squares-Fit RMSD Model

A translation-only least-squares fit model can be used to illustrate how the global RMSD between two molecules can increase with size, even if the local RMSD remains constant. Consider an N -atom model in which the test structure has initial coordinates of

$x_k = x_k^0 \pm D_x$, $y_k = y_k^0 \pm D_y$, and $z_k = z_k^0 \pm D_z$, where (x_k^0, y_k^0, z_k^0) are the coordinates of the k th atom in the reference structure, and D_x , D_y , and D_z are fixed distances. The test structure therefore has an atom at one corner of a rectangular box centered at each corresponding

reference atom position, at a distance of $D_r = \sqrt{D_x^2 + D_y^2 + D_z^2}$. For the N atoms, assume R_x , R_y and R_z test atoms are to the right and $L_x = N - R_x$, $L_y = N - R_y$ and $L_z = N - R_z$ atoms are to the left of the reference atoms in the x , y and z directions, respectively, and that the test atoms are assigned randomly to either the left or the right in each direction. The test system is then adjusted with a translation-only (i.e., rotationless) least-squares fit to the reference system, so that the new coordinates are

$x'_k = x_k^0 \pm D_x + \Delta x$, $y'_k = y_k^0 \pm D_y + \Delta y$, and $z'_k = z_k^0 \pm D_z + \Delta z$, where Δx , Δy and Δz are the adjustments due to the fit.

The RMSD after the fit is given by

$$\begin{aligned} RMSD &= \left(\frac{\sum_{k=1}^N (x'_k - x_k^0)^2 + \sum_{k=1}^N (y'_k - y_k^0)^2 + \sum_{k=1}^N (z'_k - z_k^0)^2}{N} \right)^{1/2} = \\ &= \left(\frac{\sum_{k=1}^N (x_k^0 \pm D_x + \Delta x - x_k^0)^2 + (y_k^0 \pm D_y + \Delta y - y_k^0)^2 + (z_k^0 \pm D_z + \Delta z - z_k^0)^2}{N} \right)^{1/2} = \\ &= \left(\frac{\sum_{k=1}^N (\pm D_x + \Delta x)^2 + \sum_{k=1}^N (\pm D_y + \Delta y)^2 + \sum_{k=1}^N (\pm D_z + \Delta z)^2}{N} \right)^{1/2} = \\ &= \left((D_x - \Delta x)^2 + \frac{4D_x R_x \Delta x}{N} + (D_y - \Delta y)^2 + \frac{4D_y R_y \Delta y}{N} + (D_z - \Delta z)^2 + \frac{4D_z R_z \Delta z}{N} \right)^{1/2}, \end{aligned} \quad (A1)$$

since $L_x = N - R_x$, $L_y = N - R_y$ and $L_z = N - R_z$.

Optimizing the mean-square deviation (or, equivalently, performing a least-squares fit of the atomic positions) with respect to Δx , Δy and Δz ,

$$\begin{aligned} \frac{\partial (RMSD)^2}{\partial \Delta x} &= -2(D_x - \Delta x) + \frac{4D_x R_x}{N} = 0 \\ \frac{\partial (RMSD)^2}{\partial \Delta y} &= -2(D_y - \Delta y) + \frac{4D_y R_y}{N} = 0 \end{aligned}$$

and

$$\frac{\partial (RMSD)^2}{\partial \Delta z} = -2(D_z - \Delta z) + \frac{4D_z R_z}{N} = 0$$

so that

$$\begin{aligned}\Delta x &= \frac{D_x(N-2R_x)}{N}, \\ \Delta y &= \frac{D_y(N-2R_y)}{N}, \text{ and} \\ \Delta z &= \frac{D_z(N-2R_z)}{N}.\end{aligned}\quad (\text{A2})$$

