J. Mol. Biol. (1997) **267**, 975–1001

JMB

PDB-based Protein Loop Prediction: Parameters for Selection and Methods for Optimization

Herman W. T. van Vlijmen[1] and Martin Karplus[1,2]*

[1]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street Cambridge, MA 02138, USA*

[2]*Laboratoire de Chimie Biophysique, Institut le Bel Université Louis Pasteur 4, Rue Blaise Pascal 67000, Strasbourg, France*

**Corresponding author*

An approach to loop prediction that starts with a database search is presented and analyzed. To obtain meaningful statistics, 130 loops from 21 proteins were studied. The correlation between the internal conformation of the loop and the conformation of the neighboring stem residues was examined. Distances between C^{α} and C^{β} of the immediate neighbor residues at each end select template loops as well as more complex (e.g. three residues on either side) matching criteria. To have a high probability that the best possible loop candidate in the database is included in the set, relatively large cutoffs for matching the interatomic distances of the stem residues have to be used in the template loop selection procedure; for loops of length 5, this results in an average of 1000 loops and for loops of length 9, the number is about 1500. The required number increases only slowly with loop length, in contrast to the exponential time increase involved in direct searches of the conformational space. The best loops among the large number of candidates can be determined by ranking them with the standard CHARMM non-bonded energy function (without electrostatics) applied to the backbone and C^{β} atoms. The same representation (backbone plus C^{β}) can be used to optimize the loop orientations relative to the rest of the protein by constrained energy minimization. Target loops that have many non-bonded contacts with the protein yield better results so that analysis of the non-bonded contacts of the selected template loops is useful in determining the expected accuracy of a prediction. The method for loop selection and optimization predicted eight (out of 18) loops of up to nine residues to an RMSD better than 1.07 Å relative to the crystal structure; for 17 of the 18 loops, one of the three lowest energy template loops had an RMSD of less than 1.79 Å.

The prediction of antibody loops from a database search is more effective than that for non-antibody loops. Provided that they belong to one of the canonical classes, very similar antibody loops are certain to exist in the database. Superposition of the stem residues for antibody loops also results in a better orientation than with arbitrary target loops because the neighboring residues tend to have a more similar β-strand structure. Two H3 loops (for which no canonical structures have been proposed) were predicted with reasonable accuracy (RMSD of 0.49 Å and 1.07 Å) even though no corresponding antibody loops were in the database.

Keywords: protein loops; structure prediction; homology modeling; structural database search; energy mimimization

Introduction

Loops in proteins can be defined as segments that do not correspond to α-helical or β-strand secondary structure. They are often of functional importance and can have key roles in recognition (antibody hypervariable loops), ligand binding (e.g. triosephosphate isomerase; Joseph *et al.*, 1990), DNA-binding (M-13 phage; Coleman *et al.*, 1987), or forming enzyme active sites (e.g. serine proteases; Wlodawer *et al.*, 1989). In the context of homology modeling, the most difficult unsolved

Abbreviations used: PDB, Protein DataBank; RMSD, root-mean-square deviation; MC, Monte Carlo; SD, steepest descent.

0022–2836/97/140975–27 $25.00/0/mb960857

problem after sequence alignment, is the prediction of loop structures (Mosimann *et al*., 1995; Šali, 1995).

Loops are usually found on the surface of proteins, where they connect secondary structure elements. Three or four residue loops that connect two secondary structure elements and usually change the chain direction, are called β-turns. The β-turns can be categorized into a limited number of families (Sibanda *et al*., 1989; Mattos *et al*., 1994), and are sometimes regarded as a third class of secondary structure. For five-residue or longer loops, no simple classification has been possible (Sibanda *et al*., 1989). Leszczynski & Rose (1986) analyzed loops of 6 to 16 residues and reported their general properties, such as size, compactness, and residue composition. Because of the large degree of compactness they found for such loops, they suggested that they be regarded as a fourth class of secondary structure, with a large variability in conformation. A general classification of loops was made by Ring *et al*. (1992). They divided loops of 4 to 20 residues into three classes (strap, omega, and zeta loops) and a fourth, which is any combination of the three. They were not able to find a method for predicting the group of a particular loop sequence, based on hydrophobic periodicity or amino acid positional preferences. The only loop analysis that has considerable predictive power was developed for five of the six hypervariable antibody loops (Chothia & Lesk, 1987; Chothia *et al*., 1992).

The pioneering modeling study of Greer (1980) used a rather simple algorithm to insert loops from homologous proteins into the target protein. More recently, several methods have been introduced in attempts to obtain a more general approach for the prediction of loop conformations. Nevertheless, it is fair to say that, although the various loop construction methods have led to many interesting results, no reliable approach for longer loops (length > five residues) is available at this time (Mosimann *et al*., 1995). In most attempts, a number of (often many) loop conformations are generated that meet endpoint requirements, which are given by the residues at the termini of the ''known'' framework in the target structure. Because of the more or less fixed endpoints of a loop and its limited size, a large fraction of the available conformational space can be explored. This makes the problem much more tractable than the general protein folding problem (Creighton, 1992). All of the construction methods assume that the endpoints are known within certain tolerances. As we describe in this paper, the choice of endpoint criteria is not obvious and can have a significant effect on the results. The construction methods include more or less exhaustive dihedral angle searches (Bruccoleri & Karplus, 1985; Moult & James, 1986; Bruccoleri *et al*., 1988; Martin *et al*., 1989; Dudek & Scheraga, 1990; Borchert *et al*., 1994), minimum perturbation ''random tweak'' methods (Fine *et al*., 1986; Shenkin *et al*., 1987), molecular dynamics simulations (Bruccoleri & Karplus, 1990; Tanner *et al*., 1992), Monte Carlo searches with simulated annealing (Collura *et al*., 1993; Carlacci & Englander, 1993), dynamic programming algorithms (Vajda & DeLisi, 1990; Finkelstein & Reva, 1992), genetic algorithms (McGarrah & Judson, 1993; Ring & Cohen, 1994), bond scaling algorithms with relaxation (Zheng *et al*., 1993a,b; Rosenbach & Rosenfeld, 1995; Zheng & Kyle, 1996), and multicopy searches (Zheng *et al*., 1994). In all of these methods, the search for loop structures and their evaluation is done simultaneously. A complementary approach is based on a search of known protein structures in the Brookhaven Protein Database (PDB; Bernstein *et al*., 1977) for loop candidates and an evaluation of the resulting candidates. This method was introduced by Jones & Thirup (1986) to facilitate model building for crystallographic refinement. In that application the evaluation step is based on X-ray data for the target protein. The residues adjacent to the loop that is modeled (called ''stem residues'') are defined, and a search is made through known structures for segments whose stem residues can be superposed on those of the target. The segments are required to have the same length as the loop that is modeled (called ''target loop''), but there is no restriction on the sequence. This method of loop generation has the advantage of fast construction and a guarantee of obtaining physically reasonable mainchain conformations. Use of the PDB-based method requires a correlation between loop conformation and the properties of the stem residues. It was shown that for short loops (five residues), the search method can be very effective, as judged by the small deviation of the predicted loop conformation from the (known) target loop (Summers & Karplus, 1990). For longer loops no satisfactory conformations could be found in many cases (Summers & Karplus, 1990; Tramontano & Lesk, 1992); i.e. the superposition of the stem residues often resulted in loop conformations that did not match the target loop. A recent alternative PDB-based method constructs the loops using a ϕ_{i+1},ψ_i dimer database (Sudarsanam *et al*., 1995). For three short target loops (five residues), they obtained promising prediction results (RMSD < 0.54 Å) by selecting the constructed loops with the lowest RMSD of the stem residues with respect to the target loop. Thus, this method appears to be similar in efficacy to that of Summers & Karplus (1990). Another method that uses information present in the PDB applied a neural network to predict H3 loops of a set of antibodies (Reczko *et al*., 1995). The neural network was trained on a set of 1976 loops selected from the PDB on the basis of their similarity to known H3 loops. The input of the neural network consisted of the target loop sequence. For seven out of 20 target loops with lengths of five to nine residues, they obtained predicted loops with RMSDs of 2 Å or less, after loop superposition.

Given a set of conformations obtained by one of the methods described above, it is necessary to se-

lect those (the one) that are (is) likely to be the best loop candidate(s) for the target protein. There are two aspects in this selection. The first concerns the RMSD of the loops when they are superposed and the second concerns their orientation in space. Both aspects make use of energy functions of various types, including those limited to dihedral angle space (Dudek & Scheraga, 1990; Collura *et al*., 1993, 1994; Carlacci & Englander, 1993), and those in full Cartesian space (Summers & Karplus, 1990; Tanner *et al*., 1992; Smith & Honig, 1994). Solvent effects have been introduced by addition of explicit water molecules (Tanner *et al*., 1992), the use of simplified solvent models (Collura *et al*., 1994), and the calculation of the solvation free energy by use of the Poisson-Boltzmann equation and surface area-based terms (Smith & Honig, 1994). None of these evaluation methods have been applied to a wide enough range of loops to provide a meaningful test.

In the present study, we develop an approach for the PDB-based generation of a set of candidates for the target loop, for the evaluation of the resulting candidates, and for the optimization of the position of best candidates in the target protein. Correlations between the best candidate loop conformations and the stem residues are determined and the completeness of the available database for different target loop lengths is tested. For the analyses, we use 90 target loops in proteins whose structures are known at high resolution (2 Å or better). An additional 40 loops from antibodies are included to examine whether they should be considered as a special case because of the existence of canonical structures and of the success of prediction methods for them (Chothia & Lesk, 1987). The loops examined have lengths from four to 16 residues with primary emphasis on medium-sized loops in the range six to 12. By testing various approaches for loop/stem superposition and for energy function selection and optimization on such an extended set of loops, we are able to propose a method that is likely to provide useful loop predictions in a wide range of different proteins.

The following section describes the methods used for the selection of the structural database and the target loops, the generation of a set of template loops, and the calculation of several loop and stem properties, including structural similarities, sequence homologies, and interaction energies. The nature of the energy functions tested for loop evaluation and the methods for optimization of the selected structures are given. The results are presented, and a concluding discussion follows.

While this work was in progress, Fidelis *et al*. (1994) published a study of 11 target loops, ranging in length from four to six residues and considered one aspect of the loop prediction problem. They compared the performance of a systematic search in dihedral space with a search of the PDB for finding loop candidates and concluded that database search methods are useful only for loops of up to four residues. We come to the more optimistic conclusion that the present PDB is useful for up to nine residues, particularly if the candidate structures are subsequently evaluated evaluated and optimized. This is of considerable significance since exhaustive conformational search methods become very costly for longer loops.

Theory and Algorithms

Structural database selection

Loop backbone conformations were taken from the coordinates of known protein structures in the PDB. A subset of the available structures was selected, as follows. (1) Only structures of 2.0 Å resolution or better were included. (2) Homologous proteins from different sources were retained. (3) For identical proteins with diffeent ligands, oxidized *versus* reduced forms, and single site mutations, only the highest resolution structure was included. These selection criteria resulted in 173 different PDB files; the data can be obtained from the authors. Because homologous proteins are retained, the list of PDB files contains a large number of redundancies with respect to the overall protein fold. However, a similarity in overall protein fold does not imply a similarity in conformations of surface loops; an extreme example of this is provided by different hypervariable loop conformations for the very similar antibody framework. The variation in loop conformations for homologous proteins with the same framework structure raises a realistic modeling problem and provides additional data for analysis.

Each of the 173 protein structures was assigned to a single structural family. The definition of these structural families was obtained through structural superpositions of C^{α} atoms (Pascarella & Argos, 1992). Some structures in the selected set were not listed by Pascarella and Argos, and were assigned to one of the existing families where, if possible, the structural family classification of Šali & Overington (1994) was used. If this was not possible, an additional family was created. A total of 75 structural families resulted from this classification; the data can be obtained from the authors.

Selection of target loops

A set of 45 target loops was selected from the list reported by Leszczynski & Rose (1986). Only loops from structures with a resolution of 2.0 Å or better were included (Table 1). The length of the loops ranged from six to 16 residues. The residue numbers that define the target loops were taken from Leszcynski & Rose (1986). Because the exact choice of the first and last residue of a loop is not unambiguous, another 45 loops were defined by starting the previously defined target loops one residue later, and by ending one residue earlier in sequence. This resulted in a set of 90 target loops.

In addition, a set of 40 antibody hypervariable loops was selected, based on the loops that were

Table 1. List of protein loops, taken from the set of Leszczynski & Rose (1986), that were analyzed in this study

PDB	Protein		Loop no.	Residues	Loop no.	Residues
1pcy	Plastocyanin (poplar)	1.60 Å	loop1	41–56 (16)		
2act	Actinidin	1.70 Å	loop1	89–103 (15)	loop4	139–144 (6)
			loop2	141–156 (16)	loop5	198–205 (8)
			loop3	58–64 (7)		
2apr	Acid proteinase (*Rhizopus chinensis*)	1.80 Å	loop1	76–83 (8)	loop3	188–196 (9)
			loop2	128–137 (10)	loop4	202–210 (9)
2ptn	Trypsin	1.55 Å	loop1	69–80 (12)	loop2	142–152 (11)
3app	Penicillopepsin	1.80 Å	loop1	42–56 (15)	loop3	184–192 (9)
			loop2	129–137 (9)		
3est	Elastase (porcine)	1.65 Å	loop1	94–102 (11)	loop3	216–224 (11)
			loop2	142–152 (10)		
3grs	Glutathione reductase (human)	1.54 Å	loop1	83–89 (7)	loop4	139–147 (9)
			loop2	268–274 (7)	loop5	162–172 (11)
			loop3	300–307 (8)		
3sgb	Proteinase B (*Strep. griseus*)	1.80 Å	loop1	93–103 (7)	loop3	199–211 (9)
			loop2	118–124 (7)		
3tln	Thermolysin (native)	1.60 Å	loop1	55–70 (16)	loop5	91–97 (7)
			loop2	188–203 (16)	loop6	32–38 (7)
			loop3	221–233 (13)	loop7	44–53 (10)
			loop4	248–255 (8)	loop8	204–213 (10)
5cpa	Carboxypeptidase A (bovine)	1.54 Å	loop1	205–213 (9)	loop3	244–250 (7)
			loop2	231–237 (7)		
8abp	Arabinose binding protein	1.49 Å	loop1	93–99 (7)	loop3	203–208 (6)
			loop2	142–148 (7)	loop4	289–294 (6)
8gch	γ-Chymotrypsin	1.60 Å	loop1	71–79 (9)	loop2	95–102 (8)
9pap	Papain	1.65 Å	loop1	86–100 (15)	loop2	138–153 (16)

The set was chosen to have only high-resolution structures, and a variety of different loop lengths. The loop residue numbering is according to the PDB file. Loop lengths in residues are shown in parentheses.

analyzed by Tramontano & Lesk (1992). These loops are not limited to structures with resolutions of 2.0 Å or better. They were included to determine whether the existence of canonical structures for antibody hypervariable loops (Chothia & Lesk, 1987) influences the success rate of modeling these loops by database methods.

The total set of target loops consisted of 130 target loops with lengths between four and 16 residues. When antibodies were being studied, the structural database was expanded by an additional database of known antibody structures.

