Modeling of loops in protein structures

András Fiser, Richard Kinh Gian Do and Andrej Šali

Laboratories of Molecular Biophysics

Pels Family Center for Biochemistry and Structural Biology

The Rockefeller University

1230 York Ave, New York, NY 10021, USA

Correspondence to Andrej Šali

tel: (212) 327 7550; fax: (212) 327 7540

e-mail: sali@rockefeller.edu

Running title: Modeling of loops in protein structures

Manuscript length: 60 pages incl. 2 tables and 13 figures

Electronic version: `http://guitar.rockefeller.edu/ftp/protsci/`

Word processing: The manuscript is prepared in LaTex under Linux

Abbreviations used: 3D, three-dimensional; DRMS, distance root-mean-square;

NMR, nuclear magnetic resonance; PDB, Protein Data Bank; RMSD, root-mean-square deviation.

May 24, 2000

Abstract

Comparative protein structure prediction is limited mostly by the errors in alignment and loop modeling. We describe here a new automated modeling technique that significantly improves the accuracy of loop predictions in protein structures. The positions of all non-hydrogen atoms of the loop are optimized in a fixed environment with respect to a pseudo energy function. The energy is a sum of many spatial restraints that include the bond length, bond angle, and improper dihedral angle terms from the CHARMM-22 force field, statistical preferences for the mainchain and sidechain dihedral angles, and statistical preferences for non-bonded atomic contacts that depend on the two atom types, their distance through space, and separation in sequence. The energy function is optimized with the method of conjugate gradients combined with molecular dynamics and simulated annealing. Typically, the predicted loop conformation corresponds to the lowest energy conformation among 500 independent optimizations. Predictions were made for 40 loops of known structure at each length from 1 to 14 residues. The accuracy of loop predictions is evaluated as a function of thoroughness of conformational sampling, loop length, and structural properties of native loops. When accuracy is measured by local superposition of the model on the native loop, 100, 90 and 30% of 4, 8 and 12-residue loop predictions, respectively, had less than 2Å RMSD error for the mainchain N, C_α, C, and O atoms; the average accuracies were 0.59 ± 0.05, 1.16 ± 0.10 and 2.61 ± 0.16 Å, respectively. To simulate real comparative modeling problems, the method was also evaluated by predicting loops of known structure in only approximately correct environments with errors typical of comparative modeling without misalignment. When the RMSD distortion of the mainchain stem atoms is 2.5Å, the average loop prediction error increased by 180, 25 and 3% for 4, 8 and 12-residue loops, respectively. The accuracy of the lowest energy prediction for a given loop can be estimated from the structural variability among a number of low energy predictions. The relative value of the present method is gauged by (i) comparing it with one of the most successful previously described methods, and (ii) describing its accuracy in recent blind predictions of protein structure. Finally, it is shown that the average accuracy of prediction is limited primarily by the accuracy of the energy function rather than by the extent of conformational sampling.

Keywords: loop modeling / comparative or homology protein structure modeling

Introduction

Functional characterization of a protein sequence is one of the most frequent and challenging problems in biology. This task is usually facilitated by accurate three-dimensional (3D) structures of the studied protein and corresponding ligand complexes. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful 3D model for a protein (target) that is related to at least one known protein structure (template) (Browne *et al.*, 1969; Blundell *et al.*, 1987; Martí-Renom *et al.*, 2000).

Comparative modeling is limited in functional studies by its ability to predict accurately structural determinants of protein function, as well as by the conformational changes induced by ligand binding (Wang *et al.*, 1999). In the absence of an induced fit, the function of a protein is generally determined by shape, dynamics and physiochemical properties of its solvent exposed molecular surface. Likewise, functional differences between the members of the same protein family are usually a consequence of the structural differences on the protein surface. In a given fold family, structural variability is a result of substitutions, insertions and deletions of residues between members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold. Thus, loops often determine the functional specificity of a given protein framework. They contribute to active and binding sites. Examples include binding of metal ions by metal-binding proteins (Lu & Valentine, 1997), small protein toxins by their receptors (Wu & Dean, 1996), antigens by immunoglobulins (Bajorath & Sheriff, 1996), mononucleotides by a variety of proteins (Kinoshita *et al.*, 1999), protein substrates by serine proteases (Perona & Craik, 1995), and DNA by DNA-binding proteins (Jones *et al.*, 1999). Consequently, the accuracy of loop modeling is a major factor determining the usefulness of comparative models in studying interactions between the protein and its ligands. This includes the use of models for recognizing ligand binding sites (Jones & Thornton, 1997; Fetrow *et al.*, 1998; Russell *et al.*, 1998; Kleywegt, 1999; Wei *et al.*, 1999; Kasuya & Thornton, 1999) and for ligand docking computations (Kick *et al.*, 1997). Unfortunately, as was concluded at the meetings on Critical Assessment of Techniques for Protein Structure Prediction (CASP), no generally reliable method is available for constructing loops longer than 5 residues (Mosimann *et al.*, 1995; Martin *et al.*, 1997), although recently some progress has been made (van Vlijmen & Karplus, 1997; Rufino *et al.*, 1997; Oliva *et al.*, 1997; Samudrala & Moult, 1998; Rapp & Friesner, 1999).

The impact of an accurate loop modeling method would be large. Currently, approximately 40% of all protein sequences can have at least one domain modeled on a related, known protein structure (Rychlewski *et al.*, 1998; Huynen *et al.*, 1998; Jones, 1999; Sánchez & Šali, 1999). At least two thirds of the comparative modeling cases are based on less than 40% sequence identity between the target and the templates, and thus generally require loop modeling (Sánchez & Šali, 1998). Since there are over 500,000 protein sequences deposited in GENBANK and only approximately 12,000 protein structures in the Protein Data Bank (PDB) (Abola *et al.*, 1987; `http://www.rcsb.org/pdb`), the number of proteins whose structure can be modeled by comparative modeling is more than an order of magnitude larger than the number of currently known protein structures (Šali, 1998). This gap is likely to increase because the genome sequencing projects are producing a few hundred thousand protein sequences each year, while only a few thousand of them have their structures determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

Loop modeling can be seen as a mini protein folding problem. The correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Segments of up to 9 residues sometimes have entirely unrelated conformations in different proteins (Sander & Schneider, 1991; Cohen *et al.*, 1993; Mezei, 1998). Thus, the conformation of a given segment is also influenced by the core stem regions that span the loop and by the structure of the rest of a protein that cradles the loop.

Many loop modeling procedures have been described. Similarly to the prediction of whole protein structures, there are both the *ab initio* methods (Moult & James, 1986; Fine *et al.*, 1986; Bruccoleri & Karplus, 1987) and the database search techniques (Greer, 1980; Jones & Thirup, 1986; Chothia & Lesk, 1987). There are also procedures that combine the two basic approaches (Chothia *et al.*, 1986; Martin *et al.*, 1989; Mas *et al.*, 1992; van Vlijmen & Karplus, 1997).

The *ab initio* loop prediction is based on a conformational search or enumeration of conformations in a given environment, guided by a scoring or energy function. There are many such methods, exploiting different protein representations, energy function terms, and optimization or enumeration algorithms. The search algorithms include sampling of mainchain dihedral angles biased by their distributions in known protein structures (Moult & James, 1986), minimum perturbation random tweak method (Fine *et al.*, 1986; Shenkin *et al.*, 1987; Smith & Honig, 1994), systematic conformational search (Bruccoleri & Karplus, 1987; Bruccoleri *et al.*, 1988; Bruccoleri, 1993; Brower *et al.*, 1993), global energy minimization by mapping a trajectory of local minima (Dudek & Scheraga, 1990; Dudek *et al.*, 1998), importance-sampling by local minimization of randomly generated conformations (Lambert & Scheraga, 1989*a*; Lambert & Scheraga, 1989*b*; Lambert & Scheraga, 1989*c*), local energy minimization (Mattos *et al.*, 1994), molecular dynamics simulations (Bruccoleri & Karplus, 1990; Tanner *et al.*, 1992; Rao & Teeter, 1993; Nakajima *et al.*, 2000), genetic algorithms (Ring & Cohen, 1994; McGarrah & Judson, 1993), biased probability Monte Carlo search (Abagyan & Totrov, 1994; Evans *et al.*, 1995; Thanki *et al.*, 1997), Monte Carlo with simulated annealing (Higo *et al.*, 1992; Collura *et al.*, 1993; Carlacci & Englander, 1993; Vasmatzis *et al.*, 1994; Carlacci & Englander, 1996), Monte Carlo and molecular dynamics (Rapp & Friesner, 1999), extended-scaled-collective-variable Monte Carlo (Kidera, 1995), scaling relaxation and multiple copy sampling (Zheng *et al.*, 1993*a*; Zheng *et al.*, 1993*b*; Rosenfeld *et al.*, 1993; Zheng & Kyle, 1994; Zheng *et al.*, 1994; Rosenbach & Rosenfeld, 1995; Zheng & Kyle, 1996), searching through discrete conformations by dynamic programming (Vajda & DeLisi, 1990; Finkelstein & Reva, 1992), random sampling of conformations relying on dimers from known protein structures (Sudarsanam *et al.*, 1995), self-consistent field optimization (Koehl & Delarue, 1995), and an enumeration based on the graph theory (Samudrala & Moult, 1998). A variety of representations were used, such as unified atoms, all non-hydrogen atoms, non-hydrogen and "polar" hydrogen atoms, all atoms, as well as implicit and explicit solvent models. The optimized degrees of freedom include Cartesian coordinates and internal coordinates, such as dihedral angles, optimized in continuous or discrete spaces.

The second, database approach to loop prediction consists of finding a segment of mainchain that fits the two stem regions of a loop (Greer, 1980; Jones & Thirup, 1986; Cohen *et al.*, 1986; Chothia & Lesk, 1987; Tramontano *et al.*, 1989; Chothia *et al.*, 1989; Summers & Karplus, 1990; Levitt, 1992; Tramontano & Lesk, 1992; Topham *et al.*, 1993; Koehl & Delarue, 1995; Reczko *et al.*, 1995; Lessel & Schomburg, 1994; Fechteler *et al.*, 1995; Mandal *et al.*, 1996; Wintjens *et al.*, 1996; Kwasigroch *et al.*, 1996; Martin & Thornton, 1996; Pellequer & Chen, 1997; Oliva *et al.*, 1997; Donate *et al.*, 1996; Rufino *et al.*, 1997; Debnath, 1997; Wojcik *et al.*, 1999; Shepherd *et al.*, 1999; Deane & Blundell, 2000). The stems are defined as the mainchain atoms that precede and follow the loop, but are not part of it. They span the loop and are part of the core of the fold.

The search is performed through a database of many known protein structures, not only homologs of the modeled protein. Usually, many different alternative segments that fit the stem residues are obtained, and possibly sorted according to geometric criteria or sequence similarity between the template and target loop sequences. The selected segments are then superposed and annealed on the stem regions. These initial crude models are often refined by optimization of some energy function.

The database search approach to loop modeling is accurate and efficient when a specific set of loops is created to address the modeling of that class of loops, such as β-hairpins (Sibanda *et al.*, 1989) and the hypervariable regions in immunoglobulins (Chothia & Lesk, 1987; Chothia *et al.*, 1989). For immunoglobulins, an analysis of the hypervariable regions in known immunoglobulin structures resulted in rules with high prediction accuracy for other members of the family. These rules are possible because of the relatively small number of conformations for each loop and because of the dependence of loop conformation on loop length and certain key residues. The accuracy of the approach was demonstrated by a blind, validated prediction of most BR96 antibody residues involved in antigen binding (Bajorath & Sheriff, 1996). There are attempts to classify loop conformations into more general categories, thus extending the impressive performance of the key residues approach to more cases (Ring *et al.*, 1992; Rufino *et al.*, 1997; Oliva *et al.*, 1997; Wojcik *et al.*, 1999).

The database methods are limited by the exponential increase in the number of geometrically possible conformations as a function of loop length. Consequently, only segments of 7 residues or less had most of their conceivable conformations present in the database of known protein structures (Fidelis *et al.*, 1994). In contrast, 8 and 9-residue segments occurred more than once in less than 70% and 40% of the cases, respectively. These estimates depend strongly on the criteria for selecting matching conformations. When slightly stricter criteria are used, only segments of up to 4 residues have most of their conformations defined in a database of known protein structures (Lessel & Schomburg, 1994). This limits the applicability of the database search methods. The limitation is made worse by the requirement for an overlap between at least one residue in each stem and the database segment used for loop modeling. Thus, the completeness of the database for 7-residue segments allows the modeling of only up to 5-residue insertions (Claessens *et al.*, 1989). While only few insertions in a family of homologous proteins are longer than 9 residues, there are many insertions that are longer than 5 residues (Pascarella & Argos, 1992; Benner *et al.*, 1993; Flores *et al.*, 1993).

The problem of database completeness has recently been ameliorated by restrained energy minimization of the candidate loops obtained from a database search (van Vlijmen & Karplus, 1997). Both the internal conformation and global orientation relative to the rest of the protein were optimized. It was concluded that the candidate segments from a database were suitable starting points for modeling loops up to 9 residues long, but extensive optimization was required for loops longer than 4 residues.

In this paper, we take the optimization-based approach to loop modeling. The main reasons are the generality and conceptual simplicity of energy minimization, as well as the limitations on the database approach imposed by a relatively small number of known protein structures. Loop prediction by optimization is in principle applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches. Also, the optimization approach in principle allows for an improvement based on the physics of protein structure, rather than on the growth of the database. Moreover, even the database approach requires both a scoring function to sift through the many alternative loop conformations fitting the stems (Tramontano & Lesk, 1992) and an optimization procedure for relaxing the annealed database segments. Thus, loop prediction may as well rely solely on optimization of an energy function, without any dependence on loop segments from a database.

We describe and extensively evaluate a loop modeling protocol that optimizes the positions of all non-hydrogen atoms of a loop in a fixed environment. The optimization relies on conjugate gradients and molecular dynamics with simulated annealing. The optimized pseudo energy function is a sum of many terms, including some terms from the CHARMM-22 molecular mechanics force field (MacKerell *et al.*, 1998) and spatial restraints based on distributions of distances (Sippl, 1990) and dihedral angles (Cheng *et al.*, 1996) in known protein structures. The paper is organized as follows. In Theory and Algorithms, the technique is described in detail, as are the loops and criteria selected for testing the method. The Results and Discussion begin by evaluating the accuracy of loop predictions as a function of thoroughness of conformational sampling, loop length, environment distortion, and structural properties of native loops. Evaluation is based on the modeling of 40 loops of known structure at each length from 1 to 14 residues. A way to predict the accuracy of the best loop prediction is also described. This is followed by an evaluation of the modeling method by (i) comparing its predictions with those by a recently published, extensively evaluated and powerful loop modeling method (van Vlijmen & Karplus, 1997), and (ii) by summarizing its performance in recent blind predictions of protein structure at CASP3. Next, it is showed that the method is limited mostly by the accuracy of the energy function rather than the thoroughness of the optimizer. Finally, the implications for future work are discussed (Conclusion).

Theory and Algorithms

Modeling of a loop

The method for modeling a loop in a given environment is described here by specifying its three main components: (i) the representation of a protein, (ii) the restraints that define the objective or "energy" function, and (iii) the method for optimizing the energy function. The modeling method is entirely automated and is implemented in the program Modeller-5 (URL `http://guitar.rockefeller.edu`). While the most frequent application of the method is to predict single loops, it is also technically suitable for modeling any set of contiguous or non-contiguous residues or atoms (*e.g.*, several loops, a loop with a ligand, a cluster of sidechains) in the fixed environment created by the rest of the protein.

Representation of a protein

The protein is represented by all non-hydrogen atoms. An explicit treatment of all hydrogen atoms as encoded in Charmm (MacKerell *et al.*, 1998) was also tested, but did not improve prediction accuracy in our hands. No explicit solvent molecules or ligands are included in general, although they could be added in special cases. The degrees of freedom in model optimization are the Cartesian coordinates of the loop atoms. The loop atoms "feel" the other atoms in the protein, but the atoms in the "environment" of the loop do not move during optimization.

Energy function

The main aim was to maximize the accuracy of loop prediction, not to describe the physics of loop structures in proteins. Thus, the emphasis during development of the "energy" or scoring function was on statistical preferences of atoms for different geometries as obtained from the database of known protein structures, rather than on a reductionist model of physical interactions. The stereochemical features (*i.e.*, chemical bonds, bond angles, *etc*) are captured through the use of the CHARMM molecular mechanics force field (MacKerell *et al.*, 1998). The non-bonded interactions and solvation are approximated by a statistical potential of mean force for pairs of protein atoms (Sippl, 1990). In addition, the accuracy of the scoring function is improved by using statistical preferences for the mainchain and sidechain dihedral angles (Šali & Blundell, 1993). The formalism for combining and using these diverse types of information from both physics and statistics is provided by protein structure modeling by satisfaction of spatial restraints, where each individual energy term or a statistical preference is represented by a conditional probability density function for a restrained spatial feature, such as a distance between two atoms (Šali & Blundell, 1993).