Alternatively, the RMSD can be optimized directly to give the same result. Hence, the optimized coordinates of the test system are $x'_k = x_{k,F} \pm D_x + D_x(N - 2R_x)/N$, $y'_k = y_{k,F} \pm D_y + D_y(N - 2R_y)/N$, and $z'_k = z_{k,F} \pm D_z + D_z(N - 2R_z)/N$. Substituting Δx , Δy and Δz from the equations in (A2) into Eq.(A1), the RMSD in the optimized alignment is found to be

$$\frac{2}{N}(D_x^2 R_x(N - R_x) + D_y^2 R_y(N - R_y) + D_z^2 R_z(N - R_z))^{1/2}. \quad (\text{A3})$$

To calculate the expectation value of the RMSD, sum the RMSD over all states of the system and normalize by dividing by 2^{3N} possible states. For each $R_x \in \{0, 1, 2, \dots, N\}$ there

are $\binom{N}{R_x}$ equivalent x -coordinate states for the system. Hence,

$$\langle \text{RMSD} \rangle = \frac{2}{N 2^{3N}} \sum_{R_x=0}^N \sum_{R_y=0}^N \sum_{R_z=0}^N \binom{N}{R_x} \binom{N}{R_y} \binom{N}{R_z} (D_x^2 R_x(N - R_x) + D_y^2 R_y(N - R_y) + D_z^2 R_z(N - R_z))^{1/2} \quad (\text{A4})$$

This result is exact and can be calculated numerically for any specified $N \geq 1$. An approximate, closed-form analytic result can be obtained by effectively taking the square root of the average mean square deviation,

$$\begin{aligned}\langle \text{RMSD} \rangle &\approx (\langle \text{MSD} \rangle)^{1/2} \\ &= \frac{2}{N} \left(\frac{1}{2^{3N}} \sum_{R_x=0}^N \sum_{R_y=0}^N \sum_{R_z=0}^N \binom{N}{R_x} \binom{N}{R_y} \binom{N}{R_z} (D_x^2 R_x(N - R_x) + D_y^2 R_y(N - R_y) + D_z^2 R_z(N - R_z)) \right)^{1/2} \\ &= \left(\frac{(D_x^2 + D_y^2 + D_z^2)(N - 1)}{N} \right)^{1/2} \\ &= D_T \left(\frac{N - 1}{N} \right)^{1/2},\end{aligned}\quad (\text{A5})$$

where D_T is the initial atomic displacement, as mentioned earlier. Thus, the expectation value for the global RMSD increases with system size. Table 5 gives the numerical results for the exact and approximated values of the RMSD, as a fraction of D_T , for different system sizes, N . With the (trivial) exception of $N = 1$, the approximation is better with larger N . For very large N the RMSD converges to D_T .

The average squared change due to the fit is

$$\begin{aligned}
 & \langle (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \rangle \\
 &= \frac{2}{2^{3N}} \sum_{R_x=0}^N \sum_{R_y=0}^N \sum_{R_z=0}^N \binom{N}{R_x} \binom{N}{R_y} \binom{N}{R_z} \left(\left(\frac{D_x(N-2R_x)}{N} \right)^2 + \left(\frac{D_y(N-2R_y)}{N} \right)^2 + \left(\frac{D_z(N-2R_z)}{N} \right)^2 \right) \\
 &= \frac{D_x^2 + D_y^2 + D_z^2}{N} = D_T^2 / N
 \end{aligned}$$

Hence, the displacement due to fitting tends to become smaller with increasing N , and for very large N the fit effectively leaves the test structure in its initial position—i.e., with no realignment.

Since the model is homogeneous—all fragments of the structure of size $N_0 < N$ will have, on average, a constant RMSD corresponding to $\langle MSD \rangle^{1/2} = D_T ((N_0-1)/N_0)^{1/2}$ —it demonstrates that even with a constant local RMSD, the expected global RMSD increases with the size of the system, because the fit becomes worse.

1) Set-up

2) Loading of the subspaces to be searched

3) Specification of the rest of the search parameters:

- # subspaces to be searched in each region
- geometrical and IC filters
- energy cutoffs
- minimization, etc.