Loop search algorithm

The method described by Summers & Karplus (1990) was used to select loop backbones from the structural database as possible candidates for the target loop. Their approach was based on the work of Jones & Thirup (1986). A set of m target distances $t(i, j)$ were calculated between selected atoms i and j of the stem residues, separated by a n-residue gap in the target protein. The equivalent set of distances $r(P, S; k, l)$ for every n-residue segment S in each protein P in the database (template protein), is also calculated. The difference between the set of target distances and template distances (RMSD^{dist}) was defined as

$$\mathrm{RMSD}^{dist} = \left[\sum_{i\sim k, j\sim l} \frac{[t(i,j) - r(P,S;k,l)]^2}{m(2n+2)} \right]^{\frac{1}{2}} \quad (1)$$

The value of $2n + 2$ in the denominator is equal to the number of adjustable torsion angles (ϕ/ψ) that is present in a n-residue loop. If the RMSD^{dist} is lower than a specified cutoff value, the mainchain coordinates of the template loop are extracted from the PDB file. The loops are checked for having the correct geometry with respect to ϕ/ψ angles, and sp^2/sp^3 improper dihedral angles in both mainchain and sidechains, following Summers & Karplus (1990), and only loops satisfying certain criteria are included. To obtain enough antibody loops in the template loop set, the geometric criteria had to be relaxed. Most antibody loops in the PDB do not comply with the strict geometric criteria, which is a reflection of their lower resolution structures.

Here, we use the loop selection criteria in two ways. First, the method is used in comparison studies to find loops with backbone conformations that are similar to a structurally known target loop. In this case, the set of target distances are defined between atoms of the loop itself. The resulting distance comparisons are employed to evaluate correlations between loop conformation and stem conformation. Second, loop candidates are obtained for modeling an unknown loop structure (target loop) in an otherwise known protein structure. In this case, the selection is based on the RMSD^{dist} values of atoms of the stem residues with a specific length n of the loop.

Calculation of loop properties

To establish relationships between the loop conformation and other loop properties, we searched for loops that are similar to the target loops. For each

Table 2. List of antibody hypervariable loops, taken from the set of Tramontano & Lesk (1992), that were analyzed in this study

PDB	Protein		Loop no.	Residues	Loop no.	Residues
1mcp	Fab McPC603 (mouse)	2.70 Å	L1	L26–L38 (13)	H1	H26–H32 (7)
			L2	L55–L59 (5)	H2	H53–H58 (6)
			L3	L97–L102 (6)	H3	H102–H110 (9)
1rei	V_L REI Bence-Jones (human)	2.00 Å	L1	26–32 (7)	L3	91–96 (6)
			L2	49–53 (5)		
2fb4	Fab KOL (human)	1.90 Å	L1	L25–L31 (9)	H1	H26–H32 (7)
			L2	L48–L52 (5)	H2	H52–H57 (6)
			L3	L90–L97 (8)	H3	H100–H106 (15)
2fbj	Fab J539 (mouse)	1.95 Å	L1	L26–L31 (6)	H1	H26–H32 (7)
			L2	L48–L52 (5)	H2	H52–H57 (6)
			L3	L90–L95 (6)	H3	H100–H106 (7)
2hfl	Fab HyHel-5 (mouse)	2.54 Å	L1	L26–L31 (6)	H1	H26–H32 (7)
			L2	L48–L52 (5)	H2	H52–H57 (6)
			L3	L90–L94 (5)		
2rhe	V_L RHE Bence-Jones (human)	1.60 Å	L1	25–33 (9)	L3	92–99 (8)
			L2	50–54 (5)		
3fab	Fab′ NEW (human)	2.00 Å	L1	L25–L31 (10)	H1	H26–H32 (7)
					H2	H52–H56 (5)
			L3	L90–L95 (6)	H3	H98–H105 (8)
4fab	Fab 4-4-20 (mouse)	2.70 Å	L1	L26–L37 (12)	H1	H26–H32 (7)
			L2	L54–L58 (5)	H2	H53–H58 (6)
			L3	L96–L101 (6)	H3	H101–H106 (6)

The loop residues numbering is according to the PDB file, where L and H refer to light and heavy chains, respectively.

target loop between 100 and 500 most similar template loops from the structural database were collected. This was done with the loop search algorithm by using all C^{α} atoms in the target loop to define the interatomic distances. To arrive at the required set of similar template loops, the $RMSD^{dist}$ cutoff had to be chosen separately for each target loop because different loops have different numbers of similar loops in the structural database. The $RMSD^{dist}$ cutoffs ranged from 0.1 Å for loops of four residues to 1.75 Å for loops of 16 residues.

A template loop that successfully passes the $RMSD^{dist}$ cutoff and geometric criteria, was processed further. The loop C^{α} atoms were superposed on the known loop structure of the target protein, using the method of Kabsch (1976). After superposition of the loop, the relative conformations of the stems of the loop were compared. The Cartesian root-mean-square deviation (RMSD) of the loop mainchain atoms (N, C^{α}, C) was an indication of the similarity between the target loop and template loop, and was referred to as ''loop quality''. The similarities and differences of the stem residues after the superposition of the loop were examined to determine the relation between the stem residues and the loop conformation. The following comparisons were made. (1) RMSD of stem backbone atoms (N, C^{α}, C), where the stem length was varied from one to three residues (on each side of the loop). These values are called stem1, stem2, and stem3, respectively. (2) The $RMSD^{dist}$ of distances between C^{α} atoms, or between C^{α} and C^{β} atoms, of the stem residues. The length of the stem was again varied from one to three residues. These criteria are called CA1 (based on one distance), CA3 (15 distances), CACB1 (four distances), and CACB3 (60 distances).

The various fits were calculated by the linear least squares method and by the robust linear fitting algorithm MEDFIT (Press *et al.*, 1986). Since the results of the two methods were very similar, only the linear least squares fits are shown here.

The stem1–3 values depend on the superposition of the template loop and the target loop. Consequently, they cannot be used for modeling unknown loops and served only to examine correlations between the loop and the stem. The $RMSD^{dist}$ values are independent of a superposition of the template loop and target loop; i.e. they depend only on the stem residues. The correlations between the $RMSD^{dist}$ values and the loop quality were used to develop parameter values for obtaining a certain loop quality from a stem search.

The stem1–3 and $RMSD^{dist}$ values were correlated with the loop qualities for every individual target loop. A linear equation of the form:

$$P = a + b \times Q \tag{2}$$

was used with P the stem property measure, and Q is the loop quality (in Å). The significance of the correlations was measured by calculating the Pearson's linear correlation coefficient R. Significant correlation would indicate that conformationally similar loops are likely to have similar stem properties. The inverse (namely, that similar stem properties lead to similar loops), which is of primary interest, would also be true.

To use the correlations of the different stem properties to extract stem parameters for predictive loop searches, the best three template loops that were found for each target loop were considered. To avoid including corresponding loops of related proteins, we only retained template loops from unrelated proteins here. From the best three template

loops, we extracted the most restrictive (=smallest) $RMSD^{dist}$ value, i.e. the one that would result in the smallest total number of template loops from the stem search. With these stem parameters we should be able to find at least one of the three best possible loops in the database.

After collecting these parameter values for all individual target loops, they were arranged by target loop length. For every loop length, an average and a worst-case (least restrictive parameter value for that length) stem parameter was extracted. It is not possible to guarantee that the chosen stem parameter values are valid outside the specific set of target loops. However, we believe that the target set was large enough and contained a wide enough variety of loop conformations to produce reliable stem parameters; in particular, for loop lengths of five to nine, there were at least 12 target loops (including the antibody loops) in each set.

Use of energy functions

In a predictive loop modeling study, the various loop candidates must be evaluated after they have been selected from a database. This can be done most directly by an energy-based criterion. In addition, it will usually be necessary to improve the initial placement of the loop candidates by moving them rigidly or allowing them to relax. This also requires an energy related function, though not necessarily the same one as used for the initial evaluations; e.g. as in some protein-folding studies, a relatively low resolution effective energy function was used in the first step and a more complete energy function in the second (Skolnick *et al.*, 1993). A prerequisite for a good energy function is that it has a low energy value for the actual conformation of the loop, relative to the other loop candidates selected by the database search procedure. We tested the performance of several energy functions by calculating the non-bonded interaction energies between a set of selected template loops and the protein framework. As described above, each of the template loops was overlapped with the target loop so as to minimize the RMSD for the backbone atoms of the two loops.

In the interaction energy calculations, the first stem residue on both sides of the loop in the target protein framework was not included, since strongly repulsive overlaps with the template loop may result for these residues, because the template loop was fitted to the target loop without optimization, so that the loop might not fit well with respect to the stem residues of the protein framework.

Since the target loop rarely consists of the same amino acids as the template loops, a set of energy calculations were done including only the loop mainchain and C^{β} atoms, whose positions are determined by the mainchain. To include sidechains in the energy evaluation, we used several reduced sidechain models. The Levitt sidechain model has the advantage of easy construction and fast energy calculation (Levitt, 1976). In this model, the sidechains are represented by a single sphere that is placed at the position of the average centroid of the sidechain; the latter is determined from mainchain coordinates. The total interaction energy of the template loop with the rest of the target protein structure is assumed to consist of two parts: atom-based interactions between all mainchain or C^{β} atoms, and reduced sidechain interactions between sidechains. The atom-based interaction was modeled by a modified non-bonded 9-6 potential, with values for the well-depth ε and equilibrium distance σ, from the param19 parameter set of the CHARMM program (Brooks *et al.*, 1983). The calculation was simplified by using an average value of ε and σ-values for all carbon (0.1077; 2.193), nitrogen (0.2384; 1.600) and oxygen (0.3208; 1.600) atoms. The modified pairwise 9-6 potential was:

$$E = \varepsilon\left[\frac{2\sigma^9}{r^9} - \frac{3\sigma^6}{r^6}\right] \quad \text{for } E < 10 \text{ kcal/mol} \qquad (3)$$

and $E = 10$ kcal/mol if the calculated value was larger. The limit on the maximum non-bonded interaction energy allows atoms to move ''through each other'', but the modified energy function still shows which template loops have bad contacts. The Levitt sidechain energy function, which used an 8-6 potential, was modified in the same way. The parameters used in the Levitt function were taken from the original publication (Levitt, 1976). In addition to the Levitt sidechain model, we used the reduced sidechain models of Casari & Sippl (1992), and of Gerber (1992). These models also represent the sidechains by single spheres. The Casari/Sippl model puts the sidechain at the C^{β} position, and the Gerber models puts it at the C^{α} position.

Atom-based sidechain models were also used. The backbone-dependent sidechain rotamer library of Dunbrack & Karplus (1993) was employed. The sidechains were built by using the most probable χ_1/χ_2 pair for each residue as listed in the library. Other χ-angles of the sidechains were assumed to be 180°. The sidechains in the remainder of the protein were taken to have the position of the target structure. In the interaction energy calculations, we varied the number of atoms of the template loop sidechains (up to C^{β}, up to C^{γ}, etc.), so as to determine the optimal extent to which predicted sidechain atoms should be included. Also we used three different electrostatic functions: (1) no electrostatics, (2) a dielectric constant (ε) of 1 (vacuum), and (3) a sigmoidal dielectric function, with ε proportional to 1.4 times the interatomic distance and a switching function operating between 0 and 8 Å distance (A. Blondel, personal communication).

Following the calculations of the interaction energy between template loops and the target protein, we determine the percentage of template loops with a higher interaction energy than the loop of the crystal structure. Ideally, this should be close to 100%. In addition we evaluated the relation between the three reduced sidechain models and the CHARMM

all-atom model for the calculation of interaction energies. For this purpose, the interaction energies between the individual sidechains in the target loop and all non-loop sidechains in the target protein were calculated with CHARMM (param19 parameter set). In these calculations, the target loop, as well as the rest of the protein, had the crystal structure coordinates. The results for the different reduced sidechain models were compared on a residue-by-residue basis and expressed in the form of cross-correlation coefficients. Since in this last calculation known loop conformations are used, we were able to add one additional reduced sidechain model (Crippen & Viswanadhan, 1985) to gauge its performance with respect to the other three models. In the Crippen/Viswanadhan model, the spherical sidechain is placed at varying positions, which depend on the residue type. The model is less suitable for modeling unknown loops, because for some residue types sidechain atom positions beyond C^{β} have to be known to place the model sidechain.

All energy calculations were done on the initial set of 45 full-length target loops (Table 1) and on the 40 antibody loops (Table 2).

Use of stem searches for loop selection

The objective of the loop search algorithm is to be able to use it in predictive loop modeling studies. In contrast to the searches described above, we here assume that the target loop conformations are unknown. Atoms from the stem residues are selected to define the target distance parameters ($RMSD^{dist}$). Both average and worst-case values are used for the $RMSD^{dist}$ parameters. In each case, the stem residues of the template loops are superposed on the stems of the target loop, including only the atoms that define the $RMSD^{dist}$ parameter used in the stem search. For CA1, only two atoms are defined and the C^{β} atoms of the first stem residues on either side of the loop were added to enable a unique fit to be made. We also tried certain additional superposition schemes, using all backbone atoms (N, C^{α}, C) of one to three stem residues on each side of the loop.

By choosing the $RMSD^{dist}$ parameter and superposition scheme that yield the best initial placement of the template loops, the subsequent task of optimizing the loop orientation will be facilitated. These searches, which are aimed to obtain the lowest number of template loops and the best possible initial orientations, are done for all 130 template loops.

Above, we described energy functions that were used for identifying and/or optimizing template loop conformations. The ability of the modified 9-6 potential (equation (3)) to find the best possible template loop orientation was gauged by means of rigid body Monte Carlo (MC) optimization. After initial placement of a given template loop by stem superposition, it was rigidly reoriented *via* 50,000 MC steps at 5000 K. After 50,000 MC steps, the lowest energy orientation for the template loop was saved. Moves were accepted or rejected according to the Metropolis criterion (Metropolis *et al.*, 1953). Random rotations were chosen from a homogeneous distribution in the Euler angles ϕ (step taken from an interval: -30°, 30°), $\cos\theta$ (-0.1, 0.1), and ψ (-20°, 20°) (Allen & Tildesley, 1987). Translations (x, y, and z) were taken from a homogeneous distribution between -0.5 Å and 0.5 Å. Step sizes were divided by 2 if the acceptance ratio of a 5000 step window fell below 25%.

Since the modified 9-6 potential contains only nonbonded terms, additional bonded terms are needed to keep the loop termini close to the anchor residues in the target protein. This was done by introducing pseudo-bonds and pseudo-angles between C^{α} atoms of the loop and the first anchor residues of the target protein. The pseudo-bonds were between the first and last C^{α} in the loop and their respective neighbor C^{α}s in the target protein. In a similar fashion, pseudo-angles were defined. This yields two pseudo-bonds and four pseudo-angles. The bond angle potentials were represented by a harmonic term, with minima and force constants derived from a statistical analysis of 6499 coil residues in a 62-protein database (J.-M. Chandonia, unpublished results). For C^{α}–C^{α} pseudo-bonds, an average distance of 3.81 Å was found, with a standard deviation σ of 0.43 Å. This translates into a harmonic bond potential with the minimum at 3.81 Å and a force constant of 3.22 kcal/mol Å^2, identifying σ^2 with k_BT/K (k_B = Boltzmann constant, T = temperature, K = force constant). Analogously, the C^{α}–C^{α}–C^{α} pseudo-angle harmonic potential has a minimum at 109.9° and a force constant of 6.21 kcal/mol rad^2.

Optimization of the template loop orientation was also done with the full CHARMM potential function. No additional energy terms are needed here, since this potential function contains bond, angle, dihedral, and improper dihedral terms (Brooks *et al.*, 1983). Optimizations were done by minimization and by simulated annealing. During the optimizations, the rest of the protein was fixed. A strong harmonic dihedral constraint (500 kcal/mol · rad^2) was placed on all internal loop dihedral angles to emulate rigid body motion of the loop. No constraints were imposed on the first and last loop residue and on the stem residues neighboring the loop. This allowed these residues to adjust the bonds, angles, and dihedral angles that corresponded to a bad geometry. After the initial placement of the template loop, 2000 steps of Steepest Descent (SD) minimization were done, and the interaction energy between the loop and the target protein was calculated. Since the CHARMM potential function does not have an energy maximum like the modified 9-6 potential (equation (3)), atoms were unable to pass each other at close distance and a template loop could be trapped in an unfavorable orientation. To avoid this, we allowed all template loops with interaction energies of 1000 kcal/mol or more to search for a better orien-

tation by rotating them in 12 steps of 30°, around the C^{α}–C^{α} axis between the two residues bordering the loop. The orientation with the lowest interaction energy was retained. After this extensive minimization was done, consisting of another 2000 SD steps and 3000 Adopted-Basis Newton-Raphson (ABNR) steps (Brooks *et al.*, 1983).