The energy function for loop modeling is a sum of simple restraints or pseudo-energy terms, each one of which depends on a distance, angle, dihedral angle, improper dihedral angle, or a pair of dihedral angles defined by two, three, four, or eight atoms. Many combinations of different restraint types were evaluated for their performance in loop modeling. The best energy function so far is described in detail below. The energy function is

$$\begin{aligned} F \;=\; & \sum_{\text{bonds}} k_b(b-\bar{b})^2 + \sum_{\text{angles}} k_\alpha(\alpha-\bar{\alpha})^2 + \sum_{\text{dihedrals}} |k_\phi| - b_\phi \cos(n\phi+\delta) + \\ & \sum_{\text{impropers}} k_i(\theta-\bar{\theta})^2 - \sum_{\substack{\text{sidechain}\\ \text{torsions}}} \ln p_s(\chi/R) - \sum_{\text{residues}} \ln p_\Omega(\Omega/R) - \\ & \sum_{\text{residues}} \ln p_m(\Phi,\Psi/R) + \sum_{\substack{\text{non-bonded}\\ \text{atom pairs}}} \xi\left[E(a,a',d,\Delta_i) + S(r,r',d)\right] \end{aligned} \tag{1}$$

where b is a bond length, α is a covalent bond angle, ϕ is a dihedral angle other than the mainchain Φ, Ψ, Ω and sidechain χ dihedral angles, θ is an improper dihedral angle, and R is the residue type. For the non-bonded atom pairs, ξ is a scaling factor (usually 1), a and a' are the types of the atoms in the pair, d is the distance between them, Δ_i is the difference between the corresponding residue indices, and r and r' are the atomic van der Waals radii (Šali & Blundell, 1993). The sums run over all bonds, angles, dihedral angles, improper dihedral angles, and non-bonded distances that involve at least one of the loop atoms. The non-bonded atom pairs at a distance larger than 4Å are ignored. The force constants k_x and mean values $\bar{x}$ for bonds b, angles α, dihedral angles ϕ, and improper dihedral angles θ were obtained from the May 1993 version of the CHARMM-22 force field (Brooks *et al.*, 1983; MacKerell *et al.*, 1998), as were the phase shift δ and the periodicity parameter n for dihedral angles ϕ. The improper dihedral angles are used to restrain the planarity of peptide bonds and sidechain rings, as well as the chirality of chiral and pro-chiral centers (*e.g.*, C_α atoms of all residues but Gly, C_β atoms of Val and Thr).

The probability distributions for all sidechain dihedral angles (up to four per residue), $p_s(\chi/R)$, depend on the residue type (Ponder & Richards, 1987) and were obtained from a non-redundant set of known protein structures (Šali & Blundell, 1993). They are modeled by a weighted sum of Gaussian functions as described previously (Eq. 26 and Table 5 in Šali & Blundell, 1993):

$$p_s(\chi/R) = \sum_i \omega_i \frac{1}{\sigma_b\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{\Delta(\chi, \bar{\chi}_i)}{\sigma_i})^2] \tag{2}$$

where ω_i is the probability that the restrained sidechain dihedral angle is in class i ($\sum_i \omega_i = 1$) and $\Delta(x, y)$ is the shortest path around the 360° circle from angle x to angle y. Most sidechain dihedral angles can be in up to three classes, depending on the residue type. Some residues, however, have a smaller number of possible classes; for example, χ_2 in His has only two classes.

Similarly, the restraints on the mainchain dihedral angle Ω, p_Ω, were represented by a single Gaussian function centered on 180° with the standard deviation of 5°. The *cis*-proline states have not been modeled. In the set of 40 test loops of 8 residues, there are 2 and 13 *cis*- and *trans*-proline residues, respectively.

For each loop residue, the restraint on the mainchain dihedral angles Φ and Ψ is

$$p_m(\Phi, \Psi/R) = \sum_{i=1}^{m} \omega_i \, p_i \; = \; \sum_{i=1}^{m} \omega_i \frac{1}{2\pi\sigma_{\Phi,i}\sigma_{\Phi,i}\sqrt{(1-\rho_i^2)}} \; \cdot \tag{3}$$

$$\exp\left\{-\frac{1}{(1-\rho_i^2)}\left[\frac{1-\cos(\Phi-\bar{\Phi}_i)}{\sigma_{\Phi,i}^2} - \rho_i \frac{\sin(\Phi-\bar{\Phi}_i)}{\sigma_{\Phi,i}}\frac{\sin(\Psi-\bar{\Psi}_i)}{\sigma_{\Psi,i}} + \frac{1-\cos(\Psi-\bar{\Psi}_i)}{\sigma_{\Psi,i}^2}\right]\right\}$$

where m is the number of mainchain conformation classes in the Ramachandran plot (Figure 1a), ω is the weight of the corresponding conformation class (Figure 1b), the bar indicates the average

Φ and Ψ values, σ is the standard deviation, and ρ is the correlation coefficient between Φ and Ψ ($-1 \leq \rho \leq 1$). The parameters ω, $\bar{\Phi}$, $\bar{\Psi}$, σ, and ρ were obtained from a representative set of 1,000 protein structures that shared less than 60% sequence identity to each other and were determined by X-ray crystallography at resolution of 2.3Å or better. The total number of residues was 217,807. The Ramachandran plot spanned by the Φ and Ψ dihedral angles (Ramachandran *et al.*, 1963) was divided into $5^\circ \times 5^\circ$ bins. The frequency of residues in each bin was obtained separately for each of the 20 standard residue types. The peaks and valleys in the Ramachandran plots guided the partitioning of each plot into a few conformation classes (2–5 classes), again separately for each residue type; the residue types differ mostly in the boundary between the two "β" classes with the negative Φ values and positive Ψ values. The weights ω correspond to the relative frequencies of each residue in each of its classes. Next, the analytic model in Eq. 3 was fitted to the Ramachandran plots by a least-squares method (Press *et al.*, 1992), resulting into the optimal values for $\bar{\Phi}$, $\bar{\Psi}$, σ, and ρ. A comparison between the Ramachandran plot and a fitted model for four representative residue types is shown in Figure 2. The values of all parameters can be obtained from the MODELLER library files `mnch.lib`, `mnch1.lib`, and `af_mnchdef.lib`.

The first part of the non-bonded term, E, is taken from Melo & Feytmans, 1997, where it was derived as described by Sippl, 1990. E is an atomistic, distance-dependent statistical potential of mean force. Atoms in amino acid residues are classified into one of 40 atom type groups. The potential was obtained from those atom pairs in known protein structures that were separated by 11 or more residues in sequence. The original potential in the histogram form (Melo & Feytmans, 1997) was converted into cubic splines (Press *et al.*, 1992) to allow the use of first derivatives in optimization (Šali *et al.*, 1999). The second part of the non-bonded term, S, is a harmonic lower bound on non-bonded atom–atom distances (Eq. 22 in Šali & Blundell, 1993):

$$S(r, r', d) = \begin{cases} k_n \left[d - \gamma(r + r')\right]^2 \ ; & d \leq \gamma(r + r') \\ 0 \ ; & d > \gamma(r + r') \end{cases} \tag{4}$$

where d is the distance between the two atoms, k_n is the force constant (usually 59 kcal/mol), and γ is a constant (usually 0.83). S is used to compensate for noise in E at short distances of less than 3Å.

Optimization of the energy function

Optimization begins by generating an initial loop conformation (Figure 3). The atoms of the loop are positioned with uniform spacing on the line that connects the mainchain carbonyl oxygen and amide nitrogen atoms of the N- and C-terminal anchor regions, respectively. Next, the atomic positions are randomized by adding a random number distributed uniformly from -5 to 5Å to each of the Cartesian coordinates. One loop prediction consists of optimizing independently a number of such randomized initial structures, and picking as the final model the conformation that has the lowest value of the energy function. A good compromise between efficiency and performance is achieved by 50–500 independent optimizations (Results and Discussion).

It is tempting to generate initial conformations for loop atoms that are more protein-like and not random. However, the accuracy of the loop models obtained from random initial conformations is limited mostly by the accuracy of the scoring function, not the power of the optimizer (Figure 12). Thus, starting a loop prediction with random positions for loop atoms does not decrease the final accuracy of the loop models. Nevertheless, more realistic starting conformations may increase the efficiency of the loop modeling method by decreasing the need for exhaustive optimization of each starting conformation.

The procedure for optimizing a single initial conformation begins with a conjugate gradients minimization, continues with molecular dynamics with simulated annealing, and finishes by conjugate gradients again (Figure 3). The details about the optimization schedule can be found in the MODELLER file `__loop.top`. Briefly, the first conjugate gradients phase is designed to relax the system and consists of five successive minimizations of up to 200 steps each, gradually increasing the scaling factors ξ for the non-bonded restraints from 0, 0.01, 0.1, 0.5, to 1.0, respectively. In this phase, the atoms are allowed to pass very near each other without having to surmount large energy barriers. This stage is followed by a relatively fast heating up of the system consisting of 200 4 fs steps of "molecular dynamics" at 150, 250, 400, 700, and 1000 K. The heating stage is followed by the main optimization stage that consists of gradual cooling by molecular dynamics of 600 4 fs steps at 1000, 800, 600, 500, 400, and 300 K. Finally, the optimization is completed by a conjugate gradients relaxation consisting of up to 1000 steps. There are in fact two cycles of the conjugate gradients, molecular dynamics with simulated annealing, and conjugate gradients phases: In the first cycle, only those non-bonded atom pairs are considered that contain loop atoms alone (*i.e.*, the loop does not "feel" its environment). In the second cycle, the atom pairs that contain up to one environment atom are also included in the energy function (*i.e.*, the loop does "feel" its environment). It was found empirically that neglecting the environment in the first cycle results in lower final energy values than including the environment from the beginning of the optimization.

500 independent optimizations of an 8-residue loop takes from 8 to 30 hours of CPU time on an R10000-190 SGI workstation. Predictions for many loops under many different conditions were performed. This was made possible by running the computations in parallel on a cluster of SGI and PC Linux computers. An efficient and robust use of the cluster of processors was made possible by the program CLUSTOR (TurboLinux, San Francisco; `http://www.turbolinux.com`). With 30 processors, a single 8-residue loop prediction involving 500 independent optimizations takes less than one hour.

Test loop sets

Protein structures that share less than 60% sequence identity to each other and were determined by X-ray crystallography at resolution of 2.0Å or better were extracted from PDB. Helices and strands were assigned by the DSSP program (Kabsch & Sander, 1983). The regions outside helices and strands were defined as loops (Figure 4). These loops were filtered to obtain 14 test sets of loops. Each test set contained 40 loops of the same length, spanning the range from 1 to 14 residues. The test sets were obtained by random selection applying the following criteria: (i) no two loops in the test set for a given length are from the same structure; (ii) there are no overlaps between any two loops from any two sets; (iii) the N- and C-termini are not used as test loops; and (iv) also excluded are loops that span more than 9Å between their terminal C_α atoms. The last criterion was applied to maximize the number of geometrically feasible conformations for each test loop (Chan & Dill, 1989), thus maximizing the difficulty of loop modeling. The distribution of the end-to-end C_α–C_α distance for all 8-residue loops is approximately Gaussian with the mean of 13Å and standard deviation of 4Å. Only 1% of 8-residue loop–loop alignments have more than 2 positions with the same residue type in both loops in our test set.

Criteria for accuracy of loop predictions

There will generally be a wide spread in the accuracy of the predictions for different loops. Thus, it is necessary to evaluate the accuracy of a method by testing it on many different loops. We use test sets of 40 loops for each length from 1 to 14 residues, resulting in 560 test loops in total. The average and standard deviation of the accuracy for 40 test loops of the same length are the most frequently used measures of the method accuracy in this paper.

The accuracy of a single loop prediction is evaluated by comparing it with the native conformation. A large variety of reasonable criteria for comparing loop conformations exist. They include RMSD and DRMS measures (Levitt, 1983) for different sets of atoms, such as C_α, mainchain, and all atoms. The RMSD error can be calculated from the superposition of the whole structures excluding the loop ("global" superposition) or from the superposition of the compared loop atoms only ("local" superposition). In addition to Cartesian coordinates, dihedral angles, dihedral angle classes, and mainchain conformation classes can also be compared. It is not practical to use all of these criteria. Fortunately, it is also unnecessary because there is a statistical correlation between reasonable loop comparison measures (see below). Thus, a statistical description of the accuracy by one measure and the statistical relationship between that measure and all other interesting measures provides a description of the accuracy for all the measures.

We calculate RMSD for Cartesian coordinates only, and annotate the symbol by subscripts and arguments for exact definition, $\mathsf{RMSD}_{\mathsf{part,atom\_types}}(\mathsf{part}, \mathsf{atom\_types})$. The two subscripts indicate the part of the protein structure and the atom types that were used for least-squares superposition, and the two arguments indicate the part of the structure and the atom types that were compared to calculate the RMSD. For example, the "global" RMSD for the loop mainchain atoms (N, C_α, C, O) after superposition of the mainchain atoms in the stem residues on each side of the loop (usually three residues) is indicated by $\mathsf{RMSD}_{\mathsf{global}} = \mathsf{RMSD}_{\mathsf{stem,NC_\alpha CO}}(\mathsf{loop}, \mathsf{NC_\alpha CO})$. The "local" RMSD for the mainchain loop atoms superposed on themselves is indicated by $\mathsf{RMSD}_{\mathsf{local}} = \mathsf{RMSD}_{\mathsf{loop,NC_\alpha CO}}(\mathsf{loop}, \mathsf{NC_\alpha CO})$. Distortion of the two stems is indicated by the global and local superposition of the mainchain atoms in the stems, $\mathsf{RMSD}_{\mathsf{core,NC_\alpha CO}}(\mathsf{stem}, \mathsf{NC_\alpha CO})$ and $\mathsf{RMSD}_{\mathsf{stem,NC_\alpha CO}}(\mathsf{stem}, \mathsf{NC_\alpha CO})$, respectively. "Core" refers to the whole protein excluding the loop, the stems and the distorted environment. $\mathsf{RMSD}_{\mathsf{global}}$ for a loop in the native structure is obviously the same for the superposition of either the stem atoms, core atoms, or all non-loop atoms.

However, in the case of evaluating loop models in the context of approximate structures, the choice of the "core" influences the $\mathsf{RMSD}_{\mathsf{global}}$ value. In this paper, the superposition of the loop stem residues is used to obtain $\mathsf{RMSD}_{\mathsf{global}}$.

The $\mathsf{RMSD}_{\mathsf{local}}$ was picked as the primary measure of loop accuracy. Its correlation with a number of other measures is strong, corresponding to Pearson correlation coefficients larger than 0.8 (Figure 5). The $\mathsf{RMSD}_{\mathsf{local}}$ for local superposition of all atoms in an 8-residue loop is approximately twice that for the superposition of the mainchain atoms only. The $\mathsf{RMSD}_{\mathsf{global}}$ for the mainchain atoms is about 1.5 times the $\mathsf{RMSD}_{\mathsf{local}}$. The DRMS for the mainchain atoms is approximately 0.8 times the $\mathsf{RMSD}_{\mathsf{local}}$. The $\mathsf{RMSD}_{\mathsf{local}}$ for C_α atoms is almost the same as that for all the mainchain atoms. The $\mathsf{RMSD}_{\mathsf{local}}$ for N, C_α, and C is always slightly smaller than that for all the mainchain atoms. These correlations were obtained from a comparison of the loop modeling predictions with the corresponding native loop structures. Thus, the correlations can be used to asses the present method (for 8-residue long loops) by all the criteria plotted in Figure 5, even when only the $\mathsf{RMSD}_{\mathsf{local}}$ numbers are available.

Loop modeling errors can be approximately deconvoluted into two contributions: (i) errors in conformation and (ii) errors in orientation of the loop relative to the rest of the protein (van Vlijmen & Karplus, 1997; Martin *et al.*, 1997). While $\mathsf{RMSD}_{\mathsf{local}}$ is a good measure of the accuracy of conformation, it does not depend on the relative orientation of the loop. On the other hand, $\mathsf{RMSD}_{\mathsf{global}}$ depends on both the conformational and orientational accuracies. However, the choice of the superposed atoms before the $\mathsf{RMSD}_{\mathsf{global}}$ calculation is generally arbitrary. For example, when a small domain containing the loop is shifted relative to the rest of the protein, the $\mathsf{RMSD}_{\mathsf{global}}$ of the loop will indicate a large error even when the loop is modeled perfectly in the context of the small domain. For this reason, $\mathsf{RMSD}_{\mathsf{local}}$ was chosen as the primary evaluation measure, although the main evaluation results are also reported using $\mathsf{RMSD}_{\mathsf{global}}$ (Figures 8–11, Table 2). In practical terms, for the present method applied to 8-residues long loops in their native environments, the $\mathsf{RMSD}_{\mathsf{global}}$ is approximately 1.5 times the $\mathsf{RMSD}_{\mathsf{local}}$.

Another useful characterization of accuracy of a method is the fraction of loop predictions that fall in the good, medium and bad accuracy class (Figure 6). A good prediction has $\mathsf{RMSD}_{\mathsf{local}}$ smaller than 1Å, a bad prediction has $\mathsf{RMSD}_{\mathsf{local}}$ larger than 2Å, and a medium prediction falls between these two extremes. Examples of an 8-residue loop model in each of the three classes are shown in Figure 6. In the good class, the mainchain carbonyl oxygen atoms almost invariably point in the correct direction and most of the residue mainchain conformation classes are predicted correctly. The error approaches that in the medium resolution X-ray analysis. In the medium class, there are occasional flips of the mainchain oxygen atoms as well as errors in the residue mainchain conformation classes. However, the error is not larger than the dynamic fluctuations of most loops at room temperature (Fushman *et al.*, 1997). Both good and medium loop predictions are informative when using comparative protein structure models.

Data deposition

The modeling program MODELLER as well as the list of the loops in each of the test sets are available at the URL `http://guitar.rockefeller.edu/`. The test sets contain 40 loops each. There are 14 test sets corresponding to the loop lengths from 1 to 14 residues.