4) Search

Figure 1.

General scheme for performing a Z Module calculation. (1) The problem degrees of freedom are grouped into conformational subspaces, which can be of different sizes (i.e., different numbers of DOF); internal coordinates for the conformers to be searched are specified; the conformers are grouped into regions (here “a” or “b”). (2) The W subspaces to be explored in the search are selected or “loaded” into the module. (In this example, $W = 5$.) The loading feature allows the user to search different parts of the system without having to change the set-up. (3) The other parameters for the search are specified, including the number of subspaces (W_a and W_b) to be searched in each region for each search grid (here, $W_a = 1$; $W_b = 2$). (4), The search is carried out. For each part of the search, called a grid, the conformers

from the regions assigned to that grid are combined. (Here there are $W!/((W - W_a - W_b)! W_a! W_b!) = 5!/(2!1!2!) = 30$ total grids in the search.) See text.

2 A

2 B

Figure 2.

CheY structures predicted with the Z Module in the main set of calculations. Panel A) shows the structure predicted after the hierarchical build-up stage (Stage 1). Panel B) shows a superposition of the final predicted structure (cobalt blue) and the native CheY structure (blue-green). Note the native structure is slightly more compact.

3A

3B

3C

3D

Figure 3. Plots of energy vs. RMSD for the lowest-energy CheY structures in the first step of Stage 1, main set of Z Module calculations, in which single loops were used. A) Loop 1; B) Loop 2; C) Loop 3 and D) Loop 4.

4A

4B

Figure 4. Plot of energy vs. RMSD for the lowest-energy CheY structures in A) the last step of Stage 1, in which subspaces corresponding to the two halves of the protein were joined, and B) the last step of Stage 2, in which the whole protein structure was used in the search.

Figure 5.

Results of dihedral angle compensation studies using an analytic ring-closure algorithm for backbone dihedral angles in trimeric portions of turkey egg lysozyme (black), CheY (green), and adipocyte binding protein (red). For each alternative set of six dihedral angles, composed of 3 sets of (ϕ, ψ) combinations, the vertical axis plots the sum (Σ_2) of the absolute magnitudes of the deviations of these angles from the values in the native structure. The horizontal axis plots the sum (Σ_1) of the deviations, themselves. See also text.

Table 1

Details of the prediction protocol for both stages of the CheY calculations. Key: **step**—step number within the stage. **searches**—the number of separate searches performed in each step. **grids/search**—the number of search grids involved in each search. **loops/search**—the number of loops involved in each search calculation. This is the size of the product subspace (multiply by 4 to obtain the total number of degrees of freedom in the calculation). **conformers/search**—the number of conformers (gridpoints) evaluated in each of the searches. The total number of gridpoints per step is $\approx$ searches $\times$ conformers/search. **total CPU time**—the total CPU time for all searches in the step (in hours). The total times for each stage and for the entire prediction protocol (last row) are also indicated.

Stage 1					
step	# searches	grids/search	loops/search	conformers/search	total CPU t
1	9	1	1	160–260,000	1.26
2	4	1	2 or 3	1.0×10^6	5.92
3	2	1	4 or 5	1.0×10^6	5.06
4	1	1	9	1.0×10^6	5.02
tot time Stage 1					17.26
Stage 2					
step	searches	grids/search	loops/search	conformers/search	total CPU t
1	9	1	1	$2.5\text{--}4.5 \times 10^6$	9.13
2	2	10	5	2.5×10^6	15.14
		6	4	1.5×10^6	
3	1	1	9	4.0×10^6	19.02
tot time Stage 2					43.29
tot time both Stages					60.55

Table 2

All-atom RMSD from the native CheY structure at various stages of the calculations. Key: **loop**—the loop number. **single**—the results after the single-loop searches of Stage 1. **Stage 1**—the results for the final Stage 1 structure. **Stage 2**—the results for the final Stage 2 structure. Rows corresponding to loops 1 through 9 give the RMSD for the subsystem corresponding to the loop at each of the three stages. The **all** row gives the results for the entire structure at each stage.

loop	single	Stage 1	Stage 2
1	0.384	0.248	0.248
2	3.970	0.566	0.566
3	*0.905	0.804	0.538
	*5.318		
4	3.420	0.712	0.712
5	1.623	1.626	0.385
6	0.417	11.298	0.284
7	2.328	0.537	0.537
8	4.396	0.561	0.561
9	0.453	0.265	0.377
all	15.683	20.458	1.555

The asterisk (*) indicates degenerate conformers.