Rigid body MC optimization with the modified 9-6 potential, and minimization and MD annealing with the CHARMM potential were applied to five different target loops: 3sgb loop1 (five residues), 3app loop3 (nine), 2act loop1 (15), 1rei L3 (six), and 4fab H3 (six). For 14 additional target loops from four to nine residues in length, we used the best protocol (i.e. minimization of the CHARMM energy with dihedral constraints) to optimize the selected set of template loops.

Results

The Results section is subdivided into three parts. In the first, we analyze the loop searches based on the target loop mainchain conformations. We show the correlations that were found between the template loop quality and various stem properties. We use the correlation results to evaluate parameter values that can be used in stem searches so as to include the best loop candidates in the database. In the second, we investigate the ability of various energy functions to distinguish the correct loop from the incorrect ones; here we are concerned with the RMSD between target and template loops after the best superposition. In the third we compare the results of stem searches with the different parameters and determine the best one for predictive purposes. To test the possibility of refining the initial orientations, we use the energy functions to optimize the spatial position of the template loops in the protein for 19 target loops in the final section.

Correlations between loop quality and stem properties

We first examined the correlation between loop quality and different stem lengths between one and three residues (see below). Table 3 lists the b and R^2 values of the linear fits (equation (2)) for stem1, stem2 and stem3, averaged for each loop length. The a values, which represent the intersection of the fitted line with the y-axis, are highly variable and range between −5.0 and 5.0. The linear correlations of the loop quality for all three stem lengths are very similar; i.e. there is no indication that use of a longer stem gives better correlation with loop quality. The fact that all b coefficients (with one exception) have positive values confirms that more similar loops have more similar stem residues. The R^2 values show that the correlation between the loop quality and the stem parameters is strongest for shorter loops and decreases significantly with loop length. This agrees with the fact that shorter loops, which have fewer degrees of freedom, are more restricted by their stem conformations (Summers & Karplus, 1990). The antibody loops display no better correlations between loop quality and stem properties, as can be judged from the R^2 values of corresponding loop lengths (Table 3).

To provide some details concerning the variability of the loop/stem correlation, Figure 1 shows the

Table 3. Results for the linear fits of the stem1, stem2, and stem3 parameters to the loop quality

		stem1				stem2				stem3			
n	No.	$\langle b \rangle$	±	$\langle R^2 \rangle$	±	$\langle b \rangle$	±	$\langle R^2 \rangle$	±	$\langle b \rangle$	±	$\langle R^2 \rangle$	±
4	3	7.06	0.31	0.85	0.02	7.66	0.95	0.63	0.08	8.47	1.66	0.49	0.12
5	11	5.30	2.40	0.58	0.20	5.90	2.98	0.47	0.17	6.27	3.45	0.38	0.18
6	8	3.55	1.01	0.50	0.24	3.57	1.27	0.39	0.22	3.50	1.65	0.31	0.19
7	19	3.33	1.50	0.39	0.27	3.60	1.73	0.32	0.21	3.76	2.05	0.24	0.16
8	9	2.33	1.34	0.31	0.20	2.38	1.74	0.28	0.17	2.42	1.99	0.25	0.17
9	12	3.11	0.98	0.39	0.15	3.30	1.16	0.31	0.14	3.41	1.29	0.26	0.13
10	5	2.54	0.86	0.28	0.15	2.69	1.17	0.23	0.13	2.95	1.45	0.20	0.11
11	5	3.04	0.72	0.33	0.09	3.03	0.92	0.25	0.09	3.03	1.34	0.21	0.11
12	1	4.05		0.59		4.65		0.54		5.35		0.49	
13	4	2.39	1.23	0.19	0.13	2.45	1.34	0.15	0.10	2.45	1.40	0.13	0.09
14	5	2.55	0.36	0.26	0.08	2.57	0.42	0.21	0.10	2.39	0.60	0.17	0.11
15	3	3.81	0.21	0.36	0.06	3.88	0.45	0.33	0.04	3.62	0.64	0.24	0.07
16	5	2.20	0.74	0.19	0.14	2.03	0.85	0.14	0.14	1.99	0.82	0.12	0.13
5	9	5.18	0.27	0.80	0.20	6.70	0.35	0.80	0.19	7.39	0.49	0.71	0.20
6	13	5.49	2.33	0.49	0.22	6.48	3.05	0.43	0.23	7.61	3.70	0.40	0.21
7	8	4.57	1.06	0.46	0.11	5.90	1.58	0.40	0.10	7.18	2.01	0.39	0.09
8	3	2.26	1.03	0.33	0.20	2.50	1.18	0.23	0.14	3.04	1.66	0.20	0.13
9	3	4.79	1.38	0.54	0.02	5.78	2.04	0.47	0.07	6.72	2.61	0.43	0.09
10	1	5.23		0.77		6.05		0.64		6.95		0.59	
12	1	1.10		0.06		0.34		0.00		−0.47		0.00	
13	1	2.91		0.28		3.82		0.22		5.16		0.23	
15	1	1.63		0.17		2.49		0.21		3.00		0.18	

The columns labeled n and No. contain the number of residues in the target loop, and the number of target loops with that length, respectively. The top part of the Table, with n ranging from 4 to 16, corresponds to all non-antibody loops, the bottom part (n from 5 to 15) to the antibody loops. For each parameter, the average ($\langle\ \rangle$) b and R^2 and their respective RMSD values (±) are listed.

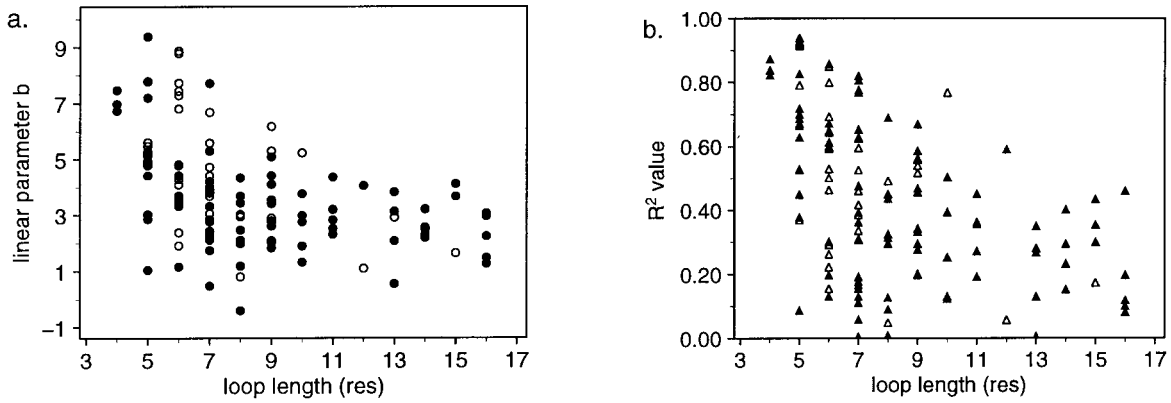

Figure 1. a, Linear parameter *b* (equation (2)) of the correlation between template loop quality (RMSD in Å) and the stem1 parameter. Every point represents the complete search for a particular target loop. Antibody loops are indicated by open symbols. b, Squares of Pearson's correlation coefficient (R^2), indicating the significance of the linear correlation.

results for the linear fit of stem1 to the loop quality for all 130 target loops, arranged by loop length; open symbols indicate antibody loops. The *b* coefficients vary between approximately 0 and 9 and the R^2 values displayed a large variation. As examples (Figure 2), the template loops found for target loop L2 of the 2fbj antibody (length five residues) clearly have a much better correlation between the loop quality and the stem1 parameter than 3grs loop1 (seven residues). However, in the case of the L2 loop of the 2fbj antibody, the set of most similar template loops are all from related proteins (i.e. other antibodies).

To illustrate the important differences between the results of template loop searches the C^{α} traces of the ten highest quality template loops are shown for two different target loops (Figures 3 and 4). In Figure 3, we show target loop 3fab L3, together with the ten best unrelated template loops. In this case, the spread in stem3 values is 1.58 to 4.31 Å. For 3grs loop3 the RMSD of the loop mainchain itself is in a comparable range (0.27 to 0.40 Å for 3grs loop3 *versus* 0.34 to 0.39 Å 3fab L3), but the range of stem3 values is much larger (5.34 to 10.39 Å; Figure 4). Together with the highly variable correlations discussed before, these figures indicate that the success of loop modeling by database screening methods is likely to depend on the particular loop being modeled. The possibility of selecting the best of the resulting loops by means of various energy functions is discussed below.

Correlations between the loop quality and the $RMSD^{dist}$ parameters CA1, CA3, CACB1, CACB3 (see earlier) are important, since these parameters are independent of the superposition of template loops on the target loop, and can therefore be used in predictive stem searches. Table 6 lists the results of the linear fits for these parameters. As for stem correlation, a large spread in the values exists within the set of target loops. Overall, the R^2 values are somewhat lower than those for the stem correlations. No significant difference is found between the parameters CA1, CA3, CACB1, and CACB3. On the basis of these results it makes no difference which $RMSD^{dist}$ is used in a stem search

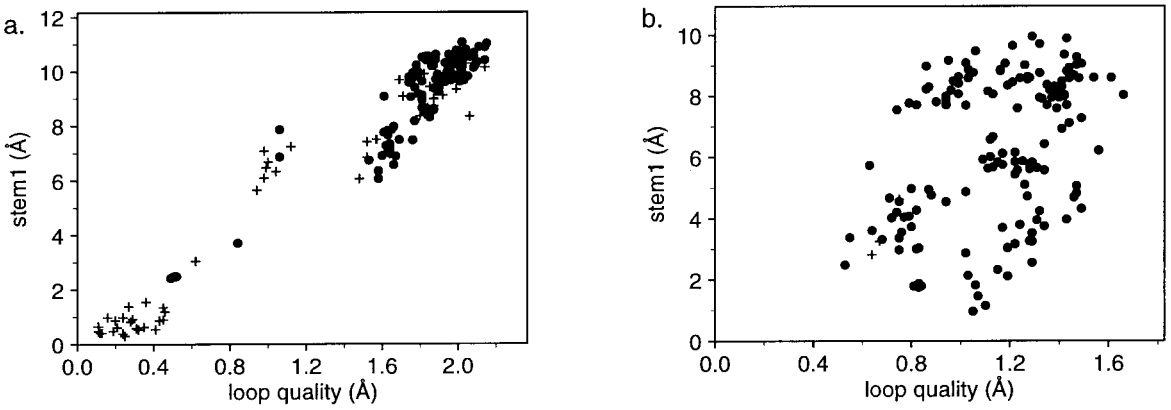

Figure 2. a, All template loops for the 2fbj L2 loop (five residues) target loop search. The stem1 value of every template loop is plotted *versus* its loop quality. Circles represent template loops from unrelated proteins, cross symbols are loops from related proteins. b, As a, but for the 3grs loop1 (seven residues) loop search.

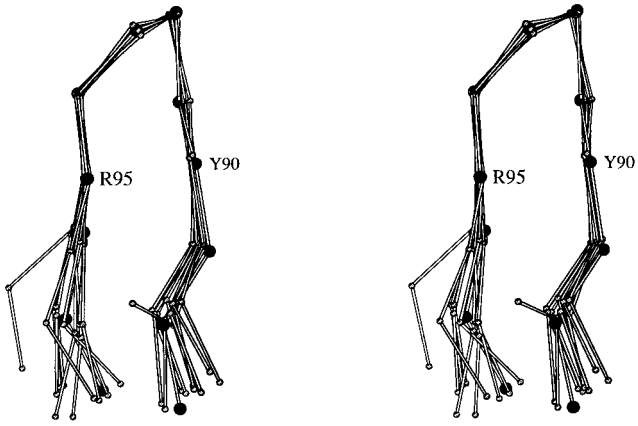

Figure 3. Wall-eyed stereo drawing of a short loop (six residues) with a relatively low spread in stem3 parameter. The ten best template loops (C^{α}-traces only) found for 3fab L3 (residues 90 to 95) are drawn with small grey atoms, and the target loop with large black atoms. The C^{α} atoms of the first and last residue in the target loop are labeled. The mainchain RMSD of these template loops ranges from 0.34 Å to 0.39 Å.

when modeling an unknown loop. Overall trends once more indicate stronger correlations for shorter loops, and a *b* coefficient that is positive except for two sets of antibody loops. The *a* values are between −1.0 and 1.0.

These results indicate, together with those of stem1-3, that the degree of conformations similarity of the stem residues is only weakly correlated with the similarity of the loop conformation. This is to be expected for long loops, which have more conformational space available, and are therefore less dependent on the stem residue positions than short loops. Although the correlations are higher for short loops (as measured by the R^2 values), stem similarity alone is still not a reliable filter for extracting high quality loops. Generally, the template loop quality range was much smaller for short loops, because there are many more similar template loops present and we limited ourselves to extracting only a few hundred. This small loop quality range in turn leads to the relatively high *b* values for the short loops (Tables 3 and 4). The often low R^2 values of the loop-stem correlations mean that although there is a statistically relevant correlation, a similarity in the stem region does not imply a similarity of the loop conformation for any particular template loop. Most often the template loops with the most similar stems do not have the highest loop quality, and *vice versa*.

Antibody loops

Hypervariable loops of antibodies have been extensively modeled by different methods (Fine *et al.*, 1986; Bruccoleri *et al.*, 1988; Tramontano & Lesk, 1992; Tanner *et al.*, 1992; Zheng *et al.*, 1993b). The existence of canonical structures for these antibody loops (other than H3) greatly facilitates the modeling if PDB-based methods are employed. Using the concept of canonical structures, Chothia *et al.* (1989) were able to get very good loop predictions for the internal structure of 19 antibody loops (i.e. 15 out of 19 mainchains with RMSD < 1 Å, after superposition of loops). The orientations of these loops after superposition of the framework residues was significantly worse. Differences in C^{α} positions between predicted and observed structures ranged from 0.4 Å to 3.5 Å.

For all 40 antibody loops, we plotted the loop qualities for the ten best template loops found (Figure 5). The loops are sorted with respect to loop name (L1, L2, etc.) on the *x*-axis. The + symbols indicate the results for searches with the extended PDB database, including all available antibody structures. The circles represent the ten best loops from unrelated proteins only. A + symbol in a circle represents an unrelated template loop that belongs to one of the ten best loops even if the extended database is used. Especially for the L1, L2, and H1 loops, the best template loops are clearly found in other antibody structures, virtually always in canonically related loops. There are a few exceptions, e.g. 1mcp L1 (13 residues), and 2hfl L1 (six residues), for which template loops from unrelated proteins are equally good. For the two other loops with canonical structures, L3 and H2, unrelated proteins deliver the majority of the ten best template loops, in about

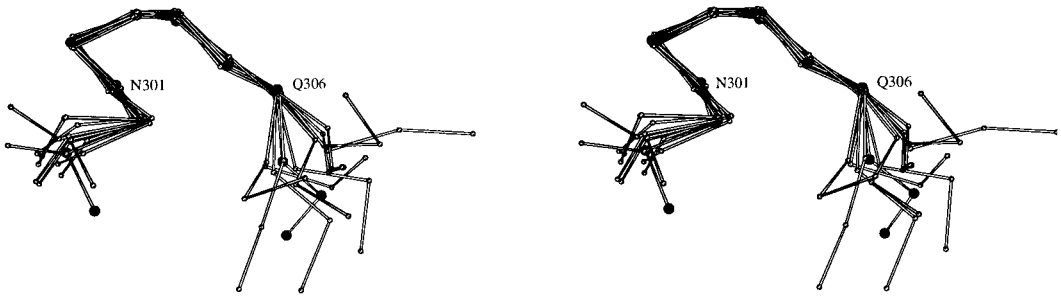

Figure 4. Example of a short loop with a high stem3 parameter spread (3grs loop 3, residue 301 to 306). The mainchain RMSD of these template loops ranges from 0.27 Å to 0.40 Å.

Table 4. Results for the linear fits of the stem parameters CA1, CA3, CACB1, and CACB3, to the loop quality

		CA1			CA3			CACB1			CACB3		
n	No.	$\langle b \rangle$	$\langle R^2 \rangle$	±	$\langle b \rangle$	$\langle R^2 \rangle$	±	$\langle b \rangle$	$\langle R^2 \rangle$	±	$\langle b \rangle$	$\langle R^2 \rangle$	±
4	3	2.10	0.49	0.26	4.43	0.32	0.21	2.34	0.47	0.24	4.68	0.34	0.21
5	11	0.91	0.31	0.25	1.55	0.17	0.18	0.07	0.30	0.24	1.58	0.17	0.18
6	8	0.52	0.29	0.21	1.31	0.17	0.12	0.55	0.30	0.17	1.35	0.18	0.11
7	19	0.14	0.11	0.16	0.45	0.11	0.13	0.13	0.11	0.13	0.44	0.11	0.13
8	9	0.32	0.17	0.17	0.80	0.11	0.13	0.32	0.17	0.16	0.81	0.12	0.14
9	12	0.29	0.11	0.07	0.54	0.05	0.04	0.29	0.11	0.08	0.57	0.06	0.05
10	5	0.26	0.10	0.07	0.69	0.06	0.06	0.23	0.09	0.06	0.64	0.06	0.05
11	5	0.31	0.09	0.12	0.83	0.07	0.08	0.30	0.08	0.10	0.80	0.07	0.08
12	1	0.36	0.14		1.11	0.12		0.39	0.14		1.10	0.12	
13	4	0.28	0.04	0.03	0.79	0.05	0.03	0.28	0.04	0.02	0.77	0.05	0.03
14	5	0.22	0.04	0.03	0.79	0.07	0.04	0.23	0.04	0.03	0.77	0.07	0.04
15	3	0.24	0.06	0.05	0.47	0.06	0.06	0.26	0.05	0.04	0.45	0.06	0.06
16	5	0.09	0.03	0.05	0.23	0.02	0.02	0.07	0.02	0.04	0.21	0.02	0.02
5	9	0.74	0.49	0.18	2.30	0.39	0.16	0.66	0.42	0.15	2.28	0.41	0.16
6	13	0.78	0.23	0.15	3.59	0.22	0.14	0.85	0.22	0.16	3.59	0.23	0.15
7	8	0.15	0.05	0.06	0.54	0.07	0.07	0.25	0.06	0.06	0.66	0.07	0.07
8	3	0.15	0.04	0.01	0.89	0.08	0.00	0.12	0.03	0.01	0.87	0.08	0.00
9	3	0.33	0.10	0.00	1.11	0.06	0.04	0.42	0.10	0.07	1.18	0.07	0.05
10	1	0.63	0.21		1.70	0.18		0.76	0.23		1.89	0.21	
12	1	0.25	0.04		0.82	0.07		0.31	0.04		1.00	0.10	
13	1	−0.11	0.01		0.34	0.01		−0.09	0.00		0.45	0.02	
15	1	0.15	0.07		0.69	0.06		0.16	0.10		0.67	0.07	

Columns are set up as in Table 3. RMSD values (±) correspond to the R^2 averages.

50% of the cases (see e.g. 2fb4 L3, 3fab H2, 4fab H2). For the H3 loop, which has not been found to have canonical structures, there is almost no improvement in the loop quality of the template loops when the extended database is used.

Determination of stem search parameter

To visualize the overall trends and completeness of the database, we plotted, for all target loops, the loop quality of the ten best-fitting template loops (Figure 6); only template loops from unrelated proteins were included. Between target loops of the same length, there are large variations in the best-fitting template loops. For instance, for 3app loop2 (nine residues), none of the template loops is better than 1.2 Å, whereas for 3grs loop4 (nine residues), there are ten unrelated template loops with a loop quality of 0.4 Å or better. In this case, the result is due to the fact that a large part of the 3grs loop has an anti-parallel β-strand character, whereas the 3app loop has an overall coil-like conformation. We found that for loops of up to and including eight residues, it makes little difference whether proteins related to the target protein are included in the database or not; i.e. the unrelated proteins

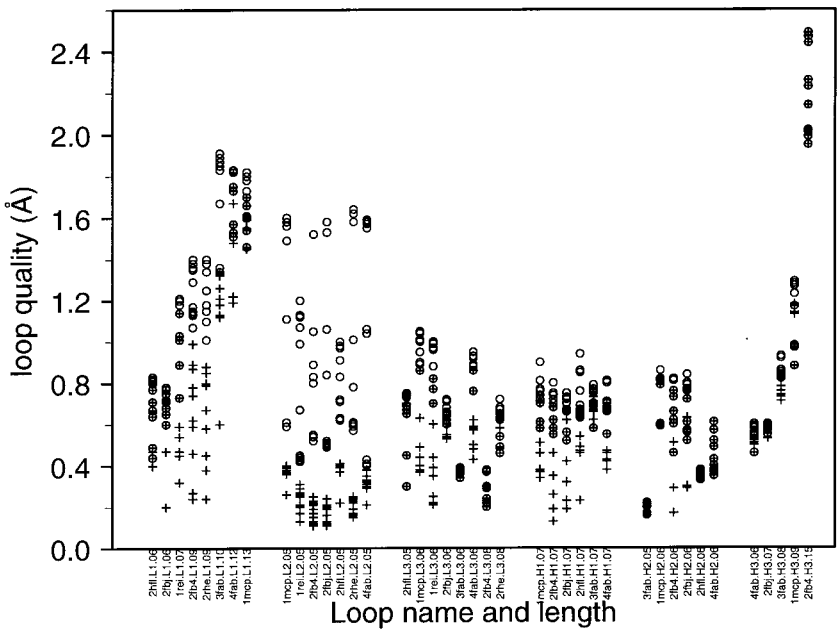

Figure 5. For all antibody loop searches, the best template loops with highest quality are shown. Cross symbols represent template loops from a database with related and unrelated proteins, and circles from a database of unrelated proteins only. The names and lengths of the different target loops are shown on the *X*-ordinate.

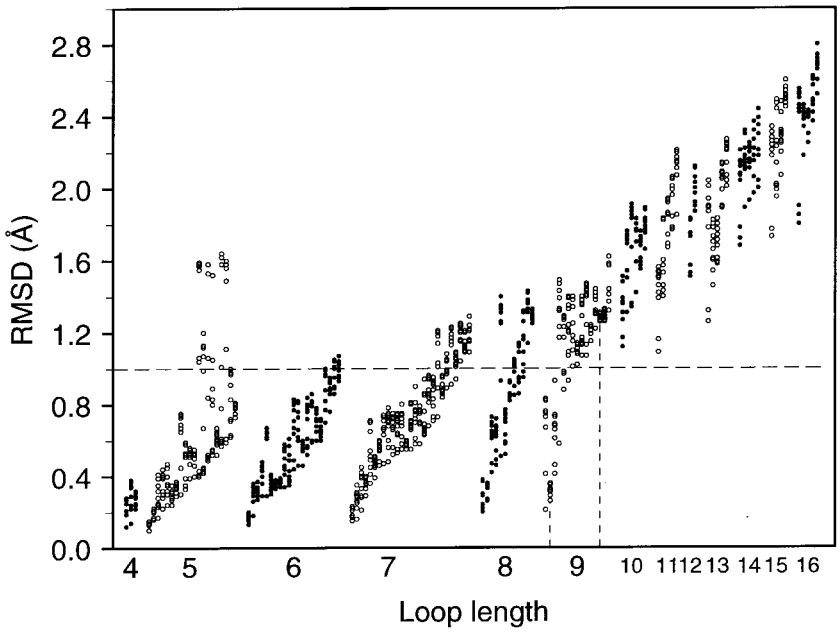

Figure 6. For all 130 target loop searches, the ten best template loops of proteins unrelated to the target protein are shown. Results are sorted according to target loop length. Every column represents one target loop. Open and filled circles are alternated between different loop lengths for clarity. All template loops below the horizontal line have an RMSD to the target loop of less than 1 Å. The results for two particular target loops are indicated by vertical broken lines: 3grs loop4 (left) and 3app loop2 (right).

contain loops that are as similar as those found in related proteins. Longer loops show a significant difference though it is still not very large; e.g. for loops of length 15, the worst case values are 2.07 Å and 2.45 Å for all cases and only unrelated proteins, respectively.

Figure 6 shows that one cannot expect to find template loops with a loop quality better than 1.0 Å RMSD for loops of more than nine residues in the available database. Also, there are large variations in best-fitting template loops for target loops of the same length. In the set of 11 loops reported by Fidelis *et al*. (1994), they find average loop qualities of 0.49 Å, 0.58 Å, and 0.59 Å, for loops of seven, eight, and nine residues, respectively. Corresponding worst-case (highest value for particular loop length) loop qualities were 0.84 Å, 0.88 Å, and 0.59 Å, respectively. Although they used the RMSD based on C^{α}-atoms only and we include all mainchain atoms, their average loop qualities are very similar to the ones reported here. However, we find higher worst-case values, most probably because of the larger number of target loops that was analyzed. This illustrates the major weakness of the PDB-based search method, which is that it is incomplete for longer loops.

From the results of the target loop searches we extracted average and worst-case parameter values (for each loop length) of CA1, CA3, CACB1, and CACB3, for use in stem searches; i.e. the $RMSD^{dist}$ values that must be allowed to obtain at least one of the three best possible template loops from unrelated proteins. The resulting $RMSD^{dist}$ search parameters display large differences between the average and worst-case values (Table 5). In the case of an unknown loop, a worst-case scenario has to be assumed, to make sure that one of the three best possibilities will be in the template loop set. The use of worst-case values means that many more template loops have to be processed, compared to a search with average parameter values. If average values are used, one of the three best loops would not be obtained in 37% of the cases, with no apparent dependence on loop length. There does not seem to be a significant increase in any of the four search parameter values with increasing loop length, especially for the average values. This suggests that the division by $2n$ (where n = number of residues in the loop) in the calculation of the $RMSD^{dist}$ values (equation (1)) removes, in an approximate sense, the dependency on loop length. There are certain loop lengths where one parameter gives significantly tighter limits, particularly for the worst case values. This is true for loops of length 7 and 11 for which CACB1 is better and for loops of length 13 and 14 for which CACB3 is better. This is most likely a reflection of the limited number of target loops. Instead of restricting the parameters to finding one of the best three loops, more relaxed criteria could have been specified. The resulting parameters would be more restrictive in the stem searches when modeling unknown loops. However, this is not likely to be useful because many target loops have only a few (<five) good-fitting template loops, while the remaining template loops in the top ten are much worse (Figure 6).

Energy based loop evaluation

It has been found that when template loops are extracted from the PDB by a stem search and fitted on the corresponding stems of the target protein, their initial orientations are not necessarily optimal (Tramontano & Lesk, 1992, Fidelis *et al*., 1994). This in accord with the weak correlation between the loop and stem conformations (e.g. see Figure 1 and Table 3). Energy functions can be used to evaluate

Table 5. Worst-case (worst) and average (aver.) $RMSD^{dist}$ values in Å, sorted by loop length n (in residues)

	CA1		CA3		CACB1		CACB3	
n	Worst	Aver.	Worst	Aver.	Worst	Aver.	Worst	Aver.
4	0.26	0.16	0.48	0.31	0.21	0.13	0.26	0.18
5	0.53	0.13	0.91	0.41	0.29	0.10	0.47	0.24
6	0.52	0.17	0.83	0.31	0.35	0.13	0.35	0.19
7	0.67	0.18	1.43	0.45	0.30	0.12	0.73	0.25
8	0.88	0.22	0.91	0.40	0.43	0.15	0.46	0.23
9	0.74	0.19	0.82	0.42	0.30	0.11	0.44	0.24
10	0.51	0.29	0.61	0.40	0.38	0.22	0.35	0.24
11	0.44	0.22	1.21	0.55	0.22	0.13	0.61	0.30
12	0.38	0.20	0.58	0.55	0.29	0.25	0.31	0.29
13	1.17	0.55	0.77	0.41	0.65	0.35	0.43	0.27
14	1.22	0.53	0.87	0.42	0.73	0.33	0.47	0.23
15	0.35	0.20	0.60	0.43	0.15	0.12	0.30	0.24
16	0.64	0.23	0.52	0.35	0.32	0.14	0.29	0.22

and refine the initial orientations. In this subsection, we calculate the interaction energies with the rest of the protein of all template loops that were found in the loop searches and determine how well these energies correlate with the similarity in the internal geometry (''quality'') of the template loop to the known target loop. As above, the template loops are superposed on the target loop prior to evaluation. The main objective of this section is to determine how well the energy function can identify native loop conformations.

Scatter plots of the modified 9-6 energy and Levitt sidechain energy *versus* template loop quality were made, similar to those in Figure 2. Varying degrees of correlation were observed. For example, the correlation of the energy *versus* loop quality for 3fab loop H1 (seven residues), is much better than that for 3fab loop H3 (eight residues; Figure 7). In general, the correlations with addition of the Levitt sidechain energy were worse than those for the 9-6 energy alone. Identical results were found when the unmodified Levitt potential was added.

Since the Levitt sidechain energy did not correlate well with the loop quality, we investigated the performance of other methods for representing the sidechains of the template loops. These methods included two additional reduced sidechain models, and all-atom sidechain representations, where the atom positions are determined by using a rotamer library. First we check how well we can do with the 9-6 energy by itself, which only contains mainchain and C^{β} atoms for the template loops. This energy is very good at singling out the correct loop for target loops of ten residues and up, where the percentage of template loops with energies lower than the correct loop is about 1% (Figure 8a). Every dot represents the result for a given target loop search. A total of 90 searches were done, corresponding to the target loops of Table 1, full length, and reduced by two residues. Since all loop searches resulted in ~200 to 400 template loops, these results mean that there are usually less than five template loops with 9-6 energies lower than that of the crystal structure. The 9-6 energy is less specific for loops with nine residues or less. In two cases, nearly 40% of the template loops have a lower energy. The poorer performance for the shorter loops can be explained by the fact that for these short loops the set of template loops is conformationally much more similar to the actual loop. For longer loops, the differences are greater and template loops are more likely to have large steric clashes with the rest of the target protein. Figure 8b presents the results of these calculations

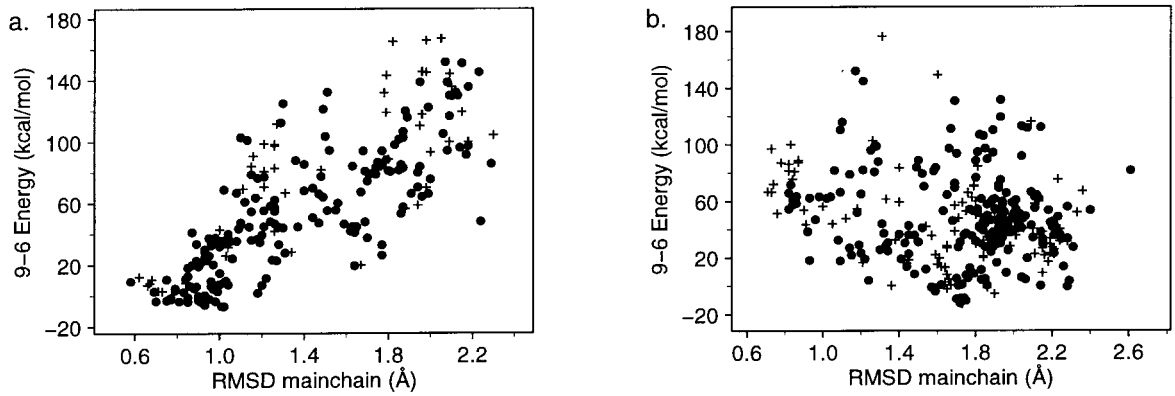

Figure 7. a, All template loops for the 3fab H1 (seven residues) target loop search. The 9-6 energy is plotted *versus* the loop quality. Circles represent loops from unrelated proteins, cross symbols represent loops from related proteins. b, As a, for the 3fab H3 (eight residues) target loop search.

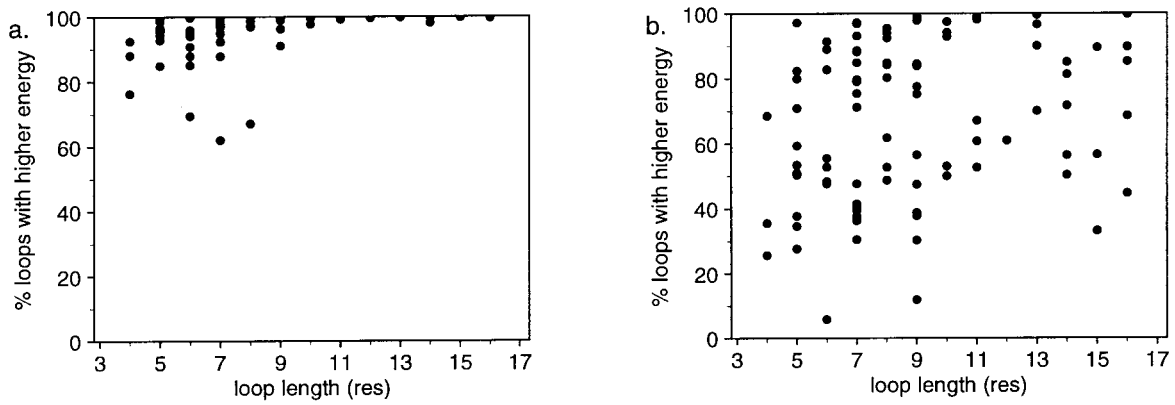

Figure 8. a, The rank of the 9-6 energy (equation (3)) of the target loop itself, among all template loops found, *versus* the length of the target loop. The template loops are selected by the search to be similar to the target loop, and are superposed on the target loop. Because at small loop lengths, many template loops are very similar to the target loop, the absolute energy differences are very small, which accounts somewhat for the low scoring loops. b, Corresponding ranks for sidechain-sidechain interaction energies calculated by the Levitt (1976) model.

for the reduced sidechain model of Levitt (1976). There seems to be virtually no discrimination between the correct loop and the set of template loop conformations, even for the longer loops. Corresponding plots for the other reduced sidechain models were qualitatively similar. This indicates that none of the reduced sidechain models can be of significant use in the selection or optimization of the loop orientations.

The energy calculations for the all-atom sidechain representations were done with the CHARMM program (Brooks *et al.*, 1983) with some alternative choices for the electrostatic part of the interaction energy. When we included only the mainchain atoms (no C^{β} atoms), the energy function was often unable to single out the correct loop, especially for loops of nine residues or less (Figure 9a). It is clear that the electrostatic part of the energy plays a significant role, since changing the dielectric function from vacuum (filled symbols) to the sigmoidal distance-dependent function (open circles), improved the results significantly. Completely turning off the electrostatics and including C^{β} atoms proved to be the best method, and the CHARMM non-bonded energy function was able to distinguish well between the crystal structure loop and the template loops (Figure 9b). The ability to identify the crystal structure loop from a set of template loops does not guarantee that the non-bonded energy function is also able to select the best template loop. In many cases, the lowest energy template loops do not have the lowest RMSD, and *vice versa*.

There is a slight improvement in results of the CHARMM energy of mainchain plus C^{β} over the 9-6 energy (cf Figures 8a and 9b). Since the same sets of atoms are included, and no electrostatics are used, the difference is caused by the fact that the CHARMM potential is more repulsive at short interatomic distances (12-6 *versus* 9-6 potential form with no modification at short distances (equation (3)), and that the CHARMM potential has more variation in atomic radii (e.g. the carbonyl carbon has a different radius from the aliphatic carbon).

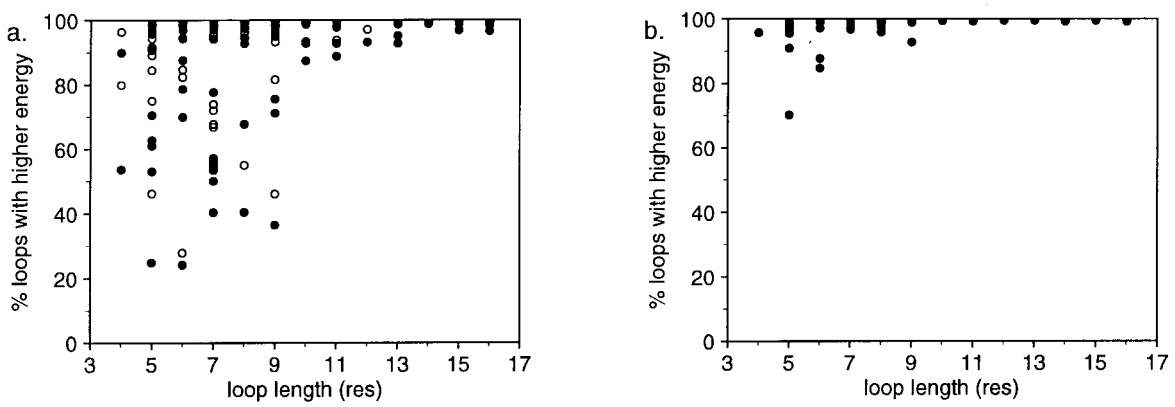

Figure 9. As Figure 8, for atom-based interaction energies calculated with the CHARMM potential. a, Only mainchain atoms (N, C^{α}, C, O) are included. Filled circles correspond to the vacuum calculation, open circles to the sigmoidal distance-dependent dielectric. b, Mainchain and C^{β} atoms are included, no electrostatics.

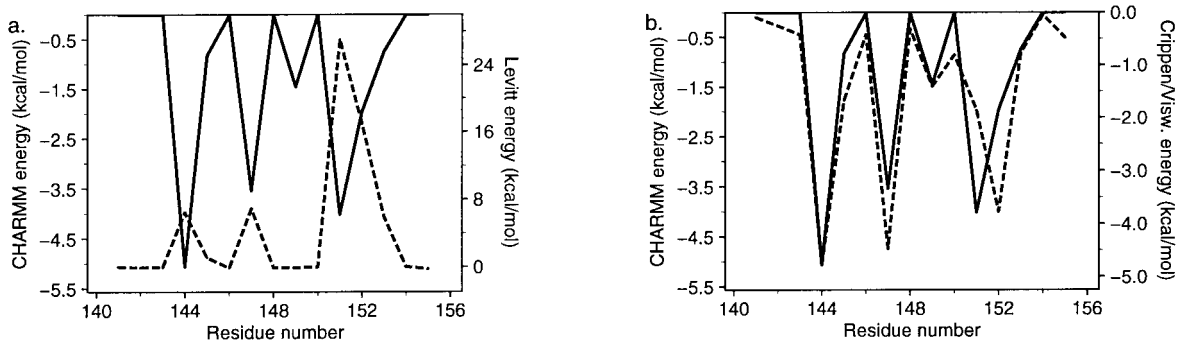

Figure 10. Individual target loop sidechain interaction energies with all other non-loop sidechains in the target protein, for the 2act loop2 (16 residues) target loop. The atom-based CHARMM energies (continuous lines) are plotted together with model sidechain energies (broken lines). a and b show results for the Levitt (1976) and Crippen/Viswanadhan (1985) models, respectively.

To include more of the sidechain, we used the rotamer library of Dunbrack & Karplus (1993). We compared these results with those obtained with the χ-angles of the crystal structure. When adding the sidechain atoms that are uniquely determined by the χ_1 angle, the results get worse and they deteriorate even more when atoms defined by the χ_2-angle are also added (results not shown). This suggests that it is better not to use the rotamer libraries at the loop orientation/optimization stage, and only to add the C^β atoms to the mainchain. The results did improve when using the crystal structure χ-angles to place the atoms further down the sidechain. When all sidechain atoms defined by χ_1 and χ_2-angles are included from the crystal structure, all crystal structure loops correspond to the best 5% (or less) of all template loops found in their respective searches (except one seven-residue target loop, which falls in the best 20%). This indicates that more accurate sidechain prediction algorithms are needed before the sidechain energies can be used. The sensitivity of the results to wrong sidechain conformation is caused by the hard sphere-like behavior of the CHARMM non-bonded potential at short distances. Short energy minimizations (50 steps SD) reduce this problem somewhat, but the results did not change significantly.

To investigate further the failure of the reduced sidechain models (e.g. Figure 8b), we made a direct comparison of the results from these models with the atom-based CHARMM potential. The models included the one by Levitt (1976), the potential function by Crippen & Viswanadhan (1985), the linearized hydrophobic potential by Casari & Sippl (1992), and the sidechain contact energy model by Gerber (1992). We compared the different energy values for the target loop whose structure is known and can therefore calculate all sidechain interaction energies. All interaction energies between individual sidechains of target loop residues with the non-loop sidechains of the target protein were calculated and compared. In particular, the CHARMM interaction energies of individual loop residues (all sidechain heavy atoms) were compared with the energies from the four different models, for all target loops of Tables 1 and 2 (a total of 85 loops). The results for target loop 2act loop2 (15 residues) highlight the differences in performance of the sidechain potential (Figure 10). The Levitt energies (Figure 10a) appear to be slightly anti-correlated with the CHARMM energy.

Table 6. Coefficients of cross-correlation between residue interaction energies, calculated with the CHARMM potential energy function, and four different reduced sidechain models

	CHM[a]	Levitt	Casari	Crippen	Gerber
CHM	0.545	−0.079	−0.117	0.328	−0.052
CHM[a]		−0.052	−0.157	0.377	−0.034
Levitt			−0.042	−0.111	−0.005
Casari				−0.008	0.042
Crippen					−0.052

CHARMM calculations were done without electrostatics (CHM) and with a sigmoidal distance-dependent dielectric (CHM[a]) (see Theory and Algorithms). The cross-correlation between variables x and y is defined as:

$$\frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2} \times \sqrt{\langle y^2\rangle - \langle y\rangle^2}},$$

where $\langle\ \rangle$ denotes average value.

This is most pronounced for sidechains that have a favorable CHARMM energy. The Crippen/Viswanadhan (Figure 10b) energies are better correlated with the CHARMM energies, and agree in most cases for the favorably interacting residues. Cross-correlations between CHARMM and sidechain model interaction energies were calculated, based on 706 individual sidechains from the target loops. Of the four different sidechain models, only the Crippen/Viswanadhan model correlates positively with the CHARMM energy (Table 6). The Levitt and Gerber energies have a near-zero correlation coefficient, and the Casari/Sippl energy is slightly anticorrelated with a coefficient of −0.12. Only minor changes occurred with a sigmoidal distance-dependent ε was used for the CHARMM energy, instead of no electrostatics term.

Cross-correlations between the different reduced sidechain models have coefficients close to zero (Table 6). In some cases this is surprising; e.g. the Casari/Sippl and Crippen/Viswanadhan energy functions are both derived from a statistical analysis of a set of proteins with known conformations. This suggests that these reduced sidechain potentials, which seem to work for identifying the native fold of a given sequence among a number of alternative protein folds (Casari & Sippl, 1992), miss specific details of the important interactions. They clearly are unreliable for predicting the conformation of a small segment, such as the loops of interest here.

Thus, the only energy functions that are useful in identifying the target loop conformation are the 9-6 and CHARMM potential, with the mainchain and C^{β} atoms of the template loop included. In many cases these energy functions are able to select the target loop structure from a set of very similar superposed template loops.

Loop selection and orientation based on stem criteria

Stem searches represent the true test for effectiveness of the PDB-based loop search method, since they correspond to real predictions of the best template loop candidates making use of the anchor residue coordinates and the target loop sequence. We now examine the performance of the loop search algorithm with the parameters extracted above by doing searches for the 130 target loops. For each target loop, a search was done using the four different $RMSD^{dist}$ parameters CA1, CA3, CACB1, and CACB3; the values used are given in Table 5. After the stem search, the template loop stems were fitted on the target frame by a least squares algorithm. Several schemes for superposing the stems were used, differing in the atoms included in the least squares fit.

An important question concerns the choice of $RMSD^{dist}$ parameter (CA1, CA3, CACB1 or CACB3). There are two issues: the number of loops that have to be included and the quality of the resulting templates. By using the worst-case parameter values displayed in Table 5, we know that the search will have at least one of the three best possible loops in its resulting set of template loops. We expect the same to hold more generally (for target loops outside this set of 130), especially for the loop lengths for which we used a large number of target loops (loop lengths five to nine all have 12 or more target loops). The total number of template loops found should be as small as possible to facilitate subsequent processing.

Using the worst-case $RMSD^{dist}$ parameter values of Table 5, we found a larger number of template loops for the parameters that are based on three stem residues on either side (CA3 and CACB3) compared to those with one residue on either side (CA1 and CACB1). The average number of template loops for CA1 and CACB1 was between 500 and 3000, with no apparent dependency on loop length. The numbers of loops are the ones obtained after the geometry check, which discards roughly 50% of the template loops in all searches. For CA3 and CACB3, the average number was usually larger, especially for loop lengths of seven (~7000 templates) and 11 (~4000 templates). For loop length 11 this is caused by one ''bad'' target loop (3est loop3), but for loop length seven there are several target loops that contributed high CA3 and CACB3 parameter values. When the average search parameters are used instead of the worst-case values, there is essentially no difference between the different parameters, and the number of templates ranged from 200 to 1500. Searches with average parameter values, however, are not guaranteed to give you at least one out of the best three template loops.

A second criterion for determining the best search parameter is the distribution of template loop qualities after the initial superposition of the stem residues. Because of the large variability in stem residue conformation for similar loop conformations (see Figure 1 and Table 3), it is very likely that a least squares superposition of stem residue atoms should be followed by further optimization. Obviously, in any such optimization it is best to start with an orientation that is close to the optimum from the initial placement. For all target loop lengths, we calculated the percentage of target loops for which the initial stem superposition resulted in at least one template loop with a mainchain RMSD of less than 1 Å, 1.5 Å, 2 Å, and 3 Å.

The reported searches were done with the worst-case CACB1 stem search parameters. Template loops from related proteins with the same sequence as the target loop were not included. For non-antibody loops, the probabilities of obtaining initial loop orientations with RMSD less than 1 Å is more than 50% only for lengths of four to six residues. For target loops of ten residues or more, we only find good initial orientations if related template loops are included. These related template loops are corresponding loops from related proteins. For the antibody loops, there is a much stronger influence of including related template loops; this sup-

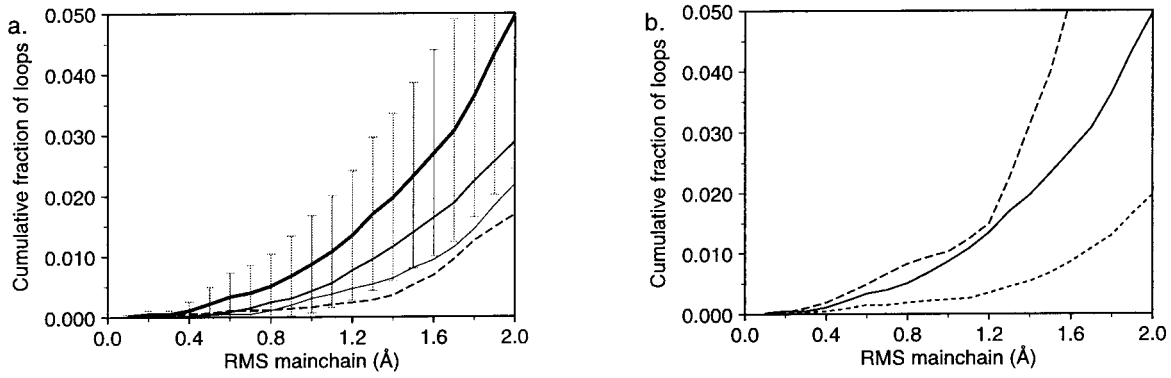

Figure 11. Cumulative initial RMSD distribution for all five-residue target loop stem searches with the worst-case CACB1 parameter. Related and unrelated template loops are included. a, For non-antibody target loops, different ways of fitting the stems are compared: using mainchain atoms of first (thick continuous line with vertical bars indicating the standard deviations), the first and second (medium), and the first three anchor residues (thin). The fourth fitting method included the C^{α} and C^{β} atoms of the first anchor residues (broken line). b, Comparison of the results for non-antibody target loops (continuous line) and antibody loops (broken) using the CACB1 parameter, and the results for non-antibody loops using the CACB3 parameter (short broken). In all cases, the mainchain atoms of the first anchor residues were fitted.

ports the concept of the existence of canonical structures. If they are included, we find initial loop RMSD values of less than 1 Å in more than 50% of the cases for lengths up to nine residues. Interestingly, for the three antibody target loops of eight residues, no canonically related loops were present in the database, but in two of the three cases unrelated template loops existed that resulted in an RMSD of less than 1 Å. In the stem superpositions, the mainchain atoms (N, C^{α}, C) of the first stem residues were used in the least squares fit, since this resulted in the best initial orientations.

Normalized initial cumulative distributions of loop RMSD values (in Å) for a loop length of five residues are plotted in Figure 11. Figure 11a shows the differences in initial RMSD values for non-antibody loops between four different ways of fitting the stems, using the mainchain atoms (N, C^{α}, C) of the first (thick line), the first and second (medium), and the first three anchor residues (thin), respectively. A fourth method uses only the atoms that defined the $RMSD^{dist}$ parameter, here C^{α} and C^{β} of the first stem residues (broken line). The results suggest that the best initial loop RMSD is obtained when the mainchain atoms of only the first stem residues are included in the least squares fit.

There seems to be no difference in initial loop RMSD between the antibody target loops (Figure 11b, broken line) and non-antibody target loops (continuous line). Up to an RMSD of 1.2 Å, they have similar initial distributions, even though the antibody searches were done with the extended database that included antibody structures. As it was shown that for non-antibody loops of four to six residues the initial orientation was often good, it is likely that the real advantages of the existence of canonical structures for antibodies become apparent for loops of more than six residues.

Results for the CA1 and CA3 parameters are virtually identical to those for CACB1 and CACB3, respectively. Although the average reduction of number of template loops by the geometry check is about 50%, its effect on the normalized distribution is negligible. There are subtle differences between the results for the four different parameters. For the loop length of five, as well as for lengths of four, six, and seven residues, the fraction of loops with a loop quality of better than 2 Å is higher for the searches with one stem residue (CA1, CACB1), than for the searches with three stem residues (CA3, CACB3). This is a direct result of the larger total number of template loops found for the CA3 and CACB3 parameters discussed before. Both the results of Figure 11 and those concerning the total number of found template loops indicate that there is no improvement of the loop search when three instead of one stem residues are used.

Optimization by use of energy functions

The initial orientation of the template loops after stem superposition is usually not good enough to be used directly in protein structure prediction. A mainchain RMSD of less than 1 Å was found consistently only for very short loops (four residues) and for antibody loops with canonical structures. Thus, in most cases there is a need to optimize the loop orientation by energy minimization. Earlier we investigated the ability of the 9-6 potential and the CHARMM potential to identify the crystal structure loop. In this section, we try to find the best energy minimization scheme to lower the loop RMSDs and to simultaneously indicate the best template loops by their energy values. We reorient the template loop, keeping it rigid or nearly rigid, by energy based methods to optimize their position in space. Such sets of optimized loops are a

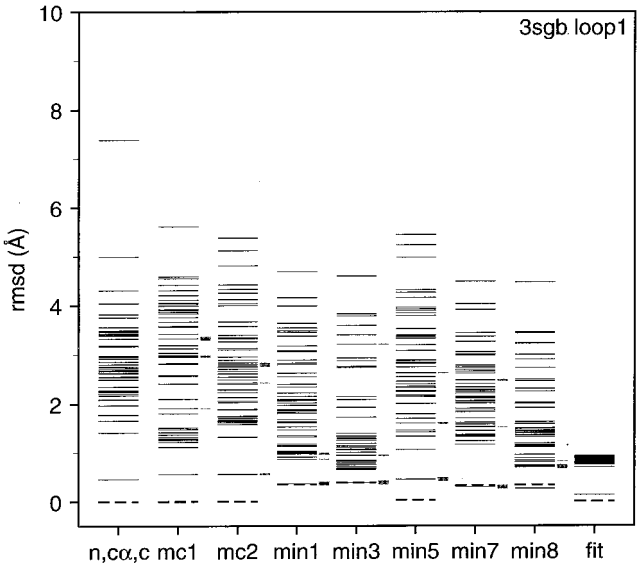

Figure 12. RMSD spectra of template loop minimization results for 3sgb loop1 (five residues). Every horizontal bar represents one template loop. Columns 1 to 9 show the 50 template loops: after initial stem superposition using N, C^{α}, and C atoms (n,cα,c); after MC optimization including mainchain and C^{β} (mc1), and including mainchain, C^{β}, and χ_1-defined atoms (mc2); after CHARMM minimization, including mainchain and C^{β} atoms, with (min1) and without (min3) dihedral constraints; after constrained CHARMM minimization with complete sidechains added according to crystal structure internal dihedrals (min5); after CHARMM minimization with sidechains constructed according to rotamer library (Dunbrack & Karplus, 1993), with (min7) and without (min8) dihedral constraints; after superposition of the loops themselves, which is the best reorientation possible for rigid template loops (fit). In columns 2 to 8, three short horizontal bars to the right side of the column indicate the template loop with the lowest (thick bar), second lowest (medium), and third lowest (thin) energy. The target loop itself also underwent minimizations and is represented by the broken bar in every column.

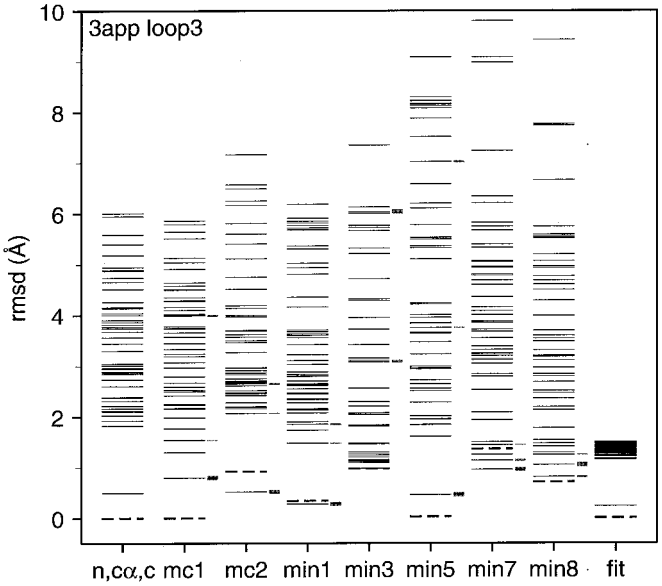

Figure 13. As Figure 12, for 3app loop3 (nine residues).

good starting point for internal minimizations. It appears from our results that such a two-step procedure is better than an unconstrained minimization of the initially placed loop.

Because this aspect of the calculation is time consuming, we applied the different minimization scheme to only five target loops, they are 3sgb loop1 (five residues), 3app loop3 (nine), 2act loop1 (15), 1rei L3 (six), and 4fab H3 (six). We then used the best minimization scheme to study 14 additional loops. To further reduce the calculation time, we kept only 50 template loops from every stem search. These were the 50 template loops with the lowest RMSD after superposition of the backbone on the target loop. This cannot be done in a real prediction, and therefore the calculations would be longer. Template loops with all backbone dihedral angles within 60° of another template loop in the set were excluded, to avoid multiple loops with the same conformation. Since these 50 loops were selected on the basis of their superposed RMSD, they are the ones with the lowest potential RMSD after reorientation and provide good test cases for the various minimization protocols. The particular loops that were selected for study are essentially arbitrary, other than that they cover the length range of the loops considered in this paper.

The results of the various minimization methods for 3sgb loop1 and 3app loop3 are summarized in Figures 12 and 13. The first column shows the RMSDs of the 50 template loop backbones (N, C^{α}, C) after superposition of the stem residues, using N, C^{α}, and C, in calculating the fit. Columns 2 and 3, labeled mc1 and mc2, list the results of two sets of rigid-body MC optimizations that were done with the 9-6 potential. The first included only the mainchain and C^{β} atoms of the template loops and the second also included all atoms defined by the χ_1 torsion angles of the sidechains. The χ_1 angles were assigned by using the backbone dependent rotamer library (Dunbrack & Karplus, 1993). Three short horizontal bars at the right side of every column indicate the three template loops with the lowest interaction energy with the target protein after minimization; the thickest bar corresponds to the lowest energy loop, the medium bar to the second lowest, and the thin bar to the third lowest. For comparison, the RMSD of the target loop with the crystal structure is represented after the same ''optimization'' by the broken line in every column. The rightmost column shows the RMSD values of the 50 template loops after superposition of the loop mainchain on to the target loop. Thus, these are the best possible RMSDs that could be reached with rigid optimizations of the template loops; in four out of five cases the optimum is an RMSD of

less than 1 Å and in the fifth (the largest loop, 2ACT loop1) it is less than 2 Å. In most cases, the MC optimization performed poorly in an overall sense. The average RMSD compared to the initial fit went up, and the lowest energy loop did not correspond to the one of the low-RMSD loops. Instead of running the full MC run at 5000 K and selecting the lowest energy orientation, we also did annealing runs, where the temperature was slowly cooled from 5000 K to 300 K. This worsened the results slightly, rather than improving it (results not shown). The MC optimization did work well for the 3app loop3 target loop, where the lowest energy template loop has an RMSD of less than 1 Å for both MC methods (Figure 13). This particular template loop is a corresponding loop from a related protein and has a nearly identical conformation (RMSD after loop superposition 0.23 Å). The MC2 optimization is somewhat better than MC1 and, except for 1rei L3, had a low RMSD for at least one of the three lowest energy structures.

The results of five different optimization schemes with the CHARMM potential are also shown (Figures 12 and 13). Columns 4 and 5, labeled min1 and min3, correspond to minimizations of the template loop with only the mainchain and C^{β} atoms (and their corresponding hydrogen atoms). Minimizations were done with (min1) and without (min3) the dihedral constraints described above. No electrostatic terms were included in the results given here, but minimizations with electrostatic terms slightly worsened the results (not shown). In accord with the results of the CHARMM potential above, the inclusion of only the mainchain and C^{β} atoms works well for identifying the template loops with low RMSDs. In all cases, the constrained minimization results in a lowest-energy loop with a relatively low RMSD. The 2act loop1 target loop is an exception, probably because even in their optimal orientations the template loops all have RMSD's of ~2 Å or more. These orientations usually correspond to high energies, because of the relatively tight loop packing (Leszczynski & Rose, 1986), and thus the minimizations will not result in low RMSDs. The unconstrained minimization slightly reduces the average RMSD values of the template loop sets, but in two cases, it results in a high-RMSD loop with the lowest energy (3app loop3, 1rei L3).

The importance of including atoms of the sidechains in template loop optimization was tested in two ways. We tried the best-case scenario with the complete sidechains in the crystal orientation on the template loops. The results of minimizing without dihedral constraints are shown in column 6 labeled min5. A more realistic scheme for modeling purposes uses sidechain dihedral angles assigned according to the backbone-dependent rotamer library (Dunbrack & Karplus, 1993); undefined dihedrals were set to 180°. The complete sidechains were built, and minimizations were done with and without dihedral constraints (column 7 and 8, labeled min7 and min8, respectively). In the minimizations with sidechains, electrostatics did not give worse results, and they were included to provide a better representation of the sidechains, which are often charged. Comparing columns 6 with columns 4, we can see that there is no significant improvement when the best-case scenario sidechains are added. The addition of the sidechains according to the rotamer library also did not improve the results. For none of the cases we found a significant lowering of the RMSD values,

Table 7. RMSD values, in Å, of mainchain atoms (N, C^{α}, C) of the three template loops with lowest energy

		mc1	mc2	min1	min2	min3	min4	ann	min5	min6	min7	min8
3sgb-1 (5)	1	3.33	2.80	*0.37	*0.25	*0.38	*0.24	*0.60	*0.45	0.73	*0.30	0.71
	2	2.97	*0.56	0.98	1.48	0.94	0.63	1.50	1.60	0.77	2.48	0.70
	3	1.91	2.43	0.86	1.73	3.21	0.63	2.03	2.63	*0.27	1.52	0.81
3app-3 (9)	1	*0.80	*0.52	*0.28	0.29	*6.04	6.00	*0.40	*0.46	*0.65	*0.95	1.05
	2	3.99	2.65	1.85	1.74	3.08	1.34	2.02	7.02	0.58	1.14	*0.81
	3	1.54	2.07	1.48	1.57	3.11	1.21	2.82	3.75	1.30	1.44	1.25
2act-1 (15)	1	7.68	3.32	5.99	6.80	5.57	3.58	6.86	17.15	4.43	5.92	4.76
	2	5.43	19.66	15.29	7.72	4.78	6.29	5.78	8.50	10.72	5.70	3.82
	3	3.18	20.43	8.03	11.25	4.26	4.19	5.72	7.72	6.18	6.85	4.30
1rei-L3 (6)	1	6.21	6.43	*0.50	3.37	3.48	2.60	2.73	*0.45	*0.54	*0.97	5.04
	2	6.12	5.37	1.28	1.78	1.92	3.68	*1.02	*0.60	*0.60	2.65	2.40
	3	5.88	5.91	3.35	1.40	1.36	1.63	1.67	*0.44	1.66	*0.44	1.35
4fab-H3 (6)	1	3.00	1.78	1.07	1.12	1.41	1.01	2.38	1.62	0.87	1.09	0.97
	2	3.30	2.71	1.36	1.17	1.41	1.01	2.73	0.86	0.85	0.93	1.02
	3	1.84	3.21	1.20	1.17	2.33	1.01	2.47	0.71	0.85	1.11	1.57

Target loop names are abbreviated: 3sgb-1 stands for 3sgb loop1, etc. Target loop lengths in residues are shown in the second column. The third column indicates the lowest energy loop (1), second lowest (2), and third lowest (3). The columns represent different optimization schemes that were used: mc1: MC optimization with 9-6 potential, mainchain atoms + C^{β}; mc2: as mc1, but also atoms defined by χ_1 included; min1: minimization with CHARMM potential, dihedral constraints; min2: as min1, no dihedral constraints; min3: as min1, with electrostatic terms included; min4: as min2, with electrostatic terms included; ann: as min1, but with simulated annealing instead of plain minimization; min5: minimization with dihedral constraints and electrostatics, sidechains added according to crystal structure internal coordinates; min6: as min5, no dihedral constraints; min7: as min5, sidechains added according to rotamer library (Dunbrack & Karplus, 1993); min8: as min7, no dihedral constraints. Corresponding template loops from related proteins are indicated with an asterisk.

Table 8. Mainchain atom (N, C^{α}, C) RMSD values, in Å, of ten target loops from Tables 1 and 2

	2apr-1 (6)	8abp-3 (6)	2act-5 (6)	3tln-6 (7)	3grs-1 (7)
1	5.16	0.28	1.58	3.70	4.55
2	4.68 (0.09)	0.38 (0.13)	1.77 (1.78)	2.17 (0.78)	4.94 (4.07)
3	1.59 (0.77)	0.41 (0.28)	1.82 (1.86)	1.55 (2.38)	4.17 (5.53)
B	0.62 (1.79)	0.28 (0.00)	1.47 (2.23)	1.45 (4.81)	1.02 (7.30)
	5cpa-2 (7)	2fb4-H1 (7)	2fbj-H3 (7)	3tln-4 (8)	3sgb-3 (9)
1	2.14	1.62	0.49	1.83	1.79
2	1.42 (7.04)	1.97 (0.48)	0.89 (1.87)	5.23 (1.02)	3.03 (6.16)
3	1.95 (7.19)	2.46 (1.90)	1.28 (3.71)	0.61 (1.36)	2.17 (9.68)
B	1.18 (58.7)	0.36 (3.47)	0.49 (0.00)	0.61 (1.36)	1.12 (17.2)

The three template loops with lowest loop-protein interaction energy are shown. The energy difference to the minimum energy is given in parentheses (kcal/mol). The number of residues of each target loop is shown in parentheses after the loop name. The row labeled *B* shows the template loop with the lowest final RMSD.

or better correspondence between loops with low energy and loops with low RMSD.

In Table 7 we list the predictions that would be made based on the energy values. We list the RMSDs of the three template loops with the lowest energy. If the template loop is a corresponding loop from a related protein, it is marked with an asterisk. Although several template loops for the 2act loop1 target loop could theoretically reach an RMSD of less than 2 Å, the lowest RMSD that would be selected by any of these methods is 3.6 Å (min4). For the four remaining template loops, both the constrained minimization of the mainchain+ C^{β} atoms (min1), and the constrained minimization of the loops with rotamer library sidechains (min7) give satisfactory results (three out of four under 1 Å RMSD, one just above). This was also reflected in energy *versus* RMSD scatterplots of the 50 template loops, where these two minimization methods showed the best correlations (results not shown). Surprisingly, there does not seem to be a clear difference between the antibody loop with canonical structure (1rei L3) and the one without (4fab H3). Good template loops for the 4fab loop were found in corresponding positions of other antibody structures, as well as in unrelated proteins. As mentioned before, this could be because the advantage of canonical structures becomes clear only at loop lengths of seven or more residues. The dihedrally constrained CHARMM minimizations gave significantly better results than the rigid body MC runs. There are two likely causes for this. First, in the CHARMM minimizations we allow the first and last residue of the loop to be unconstrained, and connect them to the target protein with a strong bond. This moves these two residues of the template loop closer to the crystal structure orientation. For the MC optimizations, the full template loop is rigid and is held close to the target protein by relatively weak C^{α}–C^{α} bonds and angles. Second, although the template loops are constrained, they are not completely rigid, and can deform somewhat during the CHARMM minimizations. They can adjust somewhat to the rigid target protein and then find orientations with lower RMSDs; in the MC optimizations, the template loops are completely rigid. However, it is also clear that significant constraints are important in obtaining the best results. In the absence of constraints, the loop can distort locally, rather than moving globally to a lower energy position. Table 7 also shows that a simulated annealing scheme with dihedral constraints (column ann) performed worse for all loops compared to the constrained minimization.

Since the constrained minimization of the mainchain + C^{β} atoms (min1) performed best, we used it for 14 additional target loops. The first ten were selected from the set of target loops in Tables 1 and 2 with emphasis on intermediate length loops

Table 9. As Table 8, for four target loops from Fidelis *et al*. (1994)

	3dfr-1 (4)			3dfr-5 (5)			3dfr-6 (5)			3blm-1 (5)		
	Ener.	Init.	Fit	Ener.	Init.	Fit	Ener.	Init.	Fit	Ener.	Init.	Fit
1	2.64	2.56	1.32	1.62	1.68	0.93	0.47	1.33	0.37	0.82	1.27	0.60
2	0.75 (1.07)	1.65	0.60	0.93 (5.91)	0.96	0.64	0.41 (0.10)	0.75	0.36	0.86 (0.10)	1.59	0.63
3	0.68 (1.73)	1.63	0.42	1.30 (12.1)	1.42	0.66	0.42 (0.95)	0.89	0.26	0.72 (1.35)	1.55	0.64
B	0.44 (3.72)			0.93 (5.91)			0.33 (1.77)			0.65 (2.57)		
F			1.47			1.23			1.09			2.15

The column Ener. shows the three lowest-energy loops. The column Init. shows the RMSD values after initial superposition. The column labeled Fit lists the RMSD values after superposition of the minimized template loops on the target loop. The last row, labeled *F*, shows the lowest mainchain RMSD values (after loop superposition) that were found in the database search of Fidelis *et al*. (1994).

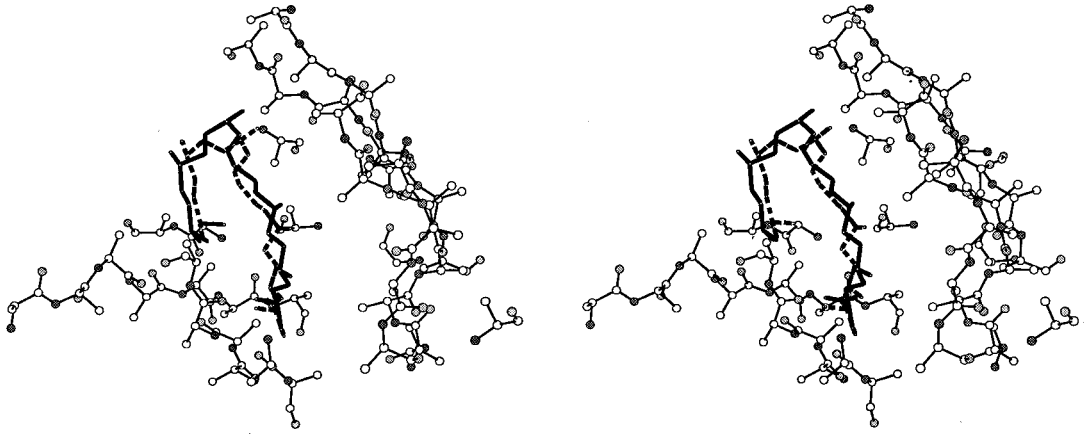

Figure 14. Wall-eyed stereo drawing of 2fbj H3 loop (seven residues) and all protein residues within 7 Å of the crystal structure loop. Only mainchain and C^{β} atoms are shown (no hydrogens). The crystal structure loop is drawn with thick continuous bonds, the predicted loop (RMSD 0.49 Å) is drawn with thick broken bonds.

(three of six residues, five of seven, one of eight and one of nine). The last four loops were shorter (four or five residues) and were taken from the paper by Fidelis *et al*. (1994). For these loops they found that a systematic search produced significantly better template loops (in terms of internal RMSD) than a database search. The database performance test done here is much more stringent, since we report the template loops with the lowest energies instead of the ones with the lowest RMSD. Moreover, we do not report the internal RMSD (RMSD after superposition of loop), but the actual RMSD of the template loop as it is built into the protein. From the Fidelis *et al*. (1994) paper, we chose dihydrofolate reductase (PDB entry 3dfr) loops f1 (residues 20 to 23), f5 (89 to 93), and f6 (120 to 124), and β-lactamase (3blm) loop f1 (164 to 168). Tables 8 and 9 list the initial and final mainchain RMSD values of the three lowest energy template loops found by the database search and subsequent optimization. For comparison with the results of Fidelis *et al*. (1994), we also report in Table 9 the RMSD values after superposition on the target loop.

None of the template loops shown in Tables 8 and 9 was a corresponding loop from a related protein. The prediction results are variable. Loops with at least one template loop below 1 Å are 8abp loop1, 2fbj H3, 3tln loop4, 3dfr loops 1, 5, and 6, and 3blm loop1. Most other target loops have at least one with an RMSD of 1.8 Å or better. A very poorly predicted loop is 3grs loop1, for which all three lowest energy template loops have an RMSD of more than 4 Å. Five of the 50 template loops minimized to an RMSD of 1.5 Å or lower, but had slightly higher energies. This target loop has a low number of non-bonded contacts (distance <5 Å) between its mainchain + C^{β} and the rest of the protein (excluding the loop itself and the two neighboring anchor residues). For individual loop lengths of five, six, and seven residues, we find consistently that the target loop with most non-bonded contacts is predicted best, and the one with least contacts is predicted worst. For five-residue loops, the best predictions were for 3dfr loop6 (average RMSD of three lowest energy template loops: 0.44 Å) which has 50 non-hydrogen non-bonded contacts within 4 Å and 169 within 5 Å. Worst predicted was 3dfr loop5 (1.28 Å) with 18 and 84 contacts, respectively. For six-residue loops, best results were obtained for 8abp loop3 (0.36 Å) with 61 and 252 contacts, and worst results for 2apr loop1 (3.81 Å), with 15 and 58 contacts. For seven-residue loops, we found best results for 2fbj H3 (0.89 Å) with 61 and 193 contacts (Figure 14), and worst results for 3grs loop1 (4.55 Å) with 14

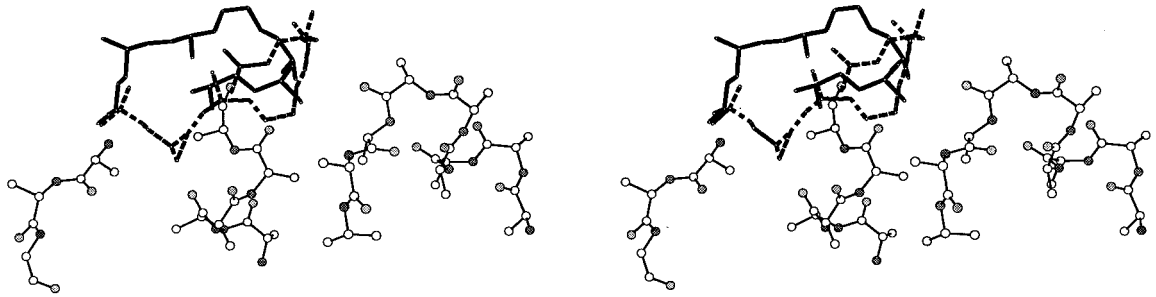

Figure 15. Drawing of 3grs loop1 (seven residues) with crystal structure loop and predicted loop (RMSD 4.55 Å), analogous to Figure 14.

and 45 contacts (Figure 15), respectively. A low number of non-bonded contacts implies that the protein framework has only a small influence on the loop conformation, and internal forces and solvent effects are relatively more important. These effects are not taken into account in the present calculations. Since the conformation of the target loop is unknown, this non-bonded contact analysis cannot be used to estimate the predictability of a particular target loop. However, we also analyzed the number of contacts of the 50 minimized template loops. If all template loops had a low number of contacts with the protein framework (0 to 30 heavy atom contacts of <4 Å) for target loops of six and seven residues, they were poorly predicted (2apr-1, 2act-5, 3grs-1). By contrast, well predicted loops (e.g. 8abp-3, 2fbj-H3) had high numbers of contacts for most templates (30 to 100). This suggests that the number of contacts that exist for the set of template loops can serve as an approximate accuracy criterion for the target loop prediction.

The results for the 3dfr and 3blm loops are better than those reported by Fidelis *et al*. (1994), even though our RMSDs are based on more stringent criteria. In all cases one of the three lowest energy template loops had an RMSD of less than 1 Å. The main cause for this improvement is our use of larger cutoffs in the stem searches. After the stem searches, all four target loops have at least one template loop with a mainchain RMSD of less than 1 Å, in contrast to results of Fidelis *et al*. (1994). The subsequent energy minimization energy minimization improves the RMSDs and makes it possible to choose the best, based on the energies.

One loop prediction takes from 50 minutes to two hours on a SGI Power Onyx work station, depending on the loop length. The actual database search takes only about five minutes, and the rest of the time is spent on minimization of the template loops. A full loop prediction would take about 20 times longer, assuming that on the order of 1000 template loops have to be minimized. To speed up the calculation, template loops with high energy could be discarded early in the minimization process.

Discussion

An approach for the predictions of loop structures is presented and analyzed. It starts with a database search of known protein structures. Given the database, there are three parts to the prediction: the selection of loop candidates, their initial orientations in the target protein, and their evaluation and optimization. A set of 130 target loops from known protein structures has been used to test various aspects of the approach.

Analysis of the present PDB database has shown that it contains templates for small to medium-sized loops of up to nine residues. Generally, one cannot expect to find template loops in the PDB with a mainchain RMSD of less than 1 Å for loops of more than nine residues, though some useful results can be obtained from the database for longer loops.

Given such a database, it can be used for finding template loop candidates only if there exist search parameters that correlate with loop quality (measured as RMSD of the superposed loop mainchain). In the cases studied here, correlations between loop quality and search parameter were variable. Strong correlations were found between the loop quality and the RMSD of the stem residues (stem1-3), but these parameters cannot be used in a stem search. For actual searches, the $RMSD^{dist}$ values (CA1, CA3, CACB1, CACB3) are most useful. The R^2s for the correlation between loop quality and $RMSD^{dist}$ were as high as 0.7 in some cases. The highest R^2 values correspond to loops of less than nine residues, and the correlations were found to be worse for longer loops; e.g. for loops of length 11, R^2 values on the order of 0.10 were obtained in general. Since the correlation is rather weak, loop prediction algorithms have to use large cutoffs to guarantee that the best possible loop in the database is included; e.g. for loops of length five, an average of 1000 loop alternatives must be included and for loops of length nine, the number is about 1500. Analysis shows that the required number increases only slowly, if at all, with loop length, in contrast to the exponential time increase involved in an exhaustive search. We found the use of sequence homology to be virtually useless in the selection of template loops. The correlation between the homology score and the template loop quality was not significant for any of the five scoring matrices that were tried (Risler *et al*., 1988; Niefind & Schomburg, 1991; Gonnet *et al*., 1992; Henikoff & Henikoff, 1992; Šali & Blundell, 1990; results not shown).

The set of hypervariable loops in antibody structures is an exception. Because of the available loop conformations on similar protein frameworks, Chothia & Lesk (1987) were able to derive rules that predict a particular loop sequence to fold in one of a limited set of conformational possibilities (canonical structures). It is likely that any loop, with a substantial amount of interaction with the protein framework, has a limited set of ''canonical'' structures, in analogy with the antibody results. On the other hand, there is no reason to assume that the known set of canonical structures for antibodies is complete (Steipe *et al*., 1992; Wu & Cygler, 1993). The availability of canonically related loops in the PDB improves the results obtained from template loop searches. It was shown that for antibodies, except for the H3 loops, the most similar template loops available in the PDB are the canonically related loops. In addition, the results after superposition of the stems are also much better for the antibody loops than for loops in general, if there are more than six residues. This can be explained by the structural invariability of the protein framework, and more specifically the

stems, of antibodies. The antibody loops that do not have a template loop with RMSD <1 Å after stem superposition, are either H3 loops (no canonical structure), long loops (>ten residues), or loops with no good canonical relative in the PDB. A recent modeling study using canonical structures and conformational search algorithms (Bajorath & Sheriff, 1996) indicates the problems that exist even under ideal conditions for antibody hypervariable loop modeling. Although loops with excellent quality were found, their orientation in space led to mainchain RMSD values of 0.34 Å (L2) to 2.85 Å (H3). The lack of good templates for H3 loops does not mean that PDB-based methods cannot be used in prediction of their conformations. By using energy minimization, both H3 loops that we tried to predict had reasonable accuracies (RMSD of 0.49 Å and 1.07 Å), even though no corresponding antibody loops were in the database.

Different loop modeling studies based on the PDB have used various numbers and schemes of stem residues at each end of the loop to select the loops (Claessens *et al*., 1989; Fidelis *et al*., 1994 (two before, one after loop); Summers & Karplus, 1990 (three, three); Tramontano & Lesk, 1992 (four, four)). The present study indicates that the particular choice has only a small effect on the results. Use of the C^{α} and C^{β} of one stem residue on each side of the loop works somewhat better than three. We see no advantage in using stems of more than one residue for selecting and positioning the loops.

Fidelis *et al*. (1994) conclude that database searches are limited to loops of four residues. From our results it appears that the present PDB is useful as a database for longer loops as well, since for our set of target loops, templates with a loop quality of around 1 Å are present in the database for loops up to nine residues.

Once a set of loops has been selected, it must be positioned in the protein. The accuracy of the initial placement of template loops by superposing stem residues was dependent on the stem superposition method; best results were obtained by superposing only the first stem residues, using the N, C^{α} and C atoms. The accuracy was determined by calculating the loop mainchain RMSD, when a known loop conformation was modeled. Usually these RMSD values were high (RMSD ~2 to 5 Å), a direct consequence of the large variation in stem conformations for similar loops. To permit selection and improve the orientation of the template loops, an optimization based on an energy function was made. The template loops can be reoriented as essentially rigid bodies, because the dihedral conformational space is already explored by doing the PDB search. It was shown that a modified 9-6 potential or a CHARMM-based interaction energy between the mainchain (N, C^{α}, C) and C^{β} of the template loop and the rest of the protein structure, can discriminate rather well between the native loop and other conformations similar to the native loop. Fine *et al*. (1986) also used a mainchain + C^{β} (although including electrostatics) and found for four antibody loops (five to 11 residues) that minimization of the crystal structure resulted in only small movements. For three of the four loops they also found reasonable correlations between RMSD (with respect to minimized loop) and energy.

A variety of schemes and potential functions (i.e. four different reduced sidechain models) were tried to account for the effect of sidechains, but were found to be ineffective. Because of the average nature of the model sidechains, close sidechain-sidechain contacts (with a favorable CHARMM energy) often give rise to high model sidechain energies. Further development in reduced sidechain models (Sun, 1993; Kolinski & Skolnick, 1992; Kolinski *et al*., 1993) may help to overcome this problem. Using predicted atom based sidechains worsened the ability of the energy function to select the crystal structure loop from the set of similar template loops.

The inclusion of an electrostatic energy term in the energy calculation worsened the results, even for the sigmoidal distance-dependent ε (Figure 9a). This sigmoidal ε was developed to take the solvent effect into account approximately. In the case of surface loops it does not perform very well, which may indicate that a high effective ε is appropriate. Smith & Honig (1994) report that the inclusion of the electrostatic solvation free energy (by solving the Poisson-Boltzmann equation) gives better results than the use of a distance-dependent ε. However, when they set the atom charges of charged sidechains to zero, to better mimic solvent screening in the distance-dependent ε calculations, the results improved. Compared to this more strongly shielded electrostatic energy, the use of the full electrostatic solvation free energy only proved marginally better.

Model building in X-ray structure refinement is a useful application (Jones & Thirup, 1986), since in addition to the interaction energy, one can use the experimental electron density to distinguish good template loop conformations from bad ones. In loop construction without X-ray density information, the use of the database search method relies heavily on the ability of available energy functions to identify good loop conformations.

After determining the ability of the different energy functions to discriminate between the crystal structure loop conformation and a set of alternate template loop conformations, we used several of the energy functions to optimize template loop positions that resulted from a least squares fit on the stem residues. Nineteen loops were used to test various optimization protocols. The best results were obtained with a dihedrally constrained minimization of the loop mainchain and C^{β} atoms. The addition of predicted sidechains did not improve the results. Constrained CHARMM minimizations outperformed the rigid body Monte Carlo method, most likely because of the flexibility of the terminal residues of the loop, and the fact that the dihedral constraints do not make the loops completely rigid. We also employed a simulated annealing

scheme with the constrained CHARMM potential, but the results were significantly worse than the minimizations. It is possible that a more extensive annealing scheme (ours consisted of a gradual cooling from 1500 K to 300 K for a total of 25 ps, followed by energy minimization) would perform better.

The results indicate that complete rigidity of the template loops will often result in relatively high energies, even if the template loop orientation is close to the target loop orientation. This is a direct result of the close packing of loops that was reported by Leszczynski & Rose (1986). A few slightly displaced atoms can increase the energy significantly and complicate the prediction.

Recently, Zheng & Kyle (1996) obtained very good results with their scaling-relaxation method for three out of four protein surface segments of seven residues (RMSD less or equal than 1 Å for the lowest energy conformer). When we tried their energy function for the 19 loops predicted in this paper, our dihedrally constrained method worked significantly better. This emphasizes the need to examine a large number of different loops to obtain a meaningful test of a prediction method.

The PDB search serves primarily as a way of sampling a set of mainchain dihedral angles that conform to the fixed endpoints of the loop. Several algorithms do the same thing by construction, without using information from the PDB. Examples are Bruccoleri & Karplus (1985), Moult & James (1986), Shenkin *et al.* (1987), Dudek & Scheraga (1990), Zheng *et al.* (1993a), and Ring & Cohen (1994). The PDB search is the fastest available method for constructing a set of template loops that can exist. It does not have the exponential increase in time with loop length of the systematic search methods. However, the resulting set may be limited; e.g. for loops of ten residues or more, the set did not include any conformation close to the actual one. Fidelis *et al.* (1994) concluded that for a set of loops of four to six residues, systematic search methods worked better than the database search method. The comparison is biased by the fact that their systematic search involved a minimization (400 steps, no electrostatics) of the mainchain + C^{β}s of each loop. Such an energy function has been shown here to be useful for identifying the correct target loop and for optimization of its position. Thus, a definitive comparison would involve optimizations in the database search, as well as for the systematic constructs, and use of a larger number of loops in the former. In addition, for the database search a stem residue RMSD cutoff greater than 1 Å (that of Fidelis *et al.* (1994)) should be used.

Besides being faster, database methods for loop construction have another advantage. Systematic construction methods mostly build the loop conformations on the known stem residue coordinates. In real homology modeling the stems are often not very well defined, since they are located at the ends of secondary structure segments. Therefore, a systematic search can miss the correct loop conformation, due to its use of the incorrect stem conformation. Database construction methods have an implicit inclusion of stem residue variability, because of the RMSD cutoffs used in the loop selection. They are, therefore, not as dependent on the correct stem conformation. Database searches are also ideally suited for the generation of loops in protein engineering applications, such as the one described by Borchert *et al.* (1994). Unlike systematic construction methods, they simultaneously offer alternatives in loop conformation and sequences that are able to adopt these conformations.

The main obstacle for successful use of database methods in loop prediction is the poor orientation of the loop that results from the stem superposition. It was shown here that energy minimization can improve the orientation significantly, and may result in good predictions. Optimizations of the initial orientation cost considerably more time than the database search itself, however, and may offset the faster search advantage of database methods over construction methods. It is very likely that the most efficient method depends on the environment of the loop. Full construction methods such as CONGEN (Bruccoleri & Karplus, 1985) take a very long time to complete if the loop is not restrained by neighboring loops or other parts of the protein, or if the loop is more than six residues long. Other construction methods (Zheng *et al.*, 1993a; Shenkin *et al.*, 1987), which appear not to increase exponentially with the number of residues, also have problems with unconstrained loops, because there is a large number of possibilities with nearly identical energies. The database search method only has a limited number of template loops to analyze, and may be faster for unconstrained loops than the other methods. This would particularly be true if the total number of existing loop conformations in all proteins is smaller than the combinatorial number. Such a limited set of conformations (families) has so far been proposed only for entire proteins (Orengo *et al.*, 1994).

The present analysis and application of various selection and optimization criteria for loop predictions starting with a PDB-based search suggests the following algorithm for loops of nine residues or less: (1) selection of template loops from the database with the CACB1 parameters listed in Table 5; (2) superposition of the template loop stems using the mainchain atoms; (3) minimization of the template loops in the target protein with the CHARMM energy function and dihedral constraints; of the sidechains, only the C^{β} atoms should be included and no electrostatic terms should be used at this stage; (4) sidechain construction and more detailed unconstrained minimizations of the template loops to arrive at the final prediction; and (5) evaluation of the results in terms of the number of non-bonded contacts. For longer and highly exposed loops the present approach can be used but the expected lack of good

template structures in the database means that the likelihood of an accurate prediction is small.

Acknowledgements

We thank Andrej Šali for helpful discussions and for making the plotting program ASGL available to us. We thank Hsiang-ai Yu for helpful advice concerning the loop search problem. The calculations were done on an SGI Predator with an R3000 processor and an SGI Power Onyx with an R8000 processor. This work was supported in part by a grant from the National Institutes of Health.

References

Allen, M. P. & Tildesley, D. J. (1987). In *Computer Simulations of Liquids*. Clarendon Press, Oxford.

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Borchert, T. V., Abagyan, R., Jaenicke, R. & Wierenga, R. K. (1994). Design, creation, and characterization of a stable, monomeric, triosephosphate isomerase. *Proc. Natl Acad. Sci. USA,* **91**, 1515–1518.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.

Bruccoleri, R. E. & Karplus, M. (1985). Chain closure with bond angle variations. *Macromolecules,* **18**, 2767–2773.

Bruccoleri, R. E. & Karplus, M. (1990). Conformational sampling using high-temperature molecular dynamics. *Biopolymers,* **29**, 1847–1862.

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

Carlacci, L. & Englander, S. W. (1993). The loop problem in proteins: a Monte Carlo simulated annealing approach. *Biopolymers,* **33**, 1271–1286.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature,* **342**, 877–883.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). Structural repertoire of the human V_H segments. *J. Mol. Biol.* **227**, 799–817.

Claessens, M., Van Cutsem, E., Lasters, I. & Wodak, S. (1989). Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng.* **2**, 335–345.

Collura, V., Higo, J. & Garnier, J. (1993). Modelling of protein loops by simulated annealing. *Protein Sci.* **2**, 1502–1510.

Collura, V. P., Greaney, P. J. & Robson, B. (1994). A method for rapidly assessing and refining simple solvent treatments in molecular modelling. Example studies on the antigen-combining loop H2 from FAB fragment McPC603. *Protein Eng.* **7**, 221–233.

Coleman, J. E., Williams, K. R., King, G. C., Prigodich, R. V., Shamoo, Y. & Konigsberg, W. H. (1987). Mapping the functional domains in single-strand {DNA}-binding proteins Gene5 and Gene32. *Protein Engineering,* (Oxender, D. L. & Fox, C. F., eds), pp. 323–336, Alan R. Liss, Inc., New York.

Creighton, T. E. (1993). In *Protein Folding*. W. H. Freeman and Company, New York.

Crippen, G. M. & Viswanadhan, V. N. (1985). Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* **25**, 487–509.

Dudek, M. & Scheraga, H. A. (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.

Dunbrack, R. L., Jr & Karplus, M. (1993). A backbone dependent rotamer library for proteins: application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.

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

Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L. & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins: Struct. Funct. Genet.* **1**, 342–362.

Finkelstein, A. V. & Reva, B. A. (1992). Search for the stable state of a short chain in a molecular field. *Protein Eng.* **5**, 617–624.

Gerber, P. R. (1992). Peptide mechanics: a force field for peptides and proteins working with entire residues as smallest units. *Biopolymers,* **32**, 1003–1017.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science,* **256**, 1443–1445.

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

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.

Joseph, D., Petsko, G. A. & Karplus, M. (1990). Anatomy of a protein conformational change: hinged ''lid'' motion of the triosephosphate isomerase loop. *Science,* **249**, 1425–1428.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A,* **32**, 922–923.

Kolinski, A. & Skolnick, J. (1992). Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. *J. Chem. Phys.* **97**, 9412–9426.

Kolinski, A., Godzik, A. & Skolnick, J. (1993). A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **98**, 7420–7432.

Leszczynski, J. F. & Rose, G. D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science,* **234**, 849–855.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1989). Modelling antibody hypervariable loops: a combined algorithm. *Proc. Natl Acad. Sci. USA,* **86**, 9268–9272.

Mattos, C., Petsko, G. A. & Karplus, M. (1994). Analysis of two-residue turns in proteins. *J. Mol. Biol.* **238**, 733–747.

McGarrah, D. B. & Judson, R. S. (1993). Analysis of the genetic algorithm method of molecular conformation determination. *J. Comp. Chem.* **14**, 1385–1395.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Mosimann, S., Meleshko, R. & James, M. N. G. (1995). A critical assessment of comparative molecular modelling of tertiary structures of proteins. *Proteins: Struct. Funct. Genet.* **23**, 301–317.

Moult, J. & James, M. N. G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Struct. Funct. Genet.* **1**, 146–163.

Niefind, K. & Schomburg, D. (1991). Amino acid similarity coefficients for protein modelling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol.* **219**, 481–497.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631–634.

Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng.* **5**, 121–137.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). In *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Reczko, M., Martin, A. C. R., Bohr, H. & Suhai, S. (1995). Prediction of hypervariable CDR-H3 loop structures in antibodies. *Protein Eng.* **8**, 389–395.

Ring, C. S. & Cohen, F. E. (1994). Conformational sampling of loop structures using genetic algorithms. *Isr. J. Chem.* **34**, 245–252.

Ring, C. S, Kneller, D. G., Langridge, R. & Cohen, F. E. (1992). Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* **224**, 685–699.

Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.* **204**, 1019–1029.

Rosenbach, D. & Rosenfeld, R. (1995). Simulations modelling of multiple loops in proteins. *Protein Sci.* **4**, 496–505.

Šali, A. (1995). Modelling mutations and homologous proteins. *Curr. Opin. Struct. Biol.* **6**, 437–451.

Šali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403–428.

Šali, A. & Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3**, 1582–1596.

Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. & Levinthal, C. (1987). Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers,* **26**, 2053–2085.

Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777.

Skolnick, J., Kolinski, A., Brooks, C. L., III, Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence. *Curr. Biol.* **3**, 414–423.

Smith, K. C. & Honig, B. (1994). Evaluation of the conformational free energies of loops in proteins. *Proteins: Struct. Funct. Genet.* **18**, 119–132.

Steipe, B., Plückthun, A. & Huber, R. (1992). Refined crystal structure of a recombinant immunoglobulin domain and a complementarity-determining region 1-grafted mutant. *J. Mol. Biol.* **225**, 739–753.

Sudarsanam, S., DuBose, R. F., March, C. J. & Srinivasan, S. (1995). Modelling protein loops using a ϕ_{i+1}, ψ_i dimer database. *Protein Sci.* **4**, 1412–1420.

Summers, N. L. & Karplus, M. (1990). Modelling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro < > non-Pro mutations. *J. Mol. Biol.* **216**, 991–1016.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Tanner, J. J., Nell, L. J. & McCammon, J. A. (1992). Anti-insulin antibody structure and conformation. II. Molecular dynamics with explicit solvent. *Biopolymers,* **32**, 23–31.

Tramontano, A. & Lesk, A. M. (1992). Common features of the conformations of antigen-binding loops in immunoglobulins and application to modelling loop conformations. *Proteins: Struct. Funct. Genet.* **13**, 231–245.

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

Wlodawer, A., Miller, M., Jakolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. H. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease, *Science,* **245**, 616–621.

Wu, S. & Cygler, M. (1993). Conformation of complementarity determining region L1 loop in murine IgG 1 light chain extends the repertoire of canonical forms. *J. Mol. Biol.* **229**, 597–601.

Zhen, Q. & Kyle, D. J. (1996). Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy

conformational samplings. *Proteins: Struct. Funct. Genet.* **24**, 209–217.
Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1993a). Loop closure via bond scaling and relaxation. *J. Comp. Chem.* **14**, 556–565.
Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1993b). Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci.* **2**, 1242–1248.
Zheng, Q., Rosenfeld, R., DeLisi, C. & Kyle, D. J. (1994). Multiple copy sampling in protein loop modelling: computational efficiency and sensitivity to dihedral angle perturbations. *Protein Sci.* **3**, 493–506.

Edited by P. E. Wright

(Received 12 November 1996; accepted 6 December 1996)