Results and Discussion

In Theory and Algorithms, we describe a technique for modeling of loops in protein structures. The method predicts the positions of all non-hydrogen atoms of a given loop in a fixed environment by optimizing a scoring or "energy" function. Many different energy functions and optimization schedules were explored. The current energy function contains terms from a molecular mechanics force field as well as restraints based on statistical distributions derived from known protein structures (Sippl, 1990; Cheng *et al.*, 1996). Bonds, angles, some dihedral angles, and improper dihedral angles are restrained by the corresponding terms in the CHARMM-22 potential function (MacKerell *et al.*, 1998). The mainchain and sidechain dihedral angles as well as non-bonded atom pairs are restrained by statistical potentials. The energy function is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing.

This section begins by a description of the average accuracy of the method as a function of degree of conformational sampling, loop length from 1 to 14 residues, and properties of the native loop, such as the mobility and compactness. To simulate real comparative modeling problems, the method was also evaluated by predicting loops of known structure in only approximately correct environments with errors typical of comparative modeling without misalignment. Next, the accuracy of the lowest energy prediction for a given loop is estimated from the structural variability among a number of low energy predictions. The relative value of the present method is gauged by (i) comparing it with one of the most successful previously described methods, and (ii) describing its accuracy in recent blind predictions of protein structures. The latter evaluation also revealed limitations in practical loop modeling, originating from problems other than those of the loop modeling method itself. Finally, to indicate future developments, it is shown that the average accuracy of prediction is limited primarily by the accuracy of the energy function rather than by the extent of conformational sampling.

Accuracy of loop predictions as a function of thoroughness of conformational sampling

As described in Theory and Algorithms, loop modeling consists of independent energy optimizations of many random initial loop conformations. The final loop prediction is the optimized conformation that has the lowest energy among all the independent loop optimizations. The scatter plots of the energy and the $\mathsf{RMSD}_{\mathsf{local}}$ error (Theory and Algorithms) for independent optimizations in a successful and unsuccessful prediction are shown in Figures 7a and b, respectively. The success corresponds to a strong correlation between the energy and RMSD. In such cases, most of the low energy predictions are accurate. On the other hand, the failure corresponds to a lack of a positive correlation between the energy and RMSD. Most of the low energy predictions are different not only from the native loop, but also from each other. Even if a geometrically good conformation is encountered during sampling, its energy is high and is thus not predicted to be the native loop.

An important methodological consideration is the amount of conformational sampling performed for each loop prediction. This is directly proportional to the number of independent loop optimizations. The larger the number of loop optimizations, the better is the average accuracy (Figure 8). For 4-residue loops, improvement beyond 50 independent optimizations is negligible (Figure 8a). On the other hand, for 8-residue loops, the average $\mathsf{RMSD}_{\mathsf{local}}$ error decreases from 1.40 ± 0.12Å to 1.16 ± 0.10Å when the number of independent optimizations is increased from 50 to 500. The average accuracy of loops longer than approximately 6 residues is likely to improve marginally even beyond 500 independent optimizations (Figures 8b and c). To be able to sample the 14 test sets of 40 loops each, it was necessary to limit most subsequent evaluations of the method accuracy to 50 independent optimizations.

Accuracy of loop predictions as a function of loop length

The difficulty of the loop modeling problem increases with loop length. For longer loops, there are more incorrect conformations that increase the demands on the optimizer to generate a good model and on the energy function to identify it. The average $\mathsf{RMSD}_{\mathsf{local}}$ error and its standard deviation rise almost linearly with loop length (Figure 9). For 50 independent optimizations, the average $\mathsf{RMSD}_{\mathsf{local}}$ error is 0.59, 1.40, and 2.96Å for 4, 8, and 12-residue loops, respectively. All predictions of 4-residue loops fall into either the good or medium accuracy class (Table 1, Figure 4). For 8-residue loops, 50, 40, and 10% of models are in the good, medium, and bad class, respectively, when 500 independent optimizations are performed. For 12-residue loops, only 30% of predictions are in the good or medium class when 500 independent optimizations are performed. In conclusion, it makes sense to model even loops 12 residues long, if the environment error is small and 30% chance of obtaining either a good or a medium loop model is acceptable.

During the course of this project, the optimization method and the energy function were improved iteratively by relying on the test set of 40 8-residue loops. Thus, it could be that the final reported results for these loops are misleadingly favorable. To check for such bias, additional 40 8-residue loops were selected as described in Theory and Algorithms and predicted by the final version of the method only. The average accuracy for 50 independent optimizations differed from the first test set only in the second decimal place. Also, the accuracy for 8-residue loops is not an outlier relative to other lengths (Figure 9). Further confidence in the statistical robustness of the presented results is provided by comparing our method with another method (van Vlijmen & Karplus, 1997) and by the accuracy of the loop models submitted to CASP3 (see below). In both cases, the accuracy of the method is consistent with the evaluation based on the test loops.

Accuracy of loop predictions as a function of environment distortion

In real comparative modeling when the native structure of the target sequence is not known, loops are not modeled in the perfectly correct environment. At best, only an approximately correct environment is available. This complication in loop modeling mimics the situation in sidechain modeling, which needs to be performed on an approximate, not exact backbone. In the case of sidechain modeling, the accuracy of the predicted sidechain packing drops rapidly when the core backbone distortion increases beyond 1Å (Chung & Subbiah, 1996). Thus, it was expected that the average accuracy of the loop modeling method would also be worse than indicated by the evaluations in the native environment described in the preceding section. To quantify the impact of the environment errors on the accuracy of loop modeling, we prepared test sets for 4, 8, and 12-residue loops in distorted environments. Each set contained the 40 test loops in 5 distorted environments each. The environment of a loop is defined in this section to include six stem residues on each side of the loop as well as all the atoms that are within 15Å of at least one loop or stem atom in the native structure. The size of the distortion $\mathsf{RMSD}_{\mathsf{core},\mathsf{NC}_\alpha\mathsf{CO}}(\mathsf{stem},\mathsf{NC}_\alpha\mathsf{CO})$ was up to almost 3Å (Fig 10). Environment errors of less than 3Å are typical of comparative models of exposed regions based on alignments without errors. The test sets were obtained by subjecting the environment atoms in the native structure to a molecular dynamics simulation at 600 K guided by the energy function in Eq. 1 and restrained by the rest of the native protein structure.

For 8-residue loops in an environment with the stem mainchain RMSD distortion of 2.5Å, the average error of loop prediction increases for 25% from 1.40Å to 1.75Å, as measured by the least-squares line through the data points (Figure 10). Nevertheless, 75% of the predictions remain in the good and medium classes, even when only 50 independent optimizations are performed (Table 1, Figure 4). The impact of environment errors is smaller for long loops than it is for short loops, presumably because the prediction accuracy for long loops is already low in the native environment. For example, for 12-residue loops, almost no decrease in prediction accuracy occurs with environment distortion. On the other hand, for 4-residue loops, the increase in the error is large, from 0.43Å to 1.21Å. However, all the 4-residue loop predictions are still in the good and medium accuracy classes, presumably due to a relatively small number of different conformations for such short loops.

The negative impact of the environment errors on the loop accuracy might be decreased by optimizing the positions of the environment atoms as well as the loop atoms. However, our preliminary attempts to do so have not improved the average accuracy of the loop models (data not shown). It appears that the optimization algorithm and/or the energy function are overwhelmed by the large number of degrees of freedom corresponding to the loop combined with its environment. For example, even when a short distance of 6Å is used to define the loop environment, there are from 17 to 36 environment residues for the 40 8-residue loops. This might correspond in difficulty at least to the standard 25 residue "loop" modeling problem (8 + 17), far beyond the range of reliable performance for any modeling method (Table 1, Figure 4).

Accuracy of loop predictions as a function of loop properties

Loops usually occur on the surface of a protein globule and have relatively few contacts with the rest of the fold. Consequently, some of the loops are structurally least well defined parts of the protein mainchain. Conformational heterogeneity of loops includes multiple local minima with rare transitions between the minima (static disorder) as well as large fluctuations around a single minimum (dynamic disorder). Both of these phenomena are reflected in relatively high crystallographic isotropic temperature factors, B_{iso}. The atomic B_{iso} are generally determined accurately in protein structures refined at resolution of 2Å or better, such as the structures from which the test loops were collected. It was expected that a loop with a high average B_{iso} will tend to be predicted less accurately than a loop with a low average B_{iso}. However, no such correlation is observed in a comparison of the normalized average B_{iso} values for different loops as a function of the prediction accuracy (Figure 11a). One possible explanation is that most of the loops with high average temperature factors fluctuate around their equilibrium conformations, not between several different local minima. Thus, the average structure determined by crystallography and the equilibrium structure approximated by the prediction will tend to correspond to the same conformation, eliminating the correlation between prediction accuracy and high mobility.

There is also no correlation between the prediction error and the number of atomic contacts within the loop (Figure 11b). The compact loops are predicted equally well as the non-compact loops. This might indicate problems with the non-bonded terms in the energy function.

It was expected that the prediction accuracy will increase with the number of atomic contacts between the loop atoms and the environment in the native conformation, due to the moulding of the loop by the fixed environment. However, there is almost no such correlation (Figure 11c). This indicates problems with the non-bonded terms in the energy function, resulting in relatively little useful information provided by the environment. Another explanation, not mutually exclusive with the first one, is that the internal mainchain preferences of the loop residues tend to be consistent with the conformation preferred by the environment (Gõ & Abe, 1984).

The number of contacts between the loop atoms and the neighboring molecules in the crystal also does not have a significant impact on the prediction accuracy (Figure 11c). This was a surprise: Loops with a higher number of inter-molecular contacts were expected to be predicted less accurately because the inter-molecular contacts are completely ignored in loop modeling. It appears that the native loop conformations tend to be consistent with the inter-molecular packing (Gõ & Abe, 1984).

Estimating the accuracy of a loop prediction

In the absence of a perfect loop modeling method, it is useful to have an estimate of the error of a given loop prediction. Methods for detecting errors in protein structure models have been reviewed (Sánchez & Šali, 1997*b*). One popular approach is based on energy profiles (Lüthy *et al.*, 1992; Sippl, 1993). In this approach, a region is predicted to be in error when its energy is above a certain cutoff. This rule is not expected to work well in the case of our loop modeling method since the best available energy function is optimized to obtain the model in the first place. There is an additional concern about the theoretical validity of the energy profiles for detecting regional errors in models. It is likely that the contributions of the individual residues to the overall free energy of folding vary widely, even when normalized by the number of atoms or interactions made. If this is correct, the correlation between the prediction errors and energy peaks is greatly weakened, resulting in the loss of predictive power of the energy profile. Despite these concerns, error profiles have been useful in some applications (Sánchez & Šali, 1997*a*; Guenther *et al.*, 1997).

The following rational is used here to estimate an error of a given loop prediction. We take a case of a loop with one dominant native conformation. In such a case, there are degrees of freedom for which the true energy function has a deep, global free energy minimum, in addition to many local minima. The global free energy minimum is the native conformation. The energy function and degrees of freedom that are explored in modeling are only an approximation of the true energy function and its degrees of freedom. However, it is plausible that the more pronounced is the free energy minimum in the modeling function, the less likely it is that the errors in the function moved the minimum away from that in the true energy function. In other words, a given fractional error in the energy surface may not move a deep minimum, while it is likely to move a shallow minimum. The many independently optimized loop conformations in a single loop prediction make it possible to estimate how pronounced the lowest free energy minimum is. If the free energy surface has multiple comparable minima without a dominant minimum, the loop modeling method will result in multiple, significantly different conformations. When there is a dominant free energy minimum, the loop modeling method produces similar low energy conformations if the optimizer works well. Conversely, the more similar are the lowest energy conformations, the more pronounced must the corresponding minimum be, and the less likely it is that the best prediction has a large error. The more different are the lowest energy conformations, the more rugged is the modeling energy surface and the less confidence one has in the lowest energy solution.

This idea was tested by analyzing the "ensembles" of 500 independent optimizations of 40 8-residue loops. The structural variation among the lowest energy solutions was defined as the average $\mathsf{RMSD}_{\mathsf{local}}$ for all pairs of the 20 lowest energy conformations, out of the 500 conformations in total. Similar results are obtained when the best 5 to 50 conformations are used (data not shown). The scatter plot of the $\mathsf{RMSD}_{\mathsf{local}}$ error of the lowest energy prediction and the structural variation of the lowest energy solutions is shown for the 40 test loops in Figure 12. When the structural variation among the lowest energy models is low, the error of the lowest energy prediction is indeed small. When the variation is large, the error can be either small or large. In this case, it is not possible to distinguish between the failure of the optimizer, errors in the energy function, or a truly promiscuous loop. There are no cases of low variation and large error, indicating that the optimizer does not often get trapped into a local minimum. For 8-residue loops, a good upper bound on the error of loop prediction is the variation among the lowest energy models multiplied by two.

Comparison with previous results

The accuracy of any new prediction method has to be compared with previous results. There are a great many existing loop modeling methods (Introduction) and it is not practical to consider all of them. Thus, we chose to compare the present results with only one, but carefully selected previous study. The requirements were that the reference method be recent, well documented, automated, tested under realistic conditions, evaluated with a reasonable number of test loops of varying length, and that it compared favorably with other prior methods. One such method is that of van Vlijmen & Karplus, 1997 (the "VK" method). The VK method first selects approximately 1,000 loop candidates from PDB, based on stem fitting, and then subjects these candidates to independent energy minimizations *in vacuo* to obtain the final prediction. The CHARMM-19 energy function without electrostatics, applied to the mainchain atoms N, C_α, C, and the first sidechain atom C_β, is supplemented by strong mainchain dihedral angle restraints to focus minimization on the conformations relatively close to the template segments. Optimization results in both the global movement of the loop relative to the rest of the protein and local relaxation of the loop conformation. The VK method compared favorably with several other optimization and database approaches (Moult & James, 1986; Fidelis *et al.*, 1994; Zheng & Kyle, 1996).

14 loops from 4 to 9 residues long were predicted by the VK method in a manner that is appropriate for comparison with the present method (Table 2). Because the original evaluation of the VK method was given in terms of $\mathsf{RMSD}_{\mathsf{stem,NC_\alpha C}}(\mathsf{loop}, \mathsf{NC_\alpha C})$, the comparison here also had to rely on this measure, not $\mathsf{RMSD}_{\mathsf{global}}$ or $\mathsf{RMSD}_{\mathsf{local}}$. In any case, $\mathsf{RMSD}_{\mathsf{stem,NC_\alpha C}}(\mathsf{loop}, \mathsf{NC_\alpha C})$ tends to be only slightly smaller than $\mathsf{RMSD}_{\mathsf{global}}$ (Figure 5f, Table 2). For 3 out of 14 loops (2act, 2fb4, 2fbj), the VK method produced a more accurate model than the present method. For two loops (8abp and 3dfr_120-124), both methods produced indistinguishably good models. For the remaining 9 loops, the present method produced more accurate models than the VK method. In the three cases where the VK method produced a better model, a good model was sampled by the present method, but was not recognized as such by the energy function. This comparison may not be entirely fair to the VK method. The reason is that the VK method would probably perform better at this point in time than it did a few years ago because the database of known protein structures, on which the VK method relies, is now significantly larger. On the other hand, as the authors themselves pointed out, the published results of the VK method were perhaps a little better than would be obtained in realistic applications because the candidate loops for energy minimization were 50 loops with minimal $\mathsf{RMSD}_{\mathsf{loop,NC_\alpha C}}(\mathsf{loop}, \mathsf{NC_\alpha C})$ from the target conformation, not 1,000 loops with best fitting stems.

It is informative to examine the rank of the lowest $\mathsf{RMSD}_{\mathsf{local}}$ conformation in the list of 500 independently optimized conformations sorted by energy, as well as the rank of the lowest energy conformation in the list of conformations sorted by $\mathsf{RMSD}_{\mathsf{local}}$ (Table 2). Conformational sampling almost always generates an extremely good solution. In only one of the 14 cases is the lowest $\mathsf{RMSD}_{\mathsf{global}}$ significantly larger than 1Å (2act). The lowest energy prediction is frequently (*i.e.*, in 8 out 14 cases) within 10% of the geometrically best conformations. Even when the lowest energy structure has a low rank in the list of structures sorted by $\mathsf{RMSD}_{\mathsf{global}}$, it can still be a good prediction (*e.g.*, 3dfr_20-23), indicating that most independent optimizations ended in the vicinity of the native structure.

Loop modeling at CASP3

Protein structure modeling methods, including comparative modeling, are tested at the bi-annual CASP meetings (Moult *et al.*, 1999). The last meeting, CASP3, was held in December of 1998 (Jones & Kleywegt, 1999). Protein modelers are challenged to model sequences with unknown 3D structure and to submit their predictions to the organizers before the meeting. In parallel, the 3D structures of the prediction targets are being determined by X-ray crystallography or NMR methods. They only become available after the models are calculated and submitted. Thus, a *bona fide* evaluation of protein structure modeling methods is possible.

One major limitation of comparative modeling is that it is frequently not possible to lower the mainchain RMSD error of a model below the RMSD between the template and target structures (Sánchez & Šali, 1997*a*). There are several ways to overcome this limitation in some cases. The first way is to use more than one template structure (Sánchez & Šali, 1997*a*). The second way of "moving" the model closer from the template to the actual target structure is to calculate good models for insertions (*i.e.*, loops). The third, ambitious way is to rely on a loop modeling method to predict relatively small shifts and distortions in a segment that is aligned to the template structure. Even in correct alignments, the aligned regions can have an $\mathsf{RMSD}_{\mathsf{global}}$ of approximately 3Å. Thus, a good loop modeling method applied to such segments is expected to improve the model relative to the template. For our loop modeling method, for example, 90% of loops of 8 residues are modeled with less than 2Å $\mathsf{RMSD}_{\mathsf{local}}$ error when the environment is correct. The application of a loop modeling method to a fully aligned segment is referred to as "distortion" modeling, rather than loop modeling.