Table 3

Dihedral angle values for the Stage 1 and Stage 2 structures in the CheY calculations. **Key:** **loop**—the number of the loop. **DOF**—the integer corresponding to the degree of freedom (dihedral angle). **native**—the value of the dihedral angle in the native structure. **Stage 1** and **Stage 2**—the value of the dihedral angle in the Stage 1 and Stage 2 structures. **error**—the error in the dihedral angle value relative to the native value, in either the Stage 1 or Stage 2 structure. The numbers in bold are the sums of the errors in the dihedral angle values over each loop, and the last row (**sum of errors**) gives the sum of the errors over each entire structure.

loop	DOF	native	Stage 1	error	Stage 2	error
1	1	-166.92	-170	-3.08	-170	-3.08
	2	147.62	160	12.38	160	12.38
	3	-84.77	-100	-15.23	-100	-15.23
	4	-3.98	0	3.98	0	3.98
				-1.95		-1.95
2	5	-101.47	-110	-8.53	-110	-8.53
	6	-5.13	40	45.13	40	45.13
	7	-86.95	-130	-43.05	-130	-43.05
	8	84.07	100	15.93	100	15.93
				9.48		9.48
3	9	-90.83	-100	-9.17	-105	-14.17
	10	-19.96	-70	-50.04	-70	-50.04
	11	-166.42	-130	36.42	-115	51.42
	12	-179.96	-170	9.96	180	-0.04
				-12.83		-12.83
4	13	80.02	80	-0.02	80	-0.02
	14	179.61	-160	20.39	-160	20.39
	15	78.58	70	-8.58	70	-8.58
	16	10.46	0	-10.46	0	-10.46
				1.33		1.33
5	17	-100.31	-90	10.31	-95	5.31
	18	101.32	110	8.68	125	23.68
	19	-141.46	-140	1.46	-165	-23.54

loop	DOF	native	Stage 1	error	Stage 2	error
	20	152.06	150	-2.06	150	-2.06
				18.39		3.39
6	21	-54.62	-150	-95.38	-55	-0.38
	22	-31.69	-170	-138.31	-25	6.69
	23	-88.66	-80	8.66	-95	-6.34
	24	-18.87	-30	-11.13	-15	3.87
				123.84		3.84
7	25	-164.61	-170	-5.39	-170	-5.39
	26	123.34	150	26.66	150	26.66
	27	-68.87	-80	-11.13	-80	-11.13
	28	130.34	120	-10.34	120	-10.34
				-0.20		-0.20
8	29	107.03	100	-7.03	100	-7.03
	30	13.12	20	6.88	20	6.88
	31	-62.31	-60	2.31	-60	2.31
	32	143.60	140	-3.60	140	-3.60
				-1.44		-1.44
9	33	-141.90	-140	1.90	-135	6.90
	34	162.89	170	7.11	175	12.11
	35	-102.30	-110	-7.70	-115	12.70
	36	173.97	170	-3.97	170	-3.97
				-2.66		2.34
	sum of errors			133.96		3.96

Table 4

Results of MC and simulated annealing studies. Key: # **loops**—the number of consecutive loops (with their flanking secondary structural elements) that are included in the calculation. For the 1-loop calculation, loop 1 was used; for the 2-loop calculation, loops 1 and 2; for the 5-loop calculation, loops 5 through 9. # **res**—the number of residues which were varied at a time (in a given MC step), over the number of total varying residues. **T**—temperature, in degrees Kelvin; SA indicates simulated annealing calculation ($T_{initial} = 10000K$ and $T_{final} = 300$); see methods. **steps**—the number of steps in the calculation (in millions). **final**—the step number at which the final structure was attained. **RMSD**—the RMSD of the final structure from the native, in Å, for the portion of the structure being studied. **CPU t**—the CPU time, in hours, on the Intel Xeon cluster, unless *, which indicates calculations performed on the AMD Opterons. For comparison, the RMSDs obtained in the Z Module calculations for the various parts of the structure were: 1-loop, 0.15 Å 2-loop, 0.27 Å; 5-loop, 2.94 Å; 9-loop, 1.56 Å. The 1-, 2-, and 5-loop Z Module results are for build-up procedures only, without refinement steps.

# loops	# res	T	steps	final	RMSD	CPU t
1	1/2	300	5.0	2057	0.39	*16.3
1	1/2	10000	5.0	4999999	11.61	
1	1/2	10000	5.0	5000000	12.35	21.0
1	2/2	300	2.5	2195214	0.38	14.6
1	2/2	10000	2.5	2499999	5.91	14.1
2	1/4	300	5.0	8765	4.48	*16.4
2	2/4	300	5.0	42164	4.48	*18.0
2	2/4	300	5.0	709921	0.36	
2	2/4	10000	5.0	5000000	3.75	24.2
2	2/4	SA	1.0	984854	0.52	7.2
2	4/4	300	1.0	636665	6.96	
2	4/4	300	1.0	47204	3.76	*4.7
2	4/4	10000	1.0	999964	12.79	
2	4/4	SA	1.0	993123	11.81	11.5
5	2/10	300	1.0	889125	15.20	7.7
5	2/10	10000	1.0	999997	21.80	8.5
5	2/10	SA	1.0	998837	10.12	7.1
5	2/10	SA	5.0	4871823	13.91	35.1
5	4/10	300	1.0	941452	18.72	
5	4/10	10000	1.0	999979	17.12	11.9
5	4/10	SA	1.0	945357	17.18	11.3
5	4/10	SA	5.0	4638264	15.75	31.5

# loops	# res	T	steps	final	RMSD	CPU t
5	8/10	300	1.0	129577	24.59	
5	8/10	10000	1.0	999964	29.58	
5	8/10	SA	1.0	710333	17.19	14.4
5	8/10	SA	5.0	4549704	17.60	70.1
9	4/18	300	1.0	941022	19.13	8.9
9	4/18	10000	1.0	265554	28.20	
9	4/18	SA	1.0	990838	22.33	
9	4/18	SA	5.0	4640265	20.13	77.0
9	8/18	10000	1.0	249895	39.09	
9	8/18	SA	1.0	836330	25.38	22.3
9	8/18	SA	1.0	886288	19.45	
9	8/18	SA	5.0	4459613	20.59	129.8
9	8/18	10000	0.1	58790	30.38	20.5

Table 5

RMSDs after optimal fit of the test and reference systems in the translation-only LSQ fit model described in Appendix 1. The results are given as a fraction of the initial atomic displacement, D_T , for different system sizes. Key: **N**—size of system (number of atoms). **exact**—exact mean RMSD, calculated from eq. (A4). **approx**—mean RMSD, approximated from $(\langle MSD \rangle)^{1/2}$ (eq. A5).

N	RMSD	
	exact (Å)	approx (Å)
1	0.00000	0.00000
2	0.50000	0.70711
3	0.70711	0.81650
4	0.80801	0.86603
5	0.86237	0.89443
10	0.94480	0.94868
20	0.97396	0.97468
30	0.98289	0.98319
50	0.98984	0.98995
100	0.99496	0.99499

Table 6

(Supplementary Material) An example of the input conformer file in (A) uncompressed and (B) compressed forms for a Z Module calculation. Key: **subsp**—subspace number. **conf**—conformer number. **DOF**—degree of freedom number. **value**—internal coordinate value corresponding to DOF, in angular degrees for angles and in Å for bond lengths.