A precursor of the present loop prediction method was used to model both insertions and fully aligned segments in our submissions to CASP3. A detailed description of our experience will be published elsewhere (A. Fiser, R. Sánchez, F. Melo, and A. Šali, in preparation). Here, we only summarize the main lessons as they pertain to the modeling of loops and distortions.

The method applied to insertions performed according to the evaluation in this paper, when errors in the loop environment were relatively small (*i.e.*, not caused by the alignment errors). A successful prediction of an 8-residue loop is shown in Figure 13a. However, for most of the CASP3 targets, the alignment was the main challenge, frequently resulting in misaligned stem residues or pieces of protein structure occupying loop's space. Even a perfect loop modeling method will of course fail in such circumstances. The implication for comparative modeling users is that loop modeling methods are currently most valuable for modeling insertions in the easy and medium difficulty cases, corresponding to target–template sequence identity of at least 30%.

The method applied to distortion modeling also performed as evaluated in this paper. A "successful" prediction of a 6-residue segment is shown in Figure 13b. It was modeled with the $\mathsf{RMSD}_{\mathsf{local}}$ error of 1.09Å. However, although the prediction was accurate by the loop modeling standards, it was a failure by the CASP3 standards. The reason was that the template segment was in fact not divergent structurally, but closer to the actual target structure than our prediction, despite a large difference between their sequences (Figure 13b). Thus, the risk taken to move the model from the template structure closer to the target structure backfired, even though a good model of the segment was produced. Clearly, distortion modeling is useful only if aligned segments that diverged beyond the error of the loop modeling method can be identified. This is currently an unsolved problem (Abagyan *et al.*, 1997). It may be as difficult to predict that a region is in error as it is to calculate its correct model. The implication for comparative modeling users is that distortion modeling is probably not worth the risk at this time.

Is accuracy limited by the optimizer or the energy function?

Any protein structure prediction consists of two main parts. The first part is generation of candidate structures, either by enumeration or conformational sampling. The second part is selection of the best candidate structure based on some scoring function. In protein structure prediction based on "energy" minimization, which includes the present loop modeling method, structure generation and selection are intertwined because the energy function partly determines what conformations are sampled by the optimization algorithm. In principle, such optimization methods can fail for one or both of the following reasons: (i) the optimizer samples the conformational space inadequately and never generates a correct conformation; or (ii) the energy function does not identify a correct conformation even if it is generated by the optimizer. In the case of our loop modeling method, the optimizer was expected to contribute significantly to the modeling errors; it is the stochastic behavior of the optimizer that results in the need for many independent optimizations of the same loop, each one starting with a different initial conformation. We show here, however, that the present loop modeling method is limited primarily by the accuracy of the energy function rather than the robustness of the optimizer.

The first indication of an inadequate energy function is provided by examples of a successful and unsuccessful loop prediction (Figure 7). For the 500 independent optimizations used to make the correct prediction in Figure 7a, the correlation between the energy and the $\mathsf{RMSD}_{\mathsf{local}}$ error is strong. Any of the approximately 30 lowest energy conformations are close to the native loop. In contrast, there is no correlation between energy and error in the case of an incorrect prediction (Figure 7b). Many conformations different from each other and from the native loop have an equally low energy score. Even though a conformation that is close to the native structure ($\mathsf{RMSD}_{\mathsf{local}}$ = 0.84Å) was sampled, the lowest energy model had one of the largest errors among the sampled models (3.17Å).

The second indication of a reasonable performance of the optimizer is the data on its convergence properties (Figure 8). Loop models were predicted from 1 to 500 independent optimizations for sets of 40 4, 8, and 12-residue loops each. There is essentially no improvement in the average accuracy of loop prediction when more than 50 independent optimizations are performed for 4-residue loops. For 8 and 12-residue loops, there is a small improvement in the average accuracy from 100 to 500 optimizations. This indicates that conformational sampling is essentially complete. Thus, it is likely that the failure of loop prediction is due to the energy function, not the optimizer.

The most direct evidence that loop prediction is limited primarily by the energy function is provided by the data on how close to the native structure are the closest encountered conformations, irrespective of their energy (Figure 8). These data were obtained for the three different loop lengths by calculating the average over 40 loops of the minimal $\mathsf{RMSD}_{\mathsf{local}}$ error among the independently optimized conformations of the same loop. This average minimal $\mathsf{RMSD}_{\mathsf{local}}$ error is of course always lower than the average $\mathsf{RMSD}_{\mathsf{local}}$ error for the lowest energy predictions. Moreover, the average minimal $\mathsf{RMSD}_{\mathsf{local}}$ error is very low indeed. For example, it is less than 1.0 and 2.0Å for 8 and 12-residue loops, respectively, even when only 50 independent optimizations are performed. The corresponding average $\mathsf{RMSD}_{\mathsf{local}}$ errors are 1.40 and 2.92 Å. If only were the energy function able to recognize the geometrically best sampled conformation, the loop modeling problem would be essentially solved for loops up to 12 residues long.

The weakness of the potential function is also illustrated by the fact that the energies of the native loop and a relaxed native loop, within 0.2Å $\mathsf{RMSD}_{\mathsf{local}}$ of the native conformation, are always higher than that of the lowest energy model (data not shown).

The final indications that the optimizer performs reasonably well, are provided by the high average accuracy of the loop predictions based on an artificially improved energy function (see below) and the lack of frequent trapping into the same bad local minimum (Figure 12).

Although we have emphasized inadequacy of our energy function, the current optimizer is not perfect. There is some improvement in the average prediction accuracy for loops longer than approximately 6 residues even when 500 independent optimizations are performed (Figure 8). Thus, there is a need for a faster optimizer with a greater radius of convergence. This would open new applications for loop modeling, such as loop modeling on a genomic scale (Sánchez & Šali, 1998). A better optimizer might also be needed to optimize new, improved, and possibly more complex energy functions.

Conclusion

We described a completely automated loop modeling algorithm that consists of an optimization of a defined segment of protein structure in a fixed environment, guided by a pseudo energy function. The method was tested on a statistically meaningful number of loops of known structure, both in the native and near-native environments. The evaluation indicated that loops of 8 residues predicted in the native environment have 90% chance to be modeled with useful accuracy ($\mathsf{RMSD}_{\mathsf{local}}$ less than 2Å). Even 12-residue loops are modeled with useful accuracy in 30% of the cases. When the RMSD distortion of the mainchain stem atoms is 2.5Å, the average loop prediction error increased by 180, 25 and 3% for 4, 8 and 12-residue loops, respectively. It is not anymore too optimistic to expect useful models for loops as long as 12 residues, if the environment of the loop is at least approximately correct. It is possible to estimate whether or not a given loop prediction is correct, based on the structural variability of the independently derived lowest energy loop conformations.

The method is flexible and can model any subset of atoms. It is technically applicable to modeling of loops with bound ligands and several interacting loops, although it has not been evaluated in such contexts yet. Moreover, the method can incorporate additional structural information, such as restraints implied by disulfide bonds and metal binding sites.

The CASP3 meeting revealed that practical applications of loop modeling are severely limited by the need to model loops in an approximately correct environment. This is necessary because the environment has to be fixed during loop modeling; if it were not, the number of optimized degrees of freedom would increase several fold, making the modeling problem generally too hard for the current methods. If there are alignment errors at the stem residues or at the other environment residues, loop modeling is not likely to result in an accurate model. This means that loop modeling is most useful for target sequences that share more than 30% sequence identity with the template structures. Another limitation exposed by our experience at CASP3 is the need for a method that determines the regions in the target sequence to be modeled as loops. While the insertions have to fall into this category, there are generally additional regions that are aligned with templates but could benefit from an accurate "loop" modeling. These regions cannot be reliably identified at the present.

It was shown that the accuracy of the predictions is limited primarily by the accuracy of the energy function, not the thoroughness of the optimizer. The improvements of the energy function that are most likely to result in better predictions include explicit modeling of the *cis*-peptide states, a more accurate representation of the Φ, Ψ angles, as well as more accurate non-bonded terms. Better Φ, Ψ restraints might be obtained by expressing them with 2D cubic splines instead of bivariate Gaussian functions, and by explicitly taking into account the dependence of the mainchain conformation of a residue on the preceding and subsequent residues in the sequence. The non-bonded terms would probably benefit from a description of hydrogen bonds and solvent that is more physical than the current statistical potential. For example, an all-hydrogen atom and solvent representations, such as the generalized Born model (Rapp & Friesner, 1999; Dominy & Brooks III, 1999), may be necessary. To test some of these suggestions, we plan to evaluate the latest generation of molecular mechanics force fields (MacKerell *et al.*, 1998; Cornell *et al.*, 1995) for loop modeling. We will begin by a detailed analysis of the correlations of the prediction accuracy with the properties and interactions of the predicted and native loops.

The differences in length and conformation of loop regions in a family of related proteins are frequently responsible for the specificity of ligand binding. Thus, accurate modeling of loops is essential for structure-based prediction of function from sequence. For example, a comparative model can sometimes be used with computational ligand docking to find a putative ligand or resolve preferences within a limited set of ligands (Ring *et al.*, 1993; Xu *et al.*, 1996). A serious complication is that the ligand may induce conformational changes in loops with which it interacts. Thus, it is not always sufficient to be able to model loops on their own, as addressed in this paper. Instead, the protein and the ligand should ideally be modeled simultaneously. Nevertheless, induced fit is frequently small. As a result, modeling of ligand binding loops in the *apo* state can still be useful for studying functional differences within a family of proteins, as well as for the "*ab initio*" prediction of protein function by recognition of ligand binding sites (Jones & Thornton, 1997; Fetrow *et al.*, 1998; Russell *et al.*, 1998; Kleywegt, 1999; Wei *et al.*, 1999; Kasuya & Thornton, 1999). The current method presents a significant improvement in the modeling of loops in protein structures.

Acknowledgment

We are grateful to Jeff Tsai for help in the initial stages of this project, and to Drs. Azat Badretdinov, Ram Samudrala, and especially Francisco Melo for several non-bonded energy functions. AF is a Burroughs Wellcome Fellow. AŠ is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow. This work has also been aided by grants from NIH (GM 54762) and NSF (BIR-9601845).

References

Abagyan R, Batalov S, Cardozo T, Totrov M, Webber J, Zhou Y. 1997. Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. *Proteins Suppl. 1*:29–37.

Abagyan R, Totrov M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol. 235*:983–1002.

Abola EE, Bernstein FC, Bryant SH, Koetzle T, Weng J. 1987. Protein data bank. In *Crystallographic databases — Information, content, software systems, scientific applications*, (Allen FH, Bergerhoff G, Sievers R, eds), pp. 107–132. Data Commission of the International Union of Crystallography Bonn/Cambridge/Chester.

Bajorath J, Sheriff S. 1996. Comparison of an antibody model with an X-ray structure; the variable fragment of BR96. *Proteins 24*:152–157.

Benner SA, Gonnet GH, Cohen MA. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol. 229*:1065–1082.

Blundell TL, Sternberg MJE, Sibanda BL, Thornton JM. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature 326*:347–352.

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

Brower RC, Vasmatzis G, Silverman M, DeLisi C. 1993. Exhaustive conformational search and simulated annealing for models of lattice peptides. *Biopolymers 33*:329–334.

Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RC. 1969. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol. 42*:65–86.

Bruccoleri BR, Karplus M. 1990. Conformational sampling using high temperature molecular dynamics. *Biopolymers 29*:1847–1862.

Bruccoleri RE. 1993. Application of systematic conformational search to protein modeling. *Molecular Simulation 10*:151–174.

Bruccoleri RE, Haber E, Novotny J. 1988. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature 335*:564–568.

Bruccoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers 26*:137–168.

Carlacci L, Englander SW. 1993. The loop problem in proteins: Monte Carlo simulated annealing approach. *Biopolymers 33*:1271–1286.

Carlacci L, Englander SW. 1996. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. *Comp. Chem. 17*:1002–1012.

Chan HS, Dill KA. 1989. Intrachain loops in polymers: effects of excluded volume. *J. Chem. Phys. 90*:492–509.

Cheng B, Nayeem A, Scheraga HA. 1996. From secondary structure to three-dimensional structure: improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins. *J. Comp. Chem. 17*:1453–1480.

Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol. 196*:901–917.

Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SEV, Poljak RJ. 1986. The predicted structure of immunoglobulin d1.3 and its comparison with the crystal structure. *Science 233*:755–758.

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. 1989. Conformation of immunoglobulin hypervariable regions. *Nature 342*:877–883.

Chung SY, Subbiah S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure 4*:1123–1127.

Claessens M, Cutsem EV, Lasters I, Wodak S. 1989. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng. 4*:335–345.

Cohen BI, Presnell SR, Cohen FE. 1993. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci. 2*:2134–2145.

Cohen FE, Abarbanel RM, Kuntz ID, Fletterick RJ. 1986. Turn prediction in proteins using a pattern-matching approach. *Biochemistry 25*:266–275.

Collura V, Higo J, Garnier J. 1993. Modeling of protein loops by simulated annealing. *Protein Sci. 2*:1502–1510.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KMJ, Fergusson DM, Spellmeyer DC, Fox DC, Caldwell JW, Kollman PA. 1995. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc. 117*:5179–5197.

Deane CM, Blundell TL. 2000. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins 40*:135–144.

Debnath AK. 1997. A 'fragment fitting approach' to model disulfide loops by utilizing homologous peptide fragments from unrelated proteins of known structures: Application to the V3 loop of the HIV-1 envelope glycoprotein gp120. *J. Mol. Model. 3*:31–47.

Dominy BN, Brooks III CL. 1999. Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B 103*:3765–3773.

Donate LE, Rufino SD, Canard LHJ, Blundell TL. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Science 5*:2600–2616.

Dudek MJ, Ramnarayan K, Ponder JW. 1998. Protein structure prediction using a combination of sequence homology and global energy minimization: II, Energy functions. *J. Comp. Chem. 19*:548–573.

Dudek MJ, Scheraga HA. 1990. Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops. *J. Comp. Chem. 11*:121–151.

Evans JS, Mathiowetz AM, Chan SI, Goddard III WA. 1995. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. *Protein Sci. 4*:1203–1216.

Fechteler T, Dengler U, Schomburg D. 1995. Prediction of protein three-dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. *J. Mol. Biol. 253*:114–131.

Fetrow JS, Godzik A, Skolnick J. 1998. Functional analysis of the *escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol. 282*:703–711.

Fidelis K, Stern PS, Bacon D, Moult J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng. 7*:953–960.

Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCP603 from many randomly generated loop conformations. *Proteins 1*:342–362.

Finkelstein AV, Reva BA. 1992. Search for the stable state of a short chain in a molecular field. *Prot. Eng. 5*:617–624.

Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci. 2*:1811–1826.

Fushman D, Cahill S, , Cowburn D. 1997. The main chain dynamics of the dynamin pleckstrin homology (PH) domain in solution: Analysis of 15N relaxation with monomer/dimer equilibration. *J. Mol. Biol. 266*:173–194.

Go N, Abe H. 1984. The consistency principle in protein structure and pathways of folding. *Adv. Biophysics 18*:149–164.

Greer J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. USA 77*:3393–3397.

Guenther B, Onrust R, Sali A, ODonnell M, Kuriyan J. 1997. Crystal structure of the δ' subunit of the clamp-loader complex of *E. coli* DNA polymerase III. *Cell 91*:335–345.

Higo J, Collura V, Garnier J. 1992. Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. *Biopolymers 32*:33–43.

Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P. 1998. Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol. 280*:323–326.

Janin J, Rodier F. 1995. Protein-protein interaction at crystal contacts. *Proteins 23*:580–587.

Jones DT. 1999. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol. 287*:797–815.

Jones S, Thornton JM. 1997. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol. 272*:133–143.

Jones S, van Heyningen P, Berman HM, Thornton JM. 1999. Protein-DNA interactions: a structural analysis. *J. Mol. Biol. 287*:877–896.

Jones TA, Kleywegt GJ. 1999. Casp3 comparative modeling evaluation. *Proteins Suppl. 3*:30–46.

Jones TH, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J. 5*:819–822.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kasuya A, Thornton JM. 1999. Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol. 286*:1673–1691.

Kick EK, Roe DC, Skillman AG, Liu G, Ewing TJ, Sun Y, Kuntz ID, Ellman JA. 1997. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol. 4*:297–307.

Kidera A. 1995. Enhanced conformational sampling in Monte Carlo simulations of proteins: applications to a constrained peptide. *Proc. Natl. Acad. Sci. USA 92*:9886–9889.

Kinoshita K, KSadanami, Kidera A, Go N. 1999. Structural motif of phosphate-binding site common to various protein superfamilies: all-againt-all structural comparison of protein-mononucleotide complexes. *Prot. Eng. 12*:11–14.

Kleywegt GJ. 1999. Recognition of spatial motifs in protein structures. *J. Mol. Biol. 285*:1887–1897.

Koehl P, Delarue M. 1995. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in protein homology modelling. *Nature Struct. Biol. 2*:163–170.

Kwasigroch JM, Chomilier J, Mornon JP. 1996. A global taxonomy of loops in globular proteins. *J. Mol. Biol. 259*:855–872.

Lambert MH, Scheraga HA. 1989*a*. Pattern recognition in the prediction of protein structure I. Tripeptide conformational probabilities calculated from the amino acid sequence. *J. Comp. Chem. 10*:770–797.

Lambert MH, Scheraga HA. 1989*b*. Pattern recognition in the prediction of protein structure II. Chain conformation from a probability-directed search procedure. *J. Comp. Chem. 10*:798–816.

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

Lessel U, Schomburg D. 1994. Similarities between protein 3D structures. *Protein Eng. 7*:1175–1187.

Levitt M. 1983. Molecular dynamics of native protein. II. Analysis and nature of motion. *J. Mol. Biol. 168*:621–657.

Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol. 226*:507–533.

Lu Y, Valentine JS. 1997. Engineering metal-binding sites in proteins. *Curr. Opin. Struct. Biol. 7*:495–500.

Luthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*:83–85.

MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr R, Evanseck J, Field M, Fischer S, Gao J, Guo H, Ha S, Joseph McCarthy D, Kuchnir L, Kuczera K, Lau F, Mattos C, Michnick

S, Ngo T, Nguyen D, Prodhom B, Reiher III W, Roux Band Schlenkrich M, Smith J, Stote Rand Straub J, Watanabe M, Wiorkiewicz Kuczera J, Yin D, Karplus M. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B 102*:3586–3616.

Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS. 1996. ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nature Biotechnology 14*:323–328.

Marti Renom MA, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes. *Ann. Rev. Biophys. Biomolec. Struct. 29*:291–325.

Martin A, MacArthur M, Thornton J. 1997. Assessment of comparative modeling in CASP2. *Proteins Suppl. 1*:14–28.

Martin ACR, Cheetham JC, Rees AR. 1989. Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl. Acad. Sci. USA 86*:9268–9272.

Martin ACR, MacArthur MW, Thornton JM. 1997. Assessment of comparative modeling in casp2. *Proteins Suppl 1*:14–28.

Martin ACR, Thornton JM. 1996. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol. 263*:800–815.

Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM. 1992. Modeling the anti-CEA antibody combining site by homology and conformational search. *Proteins 14*:483–498.

Mattos C, Petsko GA, Karplus M. 1994. Analysis of two-residue turns in proteins. *J. Mol. Biol. 238*:733–747.

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

Melo F, Feytmans E. 1997. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol. 267*:207–222.

Mezei M. 1998. Chameleon sequences in the PDB. *Prot. Eng. 11*:411–414.

Mosimann S, Meleshko R, James MNG. 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins 23*:301–317.

Moult J, Hubbard T, Fidelis K, Pedersen JT. 1999. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins Suppl. 3*:2–6.

Moult J, James MNG. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins 1*:146–163.

Nakajima N, Higo J, Kidera A. 2000. Free energy landscapes of peptides by enchanced conformational sampling. *J. Mol. Biol. 296*:197–216.

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. 1997. An automated classification of the structure of protein loops. *J. Mol. Biol. 266*:814–830.

Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. *J. Mol. Biol. 224*:461–471.

Pellequer J, Chen S. 1997. Does conformational free energy distinguish loop conformations in proteins? *Biophys. J. 73*:2359–2375.

Perona JJ, Craik CS. 1995. Structural basis of substrate specificity. *Protein Science 4*:337–360.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol. 193*:775–791.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical Recipes, 2nd edition.* Cambridge: Cambridge University Press.

Ramachandran GE, Ramakrisnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain conformations. *J. Mol. Biol. 7*:95–99.

Rao U, Teeter MM. 1993. Improvement of turn structure prediction by molecular dynamics: a case study of α-purithionin. *Protein Eng. 6*:837–847.

Rapp CS, Friesner RA. 1999. Prediction of loop geometries using a generalized Born model of solvation effect. *Proteins 35*:173–183.

Reczko M, Martin ACR, Bohr H, Suhai S. 1995. Prediction of hypervariable CDR-H3 loop structures in antibodies. *Protein Engineering 8*:389–395.

Ring CS, Cohen FE. 1994. Conformational sampling of loop structures using genetic algorithm. *Isr. J. Chem. 34*:245–252.

Ring CS, Kneller DG, Langridge R, Cohen FE. 1992. Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol. 224*:685–699.

Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE. 1993. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA 90*:3583–3587.

Ringe D, Petsko GA. 1985. Mapping protein dynamics by x-ray diffraction. *Prog. Biophys. Molec. Biol. 45*:197–235.

Rosenbach D, Rosenfeld R. 1995. Simultaneous modeling of multiple loops in proteins. *Protein Sci. 4*:496–505.

Rosenfeld R, Zheng Q, Vajda S, DeLisi C. 1993. Computing the structure of bound peptides: application to antigen recognition by class I MHCs. *J. Mol. Biol. 234*:515–521.

Rufino SD, Donate LE, Canard LHJ, Blundell TL. 1997. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modeling. *J. Mol. Biol. 267*:352–367.

Russell RB, Sasieni PD, Sternberg MJE. 1998. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol. 282*:903–918.

Rychlewski L, Zhang B, Godzik A. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des. 3*:229–238.

Sali A. 1998. 100,000 protein structures for the biologist. *Nat. Struct. Biol. 5*:1029–1032.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol. 234*:779–815.

Sali A, Sanchez R, Badretdinov AY, Fiser A, Melo F, Overington JP, Feyfant E, Marti Renom MA. 1999. MODELLER, *A Protein Structure Modeling Program, Release 5.* URL `http://guitar.rockefeller.edu/`.

Samudrala R, Moult J. 1998. A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol. 279*:287–302.

Sanchez R, Sali A. 1997*a*. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl. 1*:50–58.

Sanchez R, Sali A. 1997*b*. Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol. 7*:206–214.

Sanchez R, Sali A. 1999. MODBASE : a database of comparative protein structure models. *Bioinformatics 15*:1060–1061.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*:56–68.

Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. 1987. Predicting antibody hypervariable loop conformation: i. Ensembles of random conformations for ring-like structures. *Biopolymers 26*:2053–2085.

Shepherd AJ, Gorse D, Thornton JM. 1999. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Sci. 8*:1045–1055.

Sibanda BL, Blundell TL, Thornton JM. 1989. Conformation of β-hairpins in protein structures: a systematic classification with applications to modelling by homology, electron deosity fitting and protein engineering. *J. Mol. Biol. 206*:759–777.

Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol. 213*:859–883.

Sippl MJ. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins 17*:355–362.

Smith KC, Honig B. 1994. Evaluation of the conformational free energies of loops in proteins. *Proteins 18*:119–132.

Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA 95*:13597–13602.

Sudarsanam S, DuBose RF, March CJ, Srinivasan S. 1995. Modeling protein loops using a ϕ_{i+1}, ψ_i dimer database. *Protein Sci. 4*:1412–1420.

Summers NL, Karplus M. 1990. Modeling of globular proteins: a distance-based search procedure for the construction of insertion/deletion regions and pro $\rightarrow$ non-pro mutations. *J. Mol. Biol. 216*:991–1016.

Tanner JJ, Nell LJ, McCammon JA. 1992. Anti-insulin antibody structure and conformation. II. Molecular dynamics with explicit solvent. *Biopolymers 32*:23–32.

Thanki N, Zeelen JP, Mathieu M, Jaenicke R, Abagyan RA, Wierenga RK, Schliebs W. 1997. Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop. *Protein Eng. 10*:159–167.

Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL. 1993. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol. 229*:194–220.

Tramontano A, Chothia C, Lesk AM. 1989. Structural determinants of the conformations of medium sized loops in proteins. *Proteins 6*:382–394.

Tramontano A, Lesk AM. 1992. Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins 13*:231–245.

Vajda S, DeLisi C. 1990. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers 29*:1755–1772.

van Vlijmen HWT, Karplus M. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol. 267*:975–1001.

Vasmatzis G, Brower RC, DeLisi C. 1994. Predicting immunoglobulin-like hypervariable loops. *Biopolymers 34*:1669–1680.

Wang J, Kollman PA, Kuntz ID. 1999. Flexible ligand docking: a multistep strategy approach. *Proteins 36*:1–19.

Wei L, , Huang ES, Altman RB. 1999. Are predicted structures good enough to preserve functional sites? *Structure 7*:643–650.

Wintjens RT, Rooman MJ, Wodak SJ. 1996. Automatic classification and analysis of α–α turn motifs in proteins. *J. Mol. Biol. 255*:235–253.

Wojcik J, Mornon JP, Chomilier J. 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol. 289*:1469–1490.

Wu SJ, Dean DH. 1996. Functional significance of loops in the receptor binding domain of *Bacillus thuringiensis* CryIIIA δ-endotoxin. *J. Mol. Biol. 255*:628–640.

Xu LZ, Sanchez R, Sali A, Heintz N. 1996. Ligand specificity of brain lipid binding protein. *J.Biol.Chem. 271*:24711–24719.

Zheng Q, Kyle DJ. 1994. Multiple copy sampling: rigid versus flexible protein. *Proteins 19*:324–329.

Zheng Q, Kyle DJ. 1996. Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings. *Proteins 24*:209–217.

Zheng Q, Rosenfeld R, DeLisi C, Kyle DJ. 1994. Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations. *Protein Sci. 3*:493–506.

Zheng Q, Rosenfeld R, Vajda S, DeLisi C. 1993*a*. Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci. 2*:1242–1248.

Zheng Q, Rosenfeld R, Vajda S, DeLisi C. 1993*b*. Loop closure via bond scaling and relaxation. *J. Comp. Chem. 14*:556–565.

Tables and Figures

Loop length	Accuracy class Good	Medium	Bad
1	100.0	0.0	0.0
2	97.5	2.5	0.0
3	100.0	0.0	0.0
4	82.5	17.5	0.0
	85.0	15.0	0.0
	42.8	57.2	0.0
5	70.0	27.5	2.5
6	57.5	27.5	15.0
7	52.5	25.0	22.5
8	30.0	47.5	22.5
	50.0	40.0	10.0
	14.1	59.4	26.5
9	20.0	35.0	45.0
10	7.5	40.0	52.5
11	2.5	27.5	72.5
12	2.5	17.5	80.0
	7.5	22.5	70.0
	0.0	6.9	93.1
13	5.0	10.0	80.0
14	2.4	2.4	95.1

Table 1: The fraction of loop predictions in the three accuracy classes, as a function of the number of residues in the loop. The accuracy classes are defined in Figure 6. The standard algorithm with 50 random initial conformations (Theory and Algorithms) was applied to 40 test loops of each length. For 4, 8, and 12-residue loops, the numbers in the 2nd line give the results for 500 random initial conformations in the native environment; the numbers in the 3rd line give the results for 50 random initial conformations in the environment distorted from 1.75 to 2.75Å (Figure 10).

Loop	Reference	Present method								
		lowest energy prediction					lowest RMSD prediction			
	global*	global*	global	local	score	rank	global	local	score	rank
2apr_76-83 (8)	5.16	1.31	1.35	0.73	-48.39	5	0.94	0.50	-44.23	2
8abp_203-208 (6)	0.28	0.38	0.37	0.24	-35.34	75	0.24	0.17	-29.25	32
2act_198-205 (8)	1.58	2.04	2.21	1.60	-10.47	62	1.93	1.09	13.95	392
8tln_E32-E38 (7)	3.70	2.03	2.26	1.82	-49.45	435	0.93	0.65	-13.30	314
3grs_83-89 (7)	4.55	0.42	0.58	0.47	2.77	6	0.43	0.31	8.71	6
5cpa_231-237 (7)	2.14	0.95	1.23	1.06	-5.32	18	1.00	0.89	11.83	244
2fb4_H26-H32 (7)	1.62	4.20	4.25	2.06	3.54	282	0.52*	0.41	15.98	148
2fbj_H100-H106 (7)	0.49	0.84	1.31	1.08	3.27	48	1.03*	0.83	47.17	532
8tln_E248-E255 (8)	1.83	0.87	0.98	0.84	-26.04	182	0.70	0.61	-10.80	246
3sgb_E199-E211 (9)	1.79	0.28	0.36	0.28	-53.36	4	0.29*	0.24	-49.14	3
3dfr_20-23 (4)	2.64	1.15	1.59	1.51	29.70	676	0.35	0.20	47.88	844
3dfr_89-93 (5)	1.62	1.02	1.14	0.85	-0.06	21	0.87	0.78	16.98	690
3dfr_120-124 (5)	0.47	0.26	0.28	0.20	0.71	217	0.23	0.15	9.96	484
3blm_131-135 (5)	0.82	0.16	0.22	0.14	-32.55	103	0.16	0.11	-26.57	75

Table 2: Comparison of predictions for 14 test loops by a recent powerful method (Tables 8 and 9 in van Vlijmen & Karplus, 1997) and the present method. The modeled loop segment is indicated after the four character PDB code and its length is given in parentheses. Since the reference study was performed, structure 3tln was replaced by 8tln in PDB; thus, 8tln had to be used here. Predictions by the present method were obtained from 1000 independent optimizations, using the non-bonded distance cutoff of 7Å. The *global* and *local* column headers stand for $\mathsf{RMSD}_{\mathsf{global}}$ and $\mathsf{RMSD}_{\mathsf{local}}$, respectively. The *global** column lists $\mathsf{RMSD}_{\mathsf{stem},\mathsf{NC}_\alpha\mathsf{C}}(\mathsf{loop},\mathsf{NC}_\alpha\mathsf{C})$. The ranks of the predictions sorted by the energy and RMSD are shown for the RMSD and energy columns, respectively. If the lowest local and global RMSD conformations are the same, the global RMSD value is indicated by a star. The energy and rank of the conformation with the lowest local RMSD are given. The CPU times for predicting one loop by the reference and present methods are approximately the same, 30 hours on an R10000-190 SGI workstation.

Figure legends

Figures 1 Weights for the mainchain conformation classes. (a) The five possible mainchain conformation classes are defined as areas A, B, E, P, and L in the Ramachandran plot spanned by the Φ and Ψ mainchain dihedral angles. The plot shows the distribution for all 217,807 residues in the 1,000 representative protein structures (Theory and Algorithms). Slightly different borders between B and P classes are used for the different residue types; the borders shown are approximate. (b) The probability that a given residue type occurs in any one of the five possible mainchain conformation classes, ω (Eq. 3).

Figures 2 Actual Ramachandran plots (left side) and fitted model surfaces (right side) for four representative residue types. The actual Ramachandran plots are the natural logarithm of the observed frequency of a given pair of the mainchain dihedral angles Φ and Ψ in the 1,000 representative protein structures (Theory and Algorithms). The modeled Ramachandran plots correspond to the restraints used in modeling, $\ln p_m(\Phi, \Psi/R)$ (Eq. 3). R=Asn (9,767 residues); R=Pro (9,755); R=Gly (16,650); R=Ala (17,454).

Figures 3 Modeling of a loop. (a) Sample model. The backbone trace of the native structure of N-carbomoylsarcosine amidohydrolase (PDB code 1nba) is light blue, the native loop of 8 residues (99–106) is dark blue, the distorted environment in which the loop is generally modeled is red, and the initial loop conformation for optimization, before atomic coordinates are shifted randomly for up to 5Å, is yellow. (b) Sample optimization. The energy (thick line) and temperature (thin line) during conjugate gradients and molecular dynamics with simulated annealing are shown (Theory and Algorithms). The initial value of the energy is very high ($\approx 10^6$).

Figures 4 Distribution of loop length. Secondary structure segments in the 1,000 representative structures (Theory and Algorithms) were defined by the program DSSP (Kabsch & Sander, 1983) (white bars). All 13,444 segments that span helices and β-strands were defined as loops. Helices were defined as contiguous segments of at least 4 residues in the H, G, or I conformation. Strands were defined as contiguous segments of at least 3 residues in the E or B conformation. The bars are divided into two parts to indicate the fraction of predictions expected to be in the good and medium classes (gray) and in the bad class (white), according to the evaluation based on 50 independent optimizations in the correct environment (Table 1). The results are similar for 500 independent optimizations in the environments with the $\mathsf{RMSD}_{\mathsf{core,NC_\alpha CO}}(\mathsf{stem}, \mathsf{NC_\alpha CO})$ error between 1.75 and 2.75 Å.

Figures 5 Correlation of $\mathsf{RMSD}_{\mathsf{local}}$ with other measures of accuracy of a loop model. The three lines are shown to facilitate interpreting the correlation; they have slopes of 0.5, 1.0, and 2.0.

Figures 6 Definition of three levels of accuracy in loop modeling. Sample mainchains of a good (a), medium (b), and bad (c) loop model are shown. The defining $\mathsf{RMSD}_{\mathsf{local}}$ ranges are $<$ 1Å, 1–2Å, and $>$ 2Å, respectively. Models for residues 28–35 in NAD–dependent formate dehydrogenase (PDB code 2nac) are shown. The $\mathsf{RMSD}_{\mathsf{core,NC_\alpha CO}}(\mathsf{stem}, \mathsf{NC_\alpha CO})$ for 3-residue stems on each side of the loop is 1.2Å. The left and right figures in each panel correspond to the $\mathsf{RMSD}_{\mathsf{global}}$ and $\mathsf{RMSD}_{\mathsf{local}}$ superposition, respectively.