A.				
subsp	conf	DOF	value	
1	1	7	18.48579	
1	1	3	180.74837	
1	1	1	50.10986	
1	2	7	24.28368	
1	2	3	180.74837	
1	2	1	50.10986	
1	3	7	30.09877	
1	3	3	180.16835	
1	3	1	60.87635	
2	4	8	64.27840	
2	4	9	80.28763	
2	4	11	17.98138	
2	5	8	56.97623	
2	5	9	80.28763	
2	5	11	1.89237	
B.				
subsp	conf	DOF	value	
1	1	1	50.10986	
1	1	3	180.74837	
1	1	7	18.48578	
1	2	7	24.28367	
1	3	1	60.87635	
1	3	3	180.16835	
1	3	7	30.09877	
2	4	8	64.27840	
2	4	9	80.28763	
2	4	11	17.98138	
2	5	8	56.97623	
2	5	11	1.89237	

Table 7

(Supplementary Material) Energy analysis for predicted structures in the main set of CheY calculations. Key: **struc**—structure for which the energies are calculated; individually predicted loops (and their flanking secondary structural elements) are labeled loop 1-loop 9; “Stage 1” and “Stage 2” refer to the final predicted structures in those stages. Columns 2–6 give the energy differences (total and components) between predicted and native structures: $\Delta E = E_{\text{predicted}} - E_{\text{native}}$. (All energies in kcal/mol.) **total**—the total energy; **vdW**—the van der Waals energy; **elec**—the electrostatic energy; **dihe**—the dihedral angle energy; **solv**—the EEF1 solvation term. Column 7 (**RMSD**) gives the RMSD from native, in Å. **Panel A** shows the original results, which use the unminimized native structure as a reference. **Panel B** shows the results using the native structure after 10 steps of Steepest Descent minimization. The all-atom RMSD of this structure from the unoptimized native is 0.08 Å. The dihedral angle energy differences are identical to within 3 decimal places in the two panels because the 10-step minimization leaves the variable dihedral angles in the study essentially unchanged.

struc	total	vdW	elec	dihe	solv	RMSD
A						
loop 1	1.728	0.621	0.206	0.163	0.737	0.384
loop 2	-30.150	-15.007	0.462	0.028	-15.633	3.970
loop 3	-3.296	-1.138	0.794	-0.196	-2.756	0.905
	-3.292	1.115	-1.449	0.064	-3.022	5.318
loop 4	-6.440	-2.922	-6.839	0.319	3.003	3.420
loop 5	-13.438	-11.963	1.007	-0.134	-2.348	1.623
loop 6	4.624	0.673	7.317	0.009	-3.376	0.417
loop 7	-9.780	-9.566	-6.954	0.216	6.524	2.328
loop 8	-4.287	-3.858	-3.101	0.514	2.158	4.396
loop 9	4.136	3.311	3.318	-0.004	-2.489	0.453
Stage 1	141.915	156.124	136.141	1.246	-151.596	20.458
Stage 2	84.199	74.535	85.936	1.084	-77.356	1.555
B						
loop 1	1.206	-0.292	0.212	0.163	1.123	0.328
loop 2	-4.241	1.477	7.565	0.028	-13.311	4.106
loop 3	-1.760	-0.145	1.443	-0.196	-2.863	0.849
	-0.876	1.954	0.001	0.064	-2.896	5.314
loop 4	-3.648	0.026	-6.671	0.319	2.679	3.456
loop 5	6.090	6.029	2.316	-0.134	-2.121	1.936

struc	total	vdW	elec	dlhe	solv	RMSD
loop 6	9.986	7.662	5.572	0.009	-3.258	0.580
loop 7	-9.168	-9.444	-7.919	0.216	7.979	2.378
loop 8	26.901	29.834	-6.212	0.514	2.765	4.315
loop 9	-7.455	-4.533	0.306	-0.004	-3.224	0.530