Figures 7 Two sample loop predictions. On the left, the energy of the final conformations from 500 independent optimizations is plotted as a function of the $\mathsf{RMSD}_{\mathsf{local}}$ error. On the right, the C_α traces of the 17 lowest energy conformations are shown. (a) Successful modeling of residues 45–52 in 5p21; the RMSD error of the conformation with the lowest energy is 0.25Å. (b) Incorrect modeling of residues 34–41 in 1alc; the RMSD error of the conformation with the lowest energy is 3.17Å.

Figures 8 Accuracy of loop modeling as a function of thoroughness of optimization. For each of the 40 test loops of 4 residues (a), 8 residues (b), and 12 residues (c), 500 independent optimizations were performed. Each independent optimization started with a different random initial loop conformation in the native environment (Theory and Algorithms). Four different average RMSD measures for the mainchain loop atoms, calculated over the 40 test loops, are plotted as a function of the number of independent optimizations. Average $\mathsf{RMSD}_{\mathsf{global}}$ error for the models with the lowest energy, open circles and thick line. Average $\mathsf{RMSD}_{\mathsf{local}}$ error for the models with the lowest energy, filled circles and thin line. Average minimal $\mathsf{RMSD}_{\mathsf{global}}$ error, open diamonds and thick dashed line. Average minimal $\mathsf{RMSD}_{\mathsf{local}}$ error, filled diamonds and thin dashed line. The standard error of the mean is generally less than 0.1Å (not shown). The curves are the least-squares fits of the average RMSD to the number of independent optimizations, n_o. For example, for the 8-residue loops: $< \mathsf{RMSD}_{\mathsf{local}} > = -6.28 + \exp{(2.00 + 0.2593 n_o^{-0.8045})}$.

Figures 9 Accuracy of loop modeling in the correct environment as a function of loop length. Models were calculated for 40 loops at each length from 1 to 14 residues, as described in Theory and Algorithms. 50 independent optimizations were used to make each prediction. Average accuracy and the standard deviation of the accuracy are shown for each length for local (upper plot) and global superposition.

Figures 10 Accuracy of loop modeling as a function of environment error. The average RMSD error and the standard error of the mean are plotted as a function of the error in the loop environment. Although many atoms around the loop were distorted (Results and Discussion), errors in the environment are measured only by the mainchain RMSD of the three stem residues on each side of the loop, upon superposition of the non-distorted atoms of the native and model structures (Figure 3). Each of the 40 test loops of 4 residues (open diamonds), 8 residues (circles), and 12 residues (filled diamonds) have been distorted 5 times, resulting in 200 loop predictions in a distorted environment. Each individual loop prediction consisted of 50 independent optimizations. Upper plot shows the local, while the one below the evaluation by global superposition.

Figures 11 Accuracy of loop modeling as a function of loop properties. The RMSD of the lowest energy model from 500 independent optimizations is plotted against several loop properties. (a) Normalized average atomic isotropic temperature factor of the loop ($< B_{\text{iso}} >_{\text{loop}} - < B_{\text{iso}} >_{\text{protein}}$. Normalization was performed to improve direct comparison of isotropic temperature factors from independent structure determinations (Ringe & Petsko, 1985). (b) Number of non-bonded atomic contacts within the loop. (c) Number of intra-molecular atomic loop–environment contacts. (d) Number of inter-molecular atomic loop–environment contacts. The symmetry related protein structures were calculated by program CRYSPACK (Janin & Rodier, 1995), relying on the information in the PDB atom files. An atomic, non-bonded contact occurs when two atoms are at a distance of less than 4.5Å and are separated by more than three covalent bonds. The data are shown for 40 loops of 8 residues each. The Pearson correlation coefficient is also given in each panel.

Figures 12 Prediction of the accuracy of the lowest energy loop model. The $\mathsf{RMSD}_{\mathsf{local}}$ error of the best loop model correlates well with the structural variability of the 20 lowest energy models, out of the 500 independently calculated models. The variability of the lowest energy models is measured by the average of the $20 \times 19/2$ pairwise $\mathsf{RMSD}_{\mathsf{local}}$ differences. The correlation is shown for 40 loops of 8 residues each. For reference, a line is drawn that corresponds to $\mathsf{RMSD} = 2 < \mathsf{RMSD} >$.

Figures 13 Examples of loop and distortion modeling for CASP3. The native conformations are in blue, the models are in red. All the mainchain atoms are shown for the loops, only the backbone trace is plotted for the 3 stem residues on each side of the loops. (a) Loop prediction for residues 46–53 in target T0076. No template loop of the same length was available from the structurally defined members of the family. The $\mathsf{RMSD}_{\mathsf{local}}$ error is 1.37Å while the stem residues superpose on the native structure with $\mathsf{RMSD}_{\mathsf{stem,NC_\alpha CO}}(\mathsf{stem, NC_\alpha CO})$ of 1.52Å. (b) Distortion modeling of residues 80–85 in T0058. The closest template structure (PDB code 1akz, shown in yellow) had a segment aligned with the modeled segment without gaps, but their sequences were so different that the target segment was modeled as a loop. The template and target sequences are RPGIAI and QRPVPP, respectively. The $\mathsf{RMSD}_{\mathsf{global}}$, $\mathsf{RMSD}_{\mathsf{local}}$, and $\mathsf{RMSD}_{\mathsf{stem,NC_\alpha CO}}(\mathsf{stem, NC_\alpha CO})$ errors for the model are 1.57, 1.09, and 0.29 Å, respectively. The corresponding deviations between the template and the actual target structure are 0.99, 0.75, and 0.27Å.

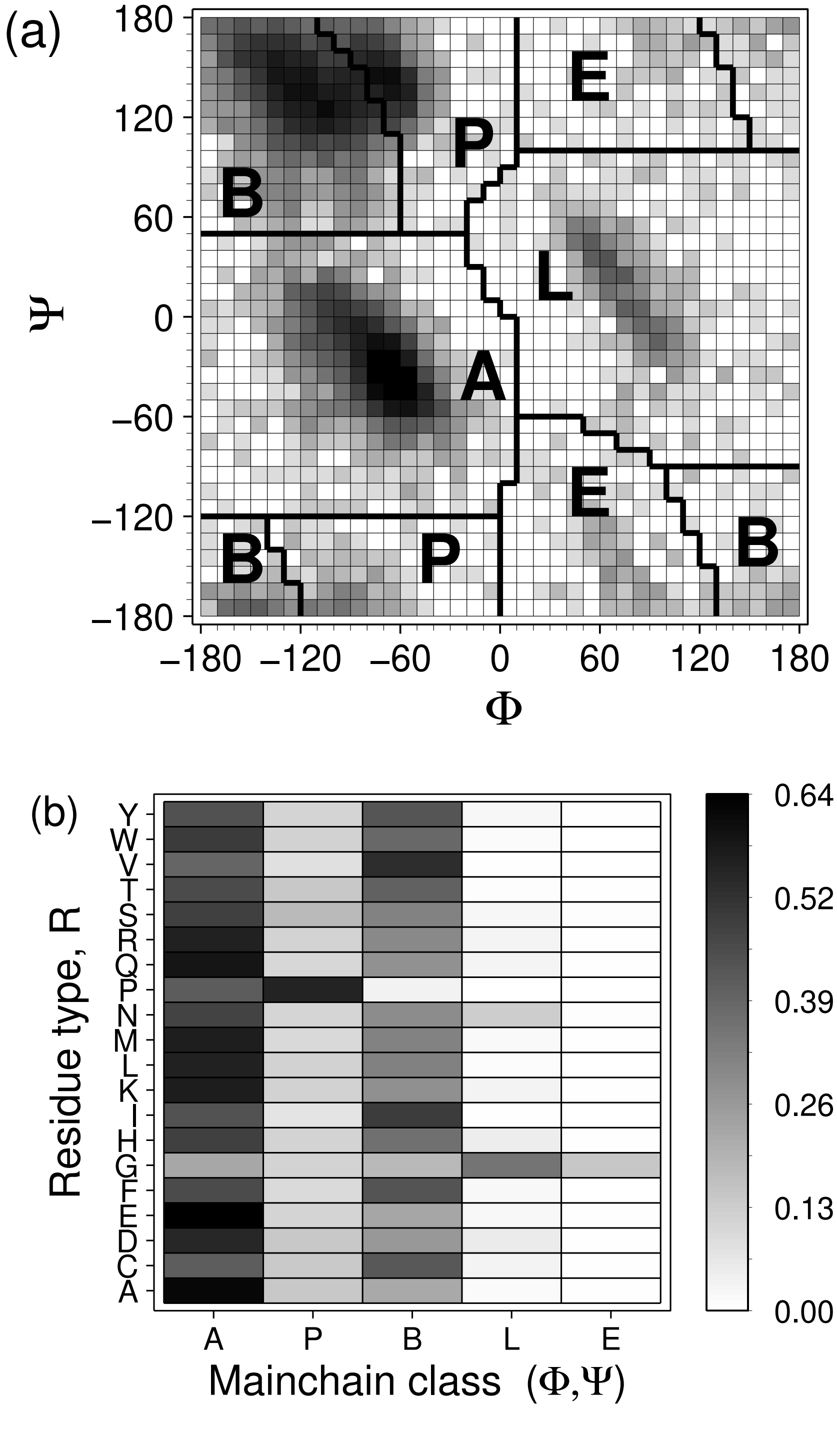

(a)
180
120
60
0
−60
−120
−180
Ψ
−180
−120
−60
0
60
120
180
Φ
B
P
E
L
A
E
B
P
B
(b)
Residue type, R
Y
W
V
T
S
R
Q
P
N
M
L
K
I
H
G
F
E
D
C
A
A
P
B
L
E
Mainchain class (Φ,Ψ)
0.64
0.52
0.39
0.26
0.13
0.00

Figure 1:

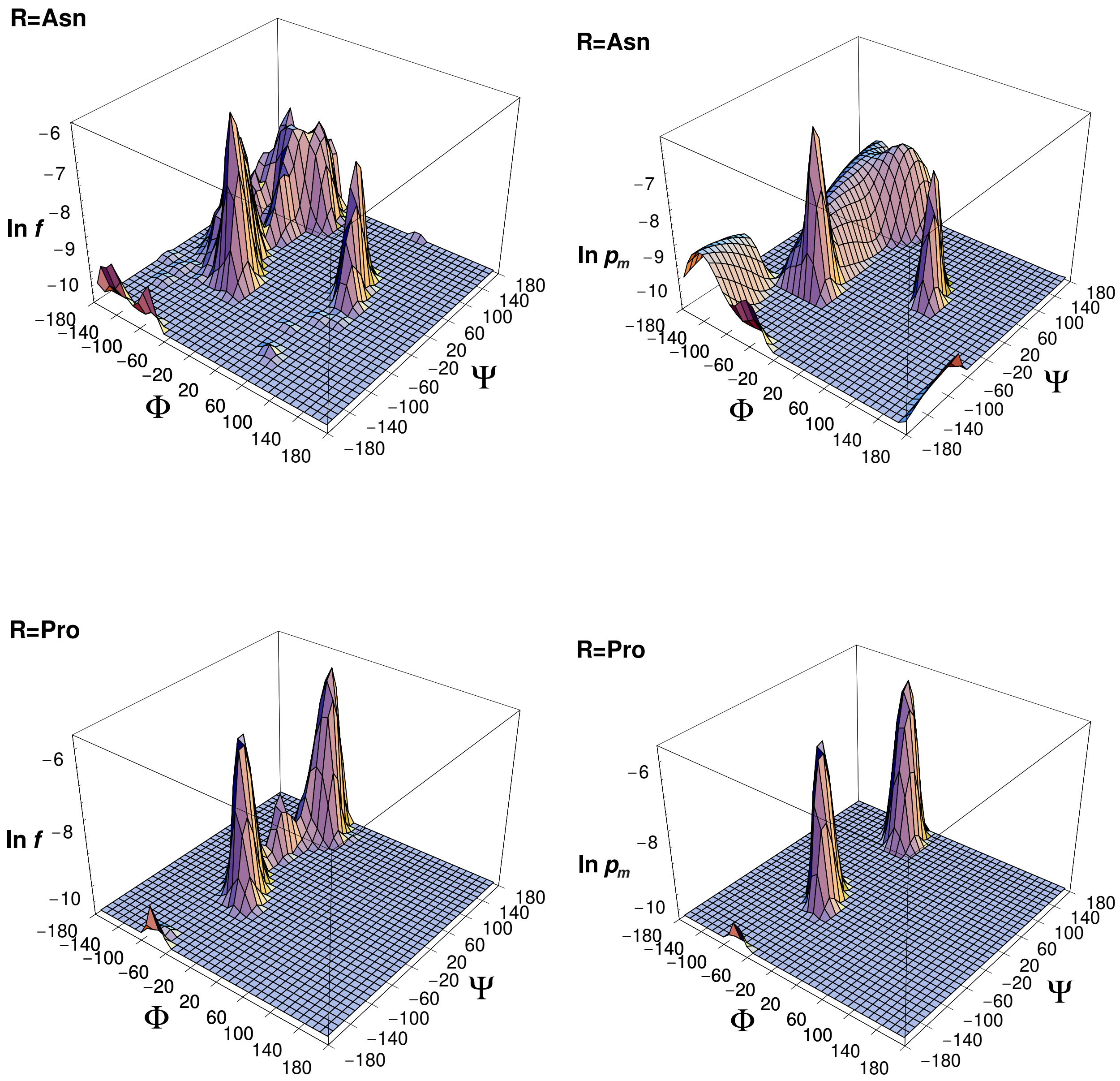


Figure 2: Continued on next page.

Figure 2 continued:

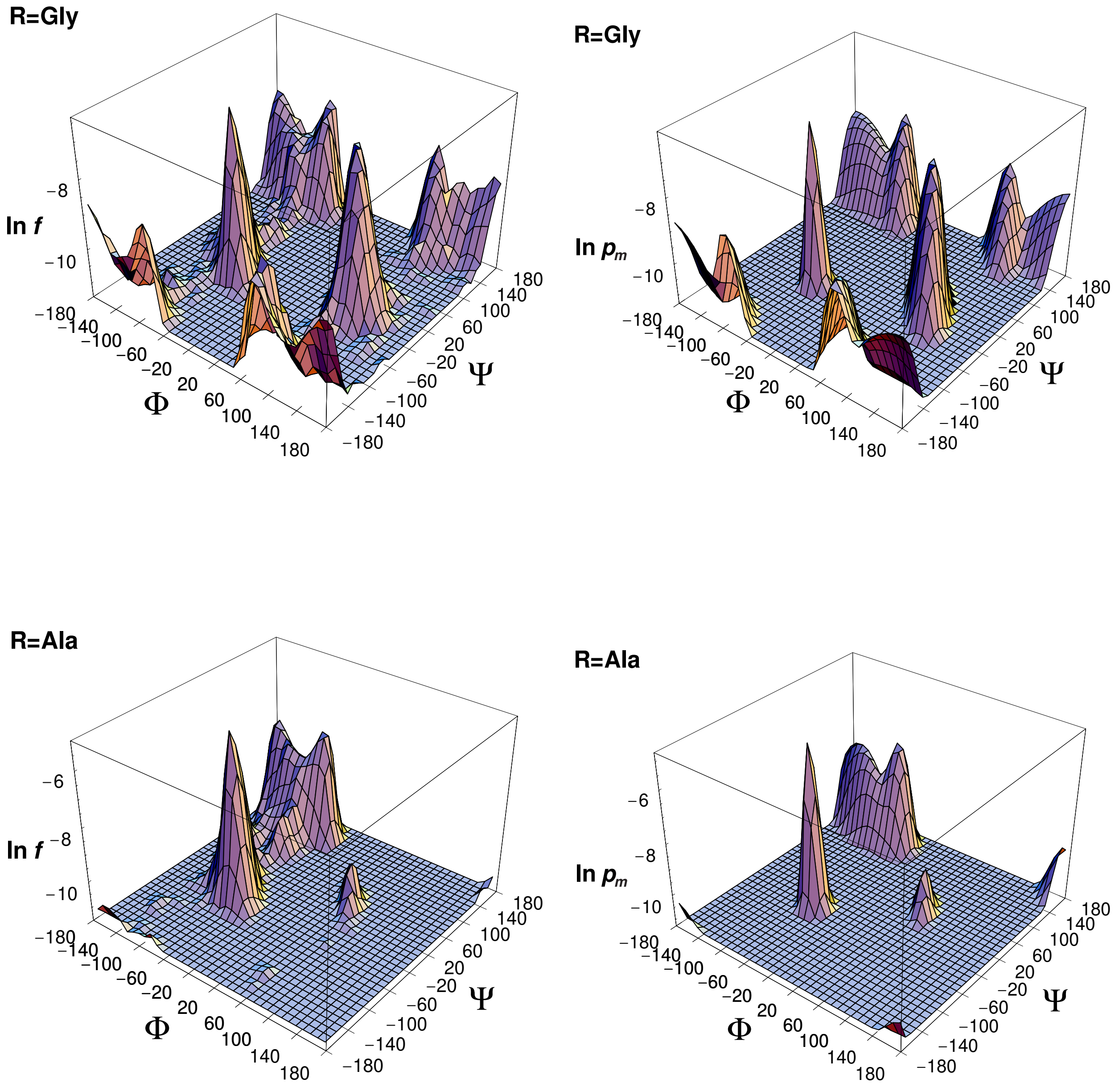
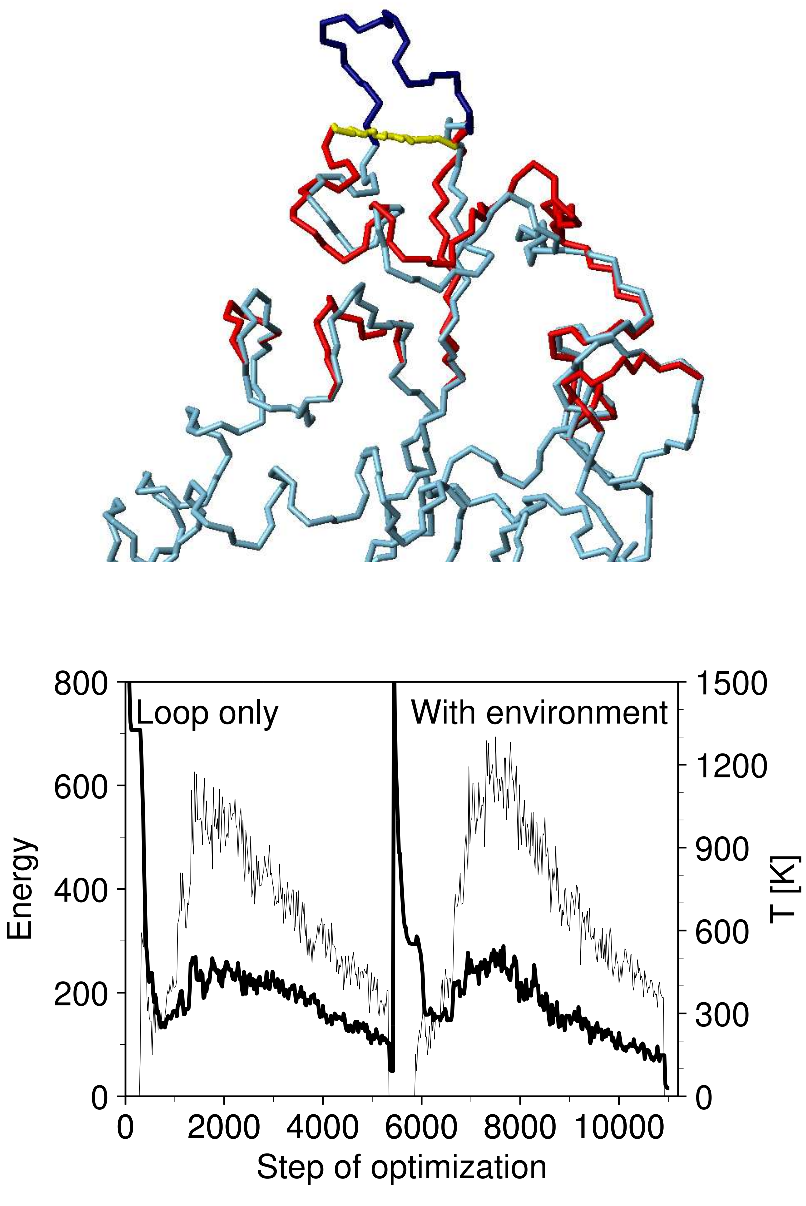


Figure 2.

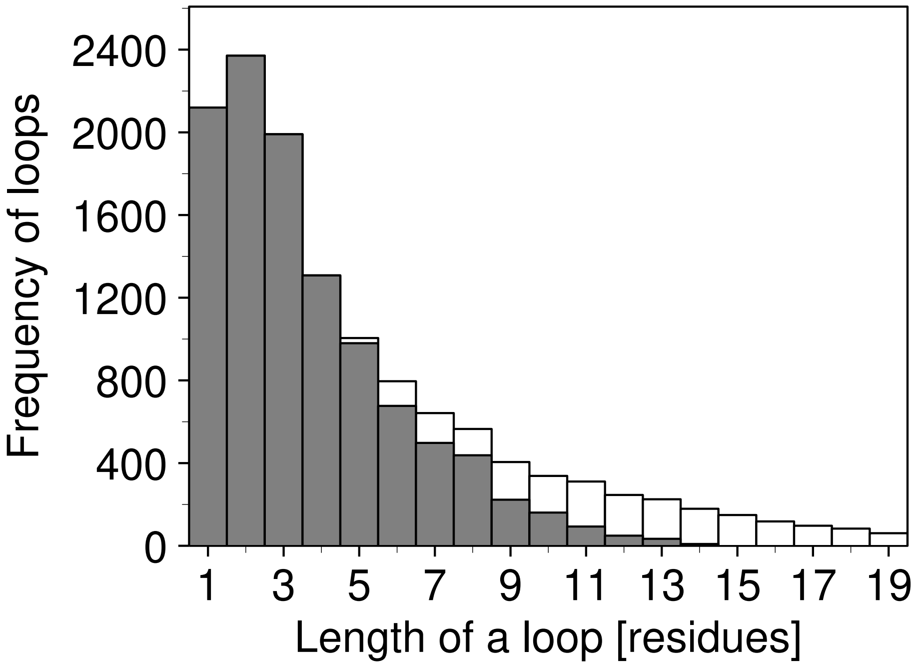
(a)

(b)



Loop only
With environment
Energy
T [K]
Step of optimization
800
600
400
200
0
1500
1200
900
600
300
0
0
2000
4000
6000
8000
10000

Figure 3:



2400
2000
1600
1200
800
400
0
Frequency of loops
1
3
5
7
9
11
13
15
17
19
Length of a loop [residues]

Figure 4:

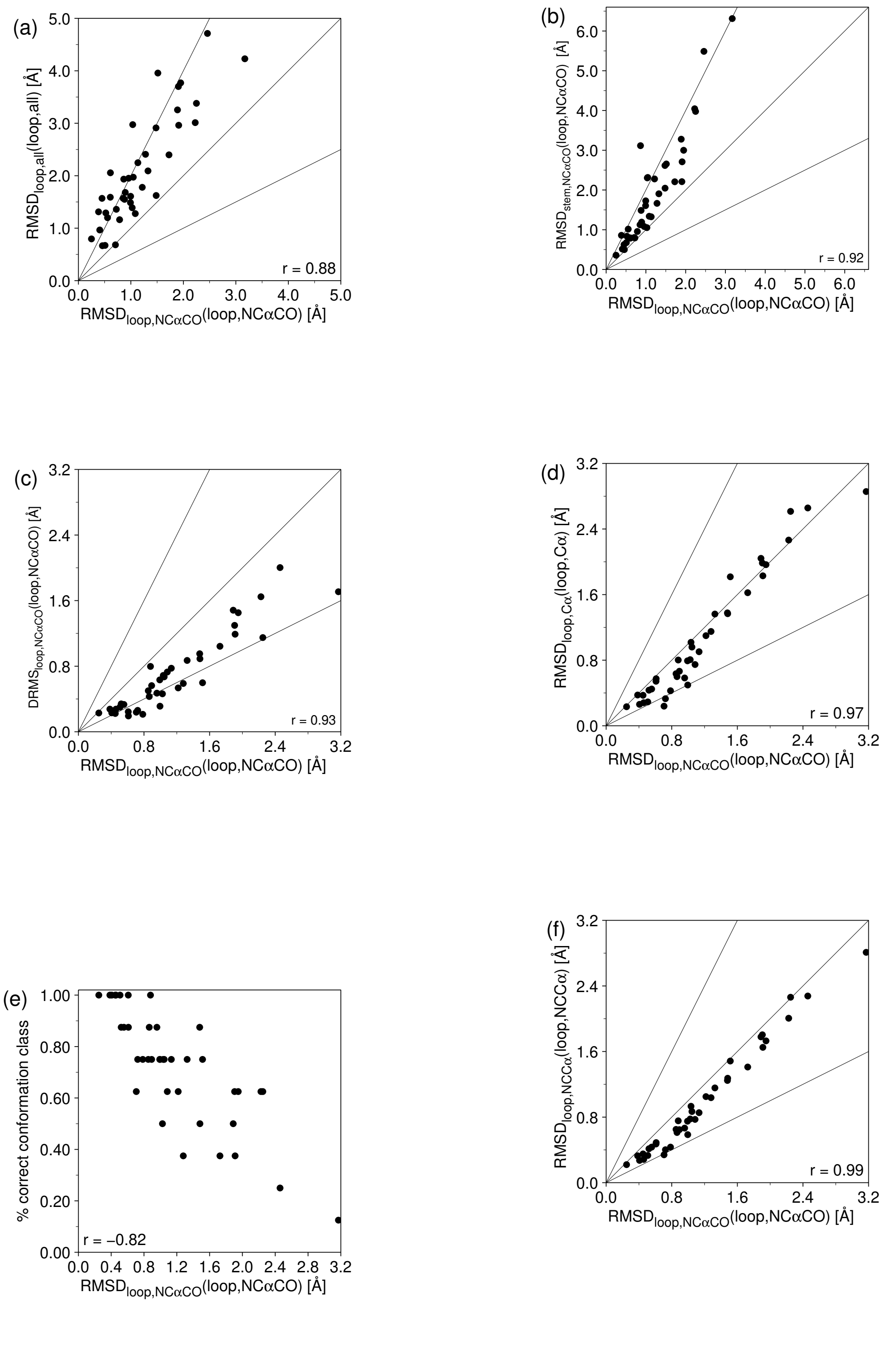
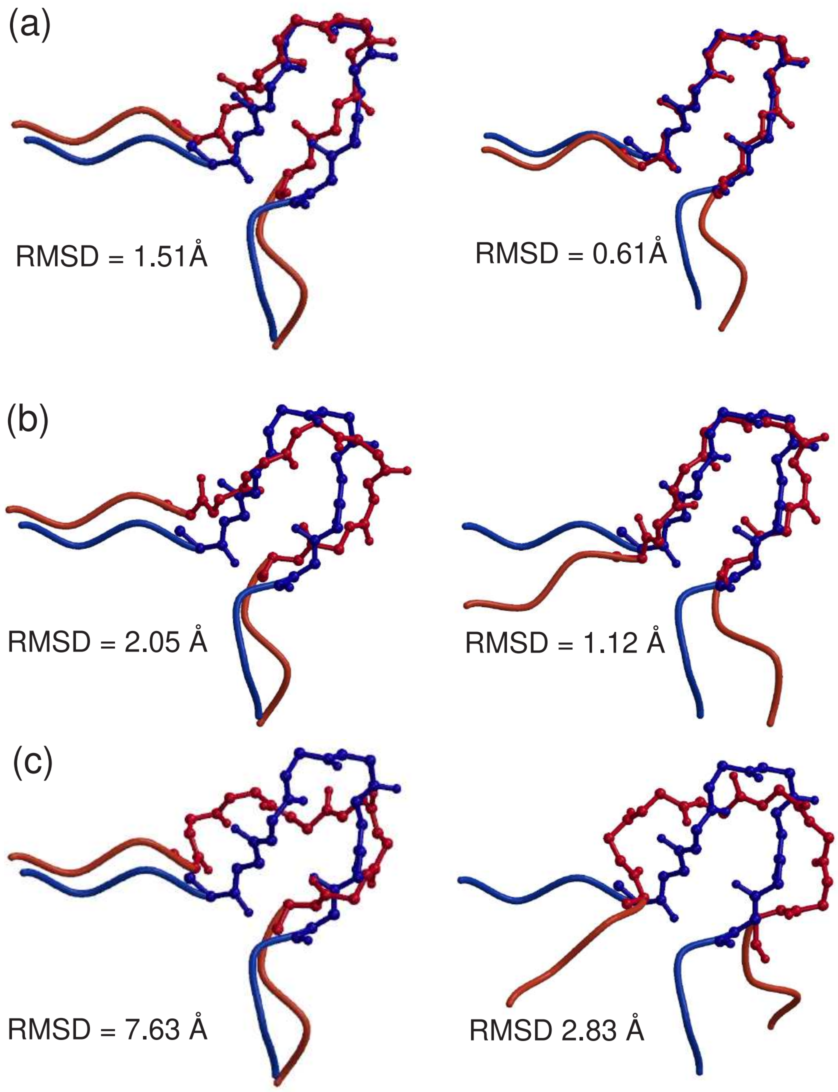

(a)
$RMSD_{loop,all}$(loop,all) [Å]
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = 0.88
(b)
$RMSD_{stem,NC\alpha CO}$(loop,NCαCO) [Å]
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = 0.92
(c)
$DRMS_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = 0.93
(d)
$RMSD_{loop,C\alpha}$(loop,Cα) [Å]
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = 0.97
(e)
% correct conformation class
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = −0.82
(f)
$RMSD_{loop,NCC\alpha}$(loop,NCCα) [Å]
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
r = 0.99

Figure 5:



(a)
RMSD = 1.51Å
RMSD = 0.61Å
(b)
RMSD = 2.05 Å
RMSD = 1.12 Å
(c)
RMSD = 7.63 Å
RMSD 2.83 Å

Figure 6:

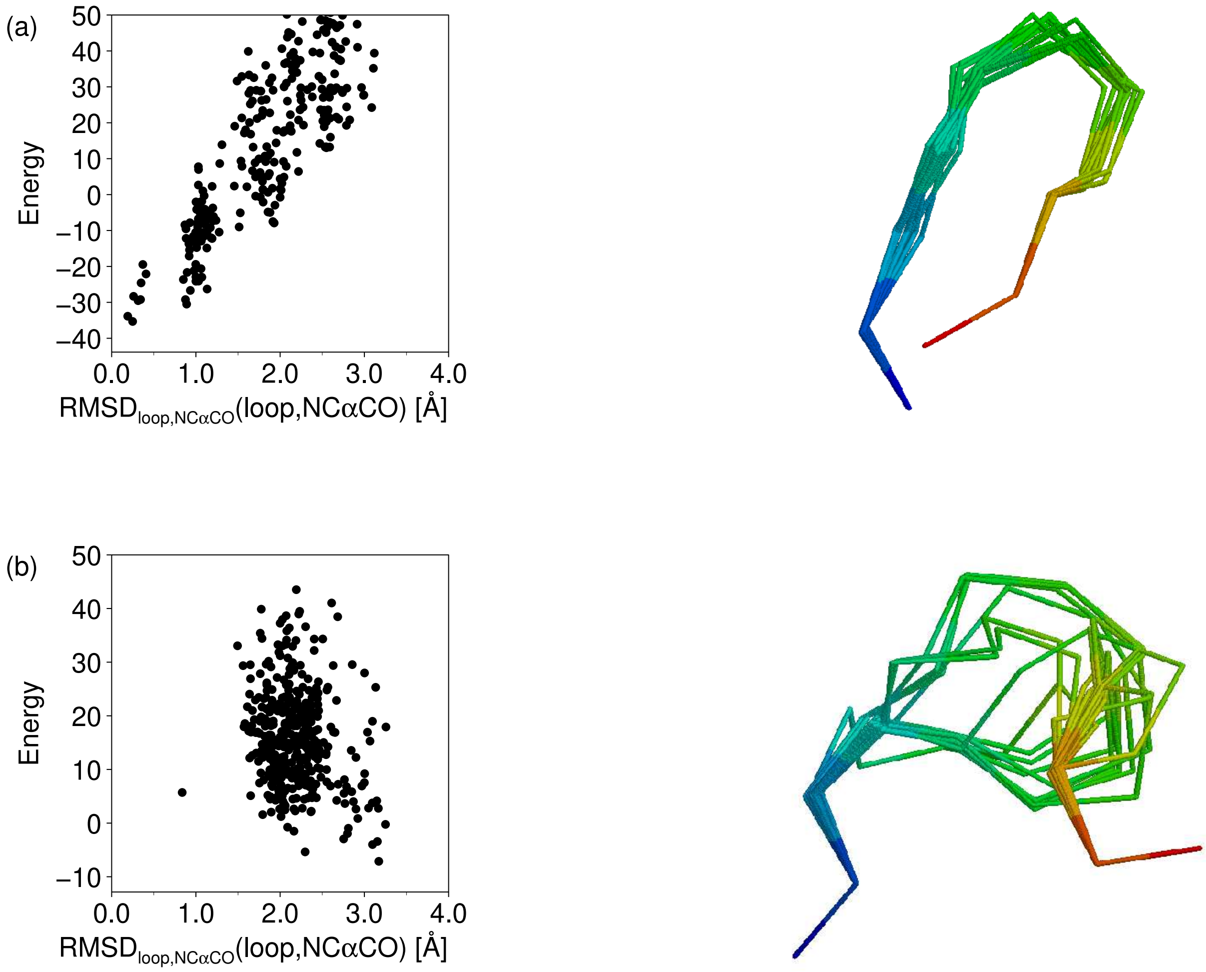

(a)
50
40
30
20
10
0
–10
–20
–30
–40
Energy
0.0
1.0
2.0
3.0
4.0
RMSD$_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
(b)
50
40
30
20
10
0
–10
Energy
0.0
1.0
2.0
3.0
4.0
RMSD$_{loop,NC\alpha CO}$(loop,NCαCO) [Å]

Figure 7:

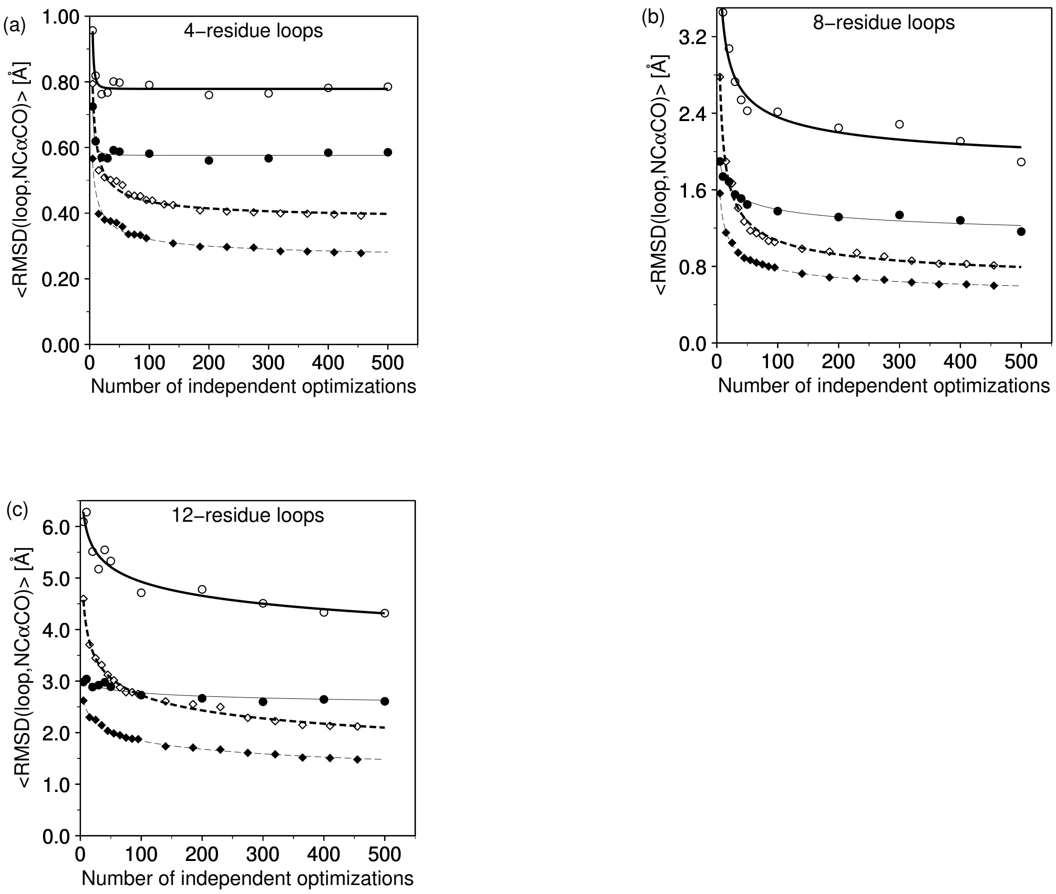
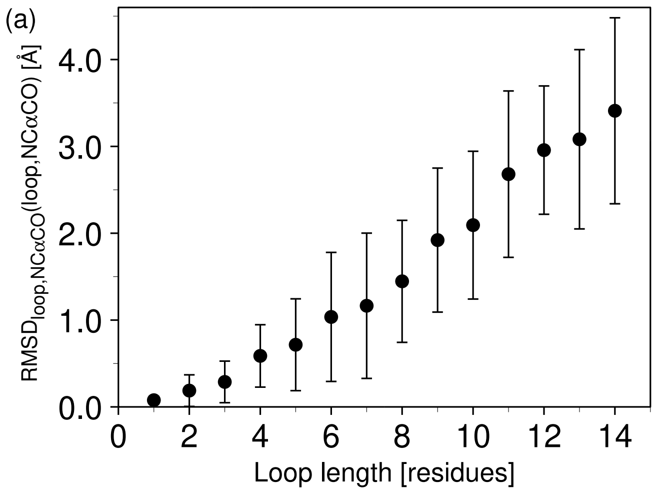

(a)
4−residue loops
<RMSD(loop,NCαCO)> [Å]
1.00
0.80
0.60
0.40
0.20
0.00
0
100
200
300
400
500
Number of independent optimizations
(b)
8−residue loops
<RMSD(loop,NCαCO)> [Å]
3.2
2.4
1.6
0.8
0.0
0
100
200
300
400
500
Number of independent optimizations
(c)
12−residue loops
<RMSD(loop,NCαCO)> [Å]
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0
100
200
300
400
500
Number of independent optimizations

Figure 8:


(a)
RMSD$_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
4.0
3.0
2.0
1.0
0.0
0
2
4
6
8
10
12
14
Loop length [residues]

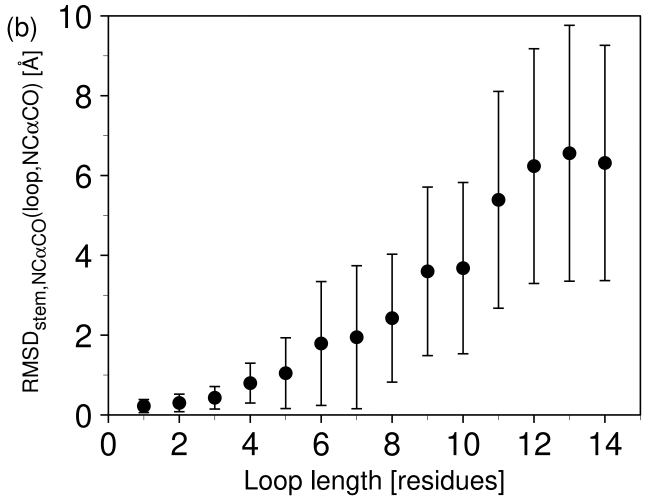

(b)
RMSD$_{stem,NC\alpha CO}$(loop,NCαCO) [Å]
10
8
6
4
2
0
0
2
4
6
8
10
12
14
Loop length [residues]

Figure 9:

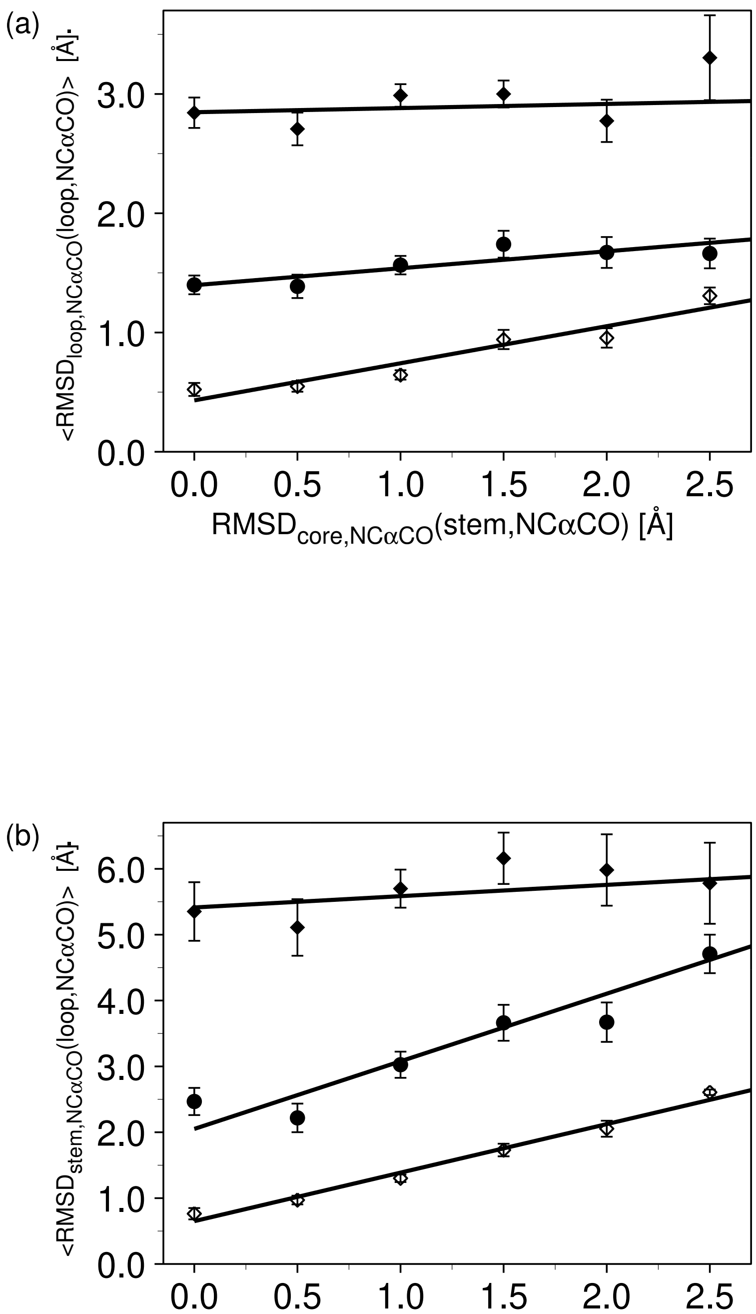
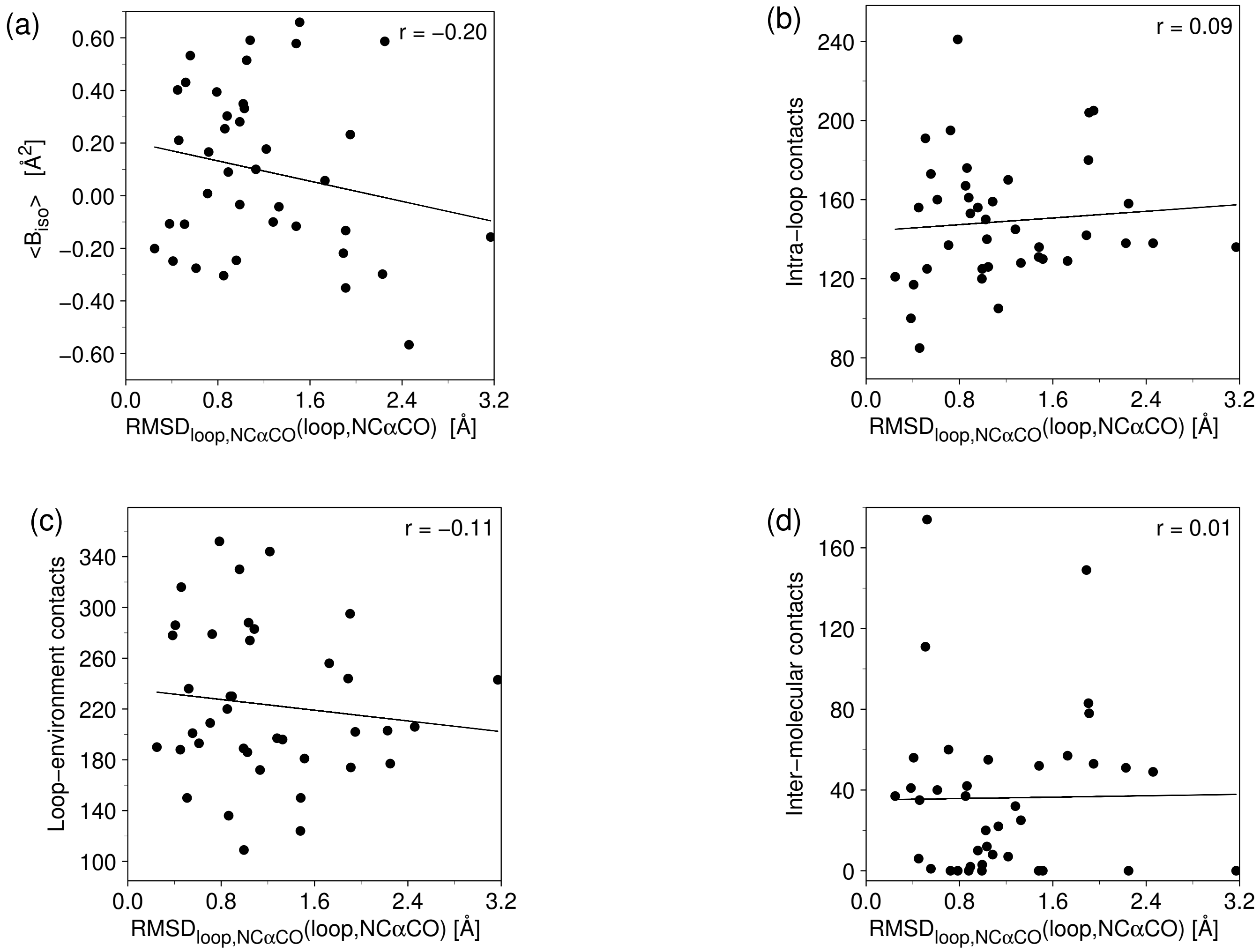
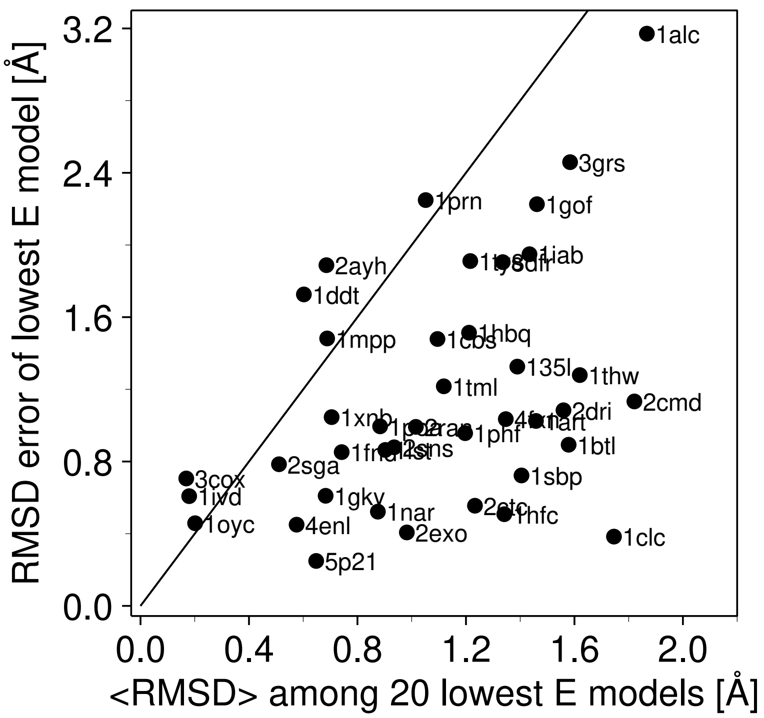

(a)
<RMSD$_{loop,NC\alpha CO}$(loop,NCαCO)> [Å]
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
2.0
2.5
RMSD$_{core,NC\alpha CO}$(stem,NCαCO) [Å]
(b)
<RMSD$_{stem,NC\alpha CO}$(loop,NCαCO)> [Å]
0.0
1.0
2.0
3.0
4.0
5.0
6.0
0.0
0.5
1.0
1.5
2.0
2.5
RMSD$_{core,NC\alpha CO}$(stem,NCαCO) [Å]

Figure 10:



(a)
r = −0.20
<B_{iso}> [$Å^2$]
0.60
0.40
0.20
0.00
−0.20
−0.40
−0.60
0.0
0.8
1.6
2.4
3.2
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
(b)
r = 0.09
Intra−loop contacts
240
200
160
120
80
0.0
0.8
1.6
2.4
3.2
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
(c)
r = −0.11
Loop−environment contacts
340
300
260
220
180
140
100
0.0
0.8
1.6
2.4
3.2
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]
(d)
r = 0.01
Inter−molecular contacts
160
120
80
40
0
0.0
0.8
1.6
2.4
3.2
$RMSD_{loop,NC\alpha CO}$(loop,NCαCO) [Å]

Figure 11:



RMSD error of lowest E model [Å]
<RMSD> among 20 lowest E models [Å]
0.0
0.4
0.8
1.2
1.6
2.0
2.4
3.2
1alc
3grs
1prn
1gof
1iab
2ayh
1ddt
1hbq
1mpp
1cbs
135l
1thw
1tml
2cmd
2dri
1xnb
1phf
1btl
2sga
3cox
1ivd
1sbp
1gky
1nar
1oyc
4enl
2exo
1hfc
1clc
5p21

Figure 12:

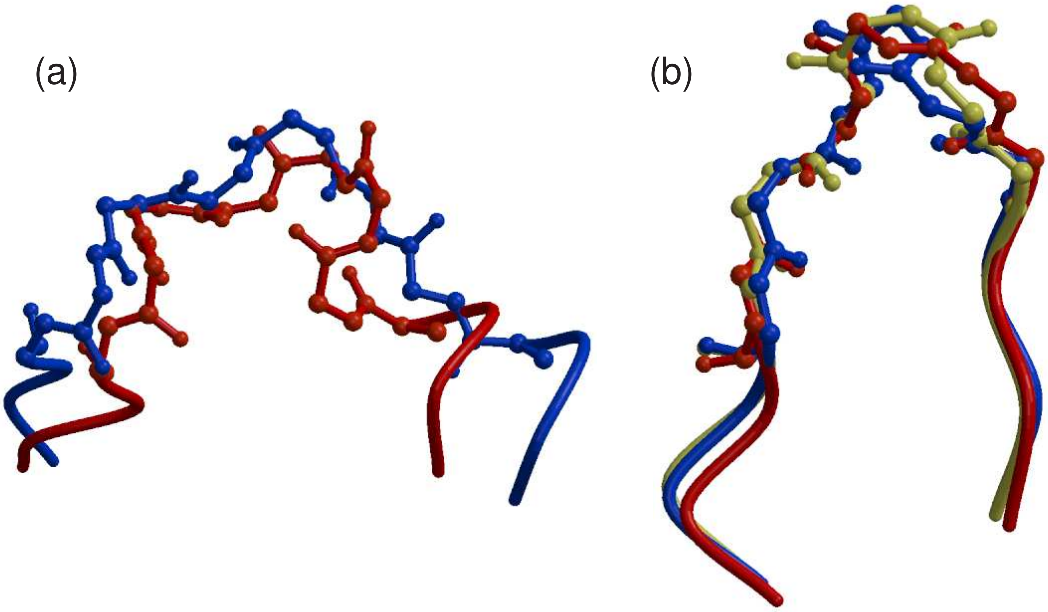


Figure 13: