PROTEINS: Structure, Function, and Bioinformatics 54:583–595 (2004)

Efficient Methods for Filtering and Ranking Fragments for the Prediction of Structurally Variable Regions in Proteins

Philipp Heuser, Gerd Wohlfahrt,[†] and Dietmar Schomburg*
University of Cologne, Institute of Biochemistry, Köln, Germany

ABSTRACT **The prediction of protein 3D structures close to insertions and deletions or, more generally, loop prediction, is still one of the major challenges in homology modeling projects. In this article, we developed ranking criteria and selection filters to improve knowledge-based loop predictions. These criteria were developed and optimized for a test data set containing 678 insertions and deletions. The examples are, in principle, predictable from the used loop database with an RMSD < 1 Å and represent realistic modeling situations. Four noncorrelated criteria for the selection of fragments are evaluated. A fast prefilter compares the distance between the anchor groups in the template protein with the stems of the fragments. The RMSD of the anchor groups is used for fitting and ranking of the selected loop candidates. After fitting, repulsive close contacts of loop candidates with the template protein are used for filtering, and fragments with backbone torsion angles, which are unfavorable according to a knowledge-based potential, are eliminated. By the combined application of these filter criteria to the test set, it was possible to increase the percentage of predictions with a global RMSD < 1 Å to over 50% among the first five ranks, with average global RMSD values for the first rank candidate that are between 1.3 and 2.2 Å for different loop lengths. Compared to other examples described in the literature, our large numbers of test cases are not self-predictions, where loops are placed in a protein after a peptide loop has been cut out, but are attempts to predict structural changes that occur in evolution when a protein is affected by insertions and deletions. Proteins 2004;54:583–595.**

Key words: protein loops; homology modeling; insertions and deletions; knowledge-based torsion potential

INTRODUCTION

The recent genome sequencing projects[1–3] have provided sequence information for large numbers of proteins. For a detailed functional characterization of these sequences, in most cases accurate, 3D structural information of the proteins is highly desirable, if not absolutely necessary. But even in the days of high-throughput methods, experimental determination of protein structures by X-ray crystallography or NMR is quite time-consuming. Thus, there is an increasing gap between the number of available protein sequences and experimentally derived protein structures, which makes it even more important to improve the methods for automatic, computer-based prediction of protein 3D structures.

There are basically three approaches to the protein structure prediction problem: de novo prediction, threading, and homology modeling. The reports from CASP4 (Critical Assessment of Techniques for Protein Structure Prediction) give a good overview of these methods and their current state.[4–6] The accuracy, and therefore the applicability, of a homology model depends heavily on the level of sequence identity.[7] The prediction of 3D protein structures by homology modeling works well for protein cores, at least when the sequence identity is above 30%. But there are two steps in the prediction process that cannot be performed satisfactorily yet, and therefore should be considered in further research: (1) the creation of a correct sequence alignment at lower homology levels, and (2) the prediction of the structure of structurally nonconserved loop regions.[6]

Loops typically connect more rigid structural elements on the protein surface and are structurally not well defined. On the other hand, loops often contribute to the function and specificity of proteins,[8] and take part in molecular recognition (e.g., in the hypervariable regions of antibodies,[9] or binding of substrates or DNA).[10] These loop regions frequently differ between the different members of a protein family. Point mutations in these regions often do not affect the overall stability of the protein, which leads to an accumulation of mutations in loop regions.[11] There have been several attempts to classify loops according to their structure,[12,13] but there is little correlation between these structure classes and the sequence, especially of longer loops.[14]

Abbreviations: 3D, three-dimensional; Δ-anchordist, difference between the distance of the template anchor groups and the corresponding distance in the fragment stems; PDB, Protein Data Bank; RMSA, RMSD of the anchor group atoms; RMSD, root-mean-square deviation; RMSL, global RMSD of the loop backbone atoms; TP, knowledge-based torsion potential; vdW, van der Waals.

*Correspondence to: Dietmar Schomburg, University of Cologne, Institute of Biochemistry, Zülpicher Str. 47, D-50674 Köln, Germany. E-mail: d.schomburg@uni-koeln.de

[†]Present address: Orion Pharma, P.O. Box 65, FIN-02101 Espoo, Finland.

Received 12 March 2003; Accepted 9 July 2003

The two basic approaches to the prediction of loop structures are conformational search and knowledge-based methods. Basics for conformational search methods have been published by Moult and James[15] and Bruccoleri and Karplus.[16] A more recent overview of those methods is found in Fiser et al.[17]

The knowledge-based approach basically consists of searching databases with fragments derived from known structures for those that fit template-defined anchor groups. For most cases, a large number of fragments matches the template stems within a certain range of accuracy, so it is essential to apply further criteria to remove unsuitable fragments (cutoff) and to sort the remaining fragments (ranking) to get a near-correct prediction on a high rank.

With increasing loop length, there are limits for both methods. The de novo methods are limited by the calculation time, since it increases exponentially with the loop length. The knowledge-based methods are primarily limited by the completeness of the fragment databases. Even with a decreasing number of newly found folds, the data banks, especially for fragments longer than 8 residues, are still incomplete.[18,19] On the other hand, knowledge-based methods do have the advantage that they guarantee the prediction of frequently occurring (i.e., energetically favored fragments).

After modeling the conserved core of a protein structure, the first step of predicting the structurally variable regions is to position the anchor groups for the loop.[19,20] After that, an appropriate fragment has to be selected from a data bank. In a previous article,[20] we were able to show that anchor-group positioning and loop selection are two independent steps in loop prediction that can be treated independently. In that article, we also suggested criteria for optimal anchor-group positioning. In this article, we are now identifying filters for the loop selection, assuming that good anchor-group positions have been identified in the previous step.

With a reasonable completeness of databases for shorter loops and simple selection criteria (usually RMSD of anchor groups), most approaches for knowledge-based loop prediction do find suitable loops, but their ranking is problematic because of the absence of fast and reliable scoring functions. Possible loop filter or scoring functions can be classified in different groups, being either suitable for removal of inappropriate candidates or even allowing a ranking of the fragments.

Simple and fast prefiltering makes use of the geometrical features of the loop anchor groups, like distances, angles, and vectors, and compares them to the stems in the template.[21] The RMSD fit of anchor groups is most frequently used for positioning but also for ranking,[22] as similarity of anchor groups often correlates with the similarity of the whole fragment. Other methods analyze the internal energy of fragments, either knowledge-based in respect to their sequence[22] or by usage of empirical force field energies.[17,23] Other methods evaluate the interactions of inserted fragments with their environment in terms of nonbonding repulsions, force field energy,[23] or propensity for that environment.[24]

In this report, four different filter criteria are evaluated: A fast prefilter for the selection of fragments from the database uses the difference of the distances between the two anchor groups in the protein template and the distance of the stems of the loops.

- Fitting of fragments and sorting of the selected fragments is performed by the use of the RMSD of the anchor groups.
- After fitting, two postfilters to remove unsuitable fragments and to enrich good fragments in the top ranks are applied:
  - Clashes of a loop candidate with the protein are analyzed.
  - Fragments with torsion angles that are unfavorable, according to a knowledge-based potential, are eliminated.

By the application of these filter criteria, it was possible to increase the percentage of successful predictions significantly. In this respect, a successful prediction is defined as such when a loop with a global RMSD < 1 Å can be found among the first ranks of the prediction for regions with insertions or deletions in the sequence alignment. This reflects the situation in real homology modeling problems.

MATERIALS AND METHODS

Test Set Generation

The test data sets used in this article are based on results of Lessel and Schomburg[19] and Wohlfahrt et al.[20] The PDB (February 1998 release) has been searched for homologous protein pairs sharing less than 50% sequence identity and differing in short segments defined by a structure–alignment algorithm,[25] as described elsewhere.[19] These protein pairs were used as template-target pairs. The loop fragments were assigned onto the unmodified template structure and then compared to the target structure. So far, 544 examples for deletions and 550 for insertions have been created. For each insertion or deletion example, all possible anchor group combinations for fragments of 3–12 amino acids length were tested.[20] This resulted in 44,957 possible anchor group combinations, which have been analyzed for the highest possible prediction quality assuming that an optimal selection algorithm exists.[20]

In previous articles, we were able to show that the result of a loop prediction strongly depends on the optimal selection of anchor groups[19] and to identify means to improve anchor group selection.[20] Therefore, we decided to use an optimal anchor group placement for this analysis (i.e., for each loop, we took those positions for the anchor groups where the database held the loop with the best possible RMSL for this target).

After the selection of the anchor group combination that gives the lowest RMSL for each example, and, of those, the ones that allow a prediction with an RMSL < 1 Å, 321 examples for insertions and 357 for deletions are left. This preselection has been performed in order to guarantee that at least one good fragment for each example exists in the

database that could be detected by the filter to be evaluated.

The test data sets are available at ftp://ftp.uni-koeln.de/institute/biochemie/pub/loop_testdata/.

Fragment Data Bank

The fragment data bank is based on all X-ray structures in the February 1998 release of the PDB with resolutions of smaller or equal to 2.0 Å and sequence identities of less than 95% determined by the Smith–Waterman algorithm[26] using standard gap penalties. After fitting N-, C_α-, and C-carbonyl atoms of two ending residues, we eliminated fragments showing RMSD values below 0.25 Å compared to other fragments considering all backbone atoms. The RMSD fit was performed following the procedure by Diamond.[27] The limit of 0.25 Å was chosen according to the estimated standard error in X-ray analysis. The data bank contains about 1.2 million fragments from 3 to 12 amino acids in length.[19]

Analysis of Prediction Quality

To analyze the quality of a prediction, we calculated the RMSL [i.e., the RMSD of all loop backbone atoms (N, Cα, carbonyl-C, and O) of the suggested fragment compared to the target loop]. The RMSL for the loops was derived by overlay of the complete structures of template and target (global RMSD) using the method of Lessel and Schomburg,[25] and not simply by comparison of the short loop fragments (local RMSD). This is necessary because it is not sufficient to determine an RMSD value solely between the inserted loop and the target loop, since an incorrect orientation with respect to the target protein (of a correct loop conformation) would not be noticed. For the RMSL, a limit of 1 Å was chosen as a criterion for successful predictions, as this accuracy is useful for most applications, even the study of catalytic mechanisms or improvement of ligands, whereas an RMSD of 1.5 Å is usually still useful for virtual screening or molecular replacement in X-ray crystallography.[7]

Filtering and Ranking of Fragments

For the preselection of appropriate fragments from the loop data bank and the fitting procedure, a modified version of a method developed by Lessel and Schomburg is used.[19]

Anchor group distance

In a first step, all fragments of the data bank with a defined "interbond" distance similar to the template (Δ-anchordist) are selected. This is defined as the distance between the middle of the bond between Cα- and carbonyl-C-atom of the N-terminal anchoring residue and the middle of the bond between N and Cα of the C-terminal anchoring residue.

RMSD of anchor groups

The subsequent criterion used for fitting and ranking of fragments is the RMSD of the anchor groups (RMSA). The fragments from the data bank are fitted onto the anchoring groups of the template protein, considering N-, Cα-, and C-atoms of both anchoring residues by using the RMS fit procedure of Diamond.[27]

Steric interactions between loop fragments and the template protein

In order to evaluate the significance of certain atoms and the influence of their radii for the filtering process, we tested three different types of collision filters, which are supposed to eliminate fragments exhibiting unfavorable interactions with the rest of the protein.

The first version of this filter analyzes only the Cα atoms of the candidate loop for steric overlaps with all atoms of the template (Cα-only filter). To determine the optimal threshold for this collision filter, we initially used a Cα-radius of 1.9 Å and assumed a common radius of 1.6 Å for the atoms of the protein target. Therefore, a clash was defined as an interatomic distance of less than 3.5 Å, and this value was reduced stepwise to 1.5 Å (C—C bond length) in order to find the optimal threshold. For the second collision filter (coll-2), all backbone-atoms and Cβ of the fragments were considered using the same radius for all loop atoms (N, Cα, carbonyl-C, O, and Cβ). Cβ atoms were generated by applying the average values for relative Cβ coordinates, which were derived from all structures in the PDB (February 1998). For this filter, a range of interatomic distances from 1.0 to 3.5 Å was evaluated.

A third filter (coll-3) uses explicit values for the radius of each atom type (Cα: 1.9 Å, N: 1.71 Å, C: 1.75 Å, O: 1.49 Å, coll3-a-Cβ: 2.2 Å, coll3-b-Cβ: 2.6 Å, coll3-c-Cβ: 3.0 Å). In order to mimic larger side-chains without the necessity to predict their explicit conformation also, larger radii were tested for Cβ. In addition, we also analyzed whether it is advantageous to allow limited clashes with the template in order to compensate for inaccurate positioning of template side-chains or loop atom.

Knowledge-based torsion potentials

Torsion potentials are computed from the relative propensities of amino acid residues for defined backbone torsion angles.[28] Therefore, the probabilities are converted into free energies by assuming that the structural states follow a Boltzmann-type distribution.[29–31]

The knowledge-based potentials used in this article are based on a set of 903 protein structures selected from the PDB (January 2001), with a sequence identity $<30\%$, determined with a resolution <2.1 Å and an R-factor <0.21.[32] All backbone torsion angles were calculated by the DSSP[33] program and sampled in 1° bins. The distribution of discrete values for ϕ and ψ for each amino acid was smoothed by the function

$$\Delta P_i = \frac{1}{2\pi\sigma^2}\, e\, \frac{\Delta\varphi^2 + \Delta\psi^2}{\sigma^2},$$

in order to take experimental inaccuracy and sparse data into account. For the standard deviation σ, we used 14. The different frequencies of occurrence for different amino acid types have been corrected by normalization before an

TABLE I. Number of Protein Pairs and Fragments, and Number and Percentage of Fragments With an RMSL < 1 Å for Insertions and Deletions for Different Loop Lengths After application of Δ-Anchordist and RMSA

	Insertions				Deletions			
Loop length	No. of protein pairs	No. of fragments	Fragments with RMSL < 1 Å		No. of protein pairs	No. of fragments	Fragments with RMSL < 1 Å	
			No.	%			No.	%
3	6	39,476	3067	7.77	16	175,275	12246	6.99
4	11	240,581	6087	2.53	24	832,560	29464	3.54
5	17	684,237	7163	1.05	41	1,607,657	9182	0.57
6	26	884,128	2132	0.24	52	2,068,884	8742	0.42
7	41	1,329,331	1837	0.14	53	1,827,928	4323	0.24
8	17	433,909	643	0.15	40	1,135,588	6421	0.57
9	34	926,486	854	0.09	36	1,183,040	1196	0.10
10	46	1,409,187	2701	0.19	47	1,300,514	3081	0.24
11	29	620,510	1115	0.18	48	1,284,527	4709	0.37
12	94	2,308,252	737	0.03	—	—	—	—

TABLE II. Number of Protein Pairs and Fragments, and Number and Percentage of Fragments With an RMSL < 1 Å for Insertions and Deletions for Different Gap Lengths After application of Δ-Anchordist and RMSA

	Insertions				Deletions			
Gap length	No. of protein pairs	No. of fragments	Fragments with RMSL < 1 Å		No. of protein pairs	No. of fragments	Fragments with RMSL < 1 Å	
			No.	%			No.	%
1	203	6,040,644	23,458	0.39	213	6,858,897	62,067	0.90
2	66	1,728,964	1806	0.10	72	2,421,190	8929	0.37
3	22	573,775	922	0.16	26	849,052	4823	0.57
4	14	244,902	128	0.05	18	471,711	1713	0.36
5	10	196,168	16	0.01	17	470,318	794	0.17
6	4	63,068	4	0.01	8	285,020	925	0.32
7	2	28,576	2	0.01	2	49,333	2	0.00

"averaged" amino acid (aa_{av}) was created as reference state for the inverse Boltzmann equation:

$$\Delta G = -kT \log P(aa)/P(aa_{av})$$

By applying the Boltzmann equation to each of the 20 amino acid types, we generated a matrix of "free energy values" for each amino acid and each pair of integer torsion angles. A second torsion potential was generated in a similar way, but in this case, only dihedral angles of those amino acids were used for the calculation of the potential, which are located in loop regions (i.e., having no helical or sheet structure according to DSSP). The torsion potentials are used to remove fragments, which show high internal energies for a certain amino acid sequence.

Combination of filters

The collision filters take the interaction between the loop and the rest of the protein into account, whereas the torsion-potential filter reflects internal properties of the loop. As both properties are assumed to be independent, a combination of the best threshold for each filter should lead to further improvement of accuracy.

RESULTS

Composition of the Test Data Set

The test set contains 321 examples for insertions and 357 examples for deletions whose fragment-length distribution is shown in Table I. It shows the amount of analyzed template-target pairs, the number of fragments tested, and the percentage of fragments with RMSL < 1 Å. After applying the Δ-anchordist and RMSA filter, there is an average of 28,000 fragments per insertion and 32,000 for the deletion examples.

Because of the preselection of optimal anchor groups, the number of protein pairs with short loops (3 and 4) is quite low. On the other hand, due to geometrical constraints, the fraction of fragments with a low RMSL after sorting by RMSA is rather high in this group compared to longer fragments. This increases the chance to find a good loop by chance for loop length 3 compared to loop length 11, for example, by one to two orders of magnitude.

As becomes obvious in Table II, short gaps are overrepresented in the data set. This reflects the distribution found in protein evolution where point mutations or short insertions/deletions appear far more often than mutations of complete fragments.[11,34] Similar to the situation at different loop length, the probability to find a good loop by chance in the fragment database decreases from gap length 1–7 by about two orders of magnitude.

Anchor Group Distance

For the selection of appropriate fragments, the difference of the fragment anchor group distance compared to

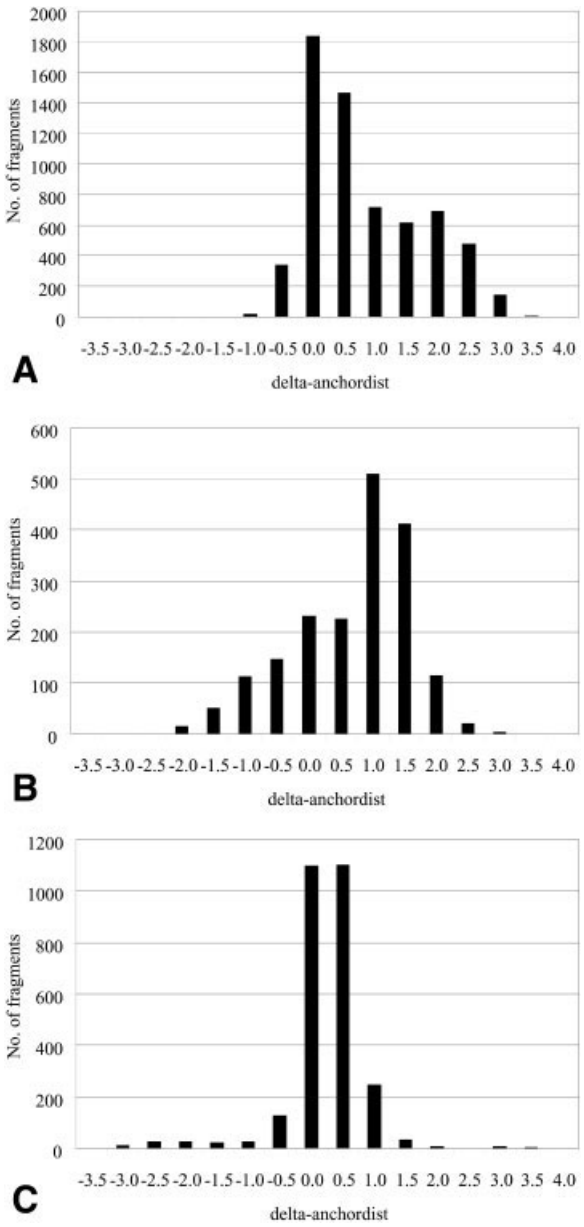

Fig. 1. All fragments with RMSL < 1 Å with respect to their Δ-anchordist. The distribution for (**A**) 4 amino acid loops, (**B**) 7 amino acid loops, and (**C**) 10 amino acid loops.

the template anchor group distance (Δ-anchordist) is a fast, geometry-based preselection criterion, which is independent of any coordinate transformation.

Figure 1 shows the relation between Δ-anchordist and RMSL for some selected loop lengths. It becomes obvious from Figure 1 that fitting loops can be found between −3 Å and +4 Å. There are significant differences for different loop lengths. For short loops, the maximum is around +1 Å, with a clear majority of fragments in the positive range. This reflects the low compressibility of short fragments. Medium-length fragments are more symmetrically distributed around small positive values, whereas the distribution is much narrower for longer fragments.

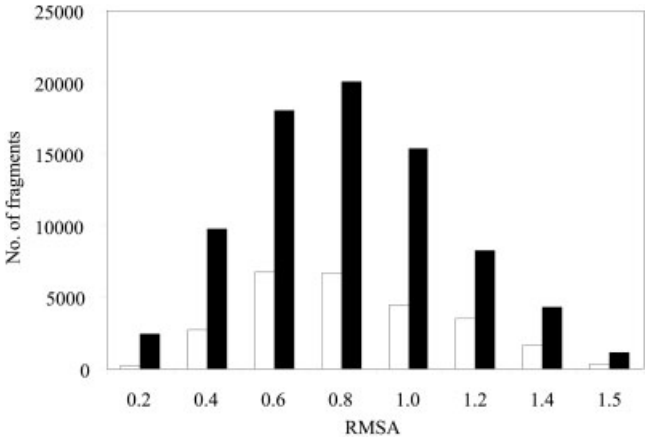

Fig. 2. Number of fragments with RMSL < 1 Å for the insertion (black) and deletion (white) examples for different RMSA values of the anchor groups.

Since only very few fragments with an RMSL < 1 Å can be found when the difference of anchor groups is larger, and on the other hand, the number of fragments dramatically increases when the range of allowed values is extended, a range between −2 Å and +3 Å was chosen for Δ-anchordist as a threshold for preselection. This range covers 99.8% of all good anchor groups.

RMSD of Anchor Groups (RMSA)

As the exact position of the fragment in the target protein is determined by the RMSD-fit of the anchor groups (RMSA),[19] we also evaluated this criterion in respect to its potential to rank those fragments.

Figure 2 shows the number of fragments with an RMSL < 1 Å at a certain RMSA value. A cutoff of 1 Å, which was originally used,[19] would remove about 27% of fragments that would still allow a successful prediction. Fragments with a low RMSL (<1 Å) can be found up to RMSA values of 1.5 Å. This larger cutoff was used in this work. The application of the optimized value for the Δ-anchordist threshold (−2 Å to +3 Å) and for the RMSA cutoff of 1.5 Å meant that for 97% (insertions) and 98% (deletions) of all targets, there is still at least one good loop in the loop data base.

Figure 3 shows the distribution of RMSA values in the loops with specific RMSL values. The first three columns represent good fragments with low RMSL values. About 21% of fragments with RMSL between 0.5 and 1.0 Å have an RMSA > 1 Å. At the same time, Figure 3 shows that an increase of the RMSA cutoff leads to a strong increase of fragments with higher RMSL values. But since the overall fraction of good fragments is rather low for most examples (Table I), it is important to avoid losing too many good fragments by a strict cutoff. Therefore, the 1.5 Å cutoff was used. The corresponding figures for the deletion examples are quite similar and lead to the same conclusion (data not shown).

After selection and RMS-fit of appropriate fragments, all possible loops are sorted by RMSA. Interestingly, the criteria applied so far work better for deletions than for

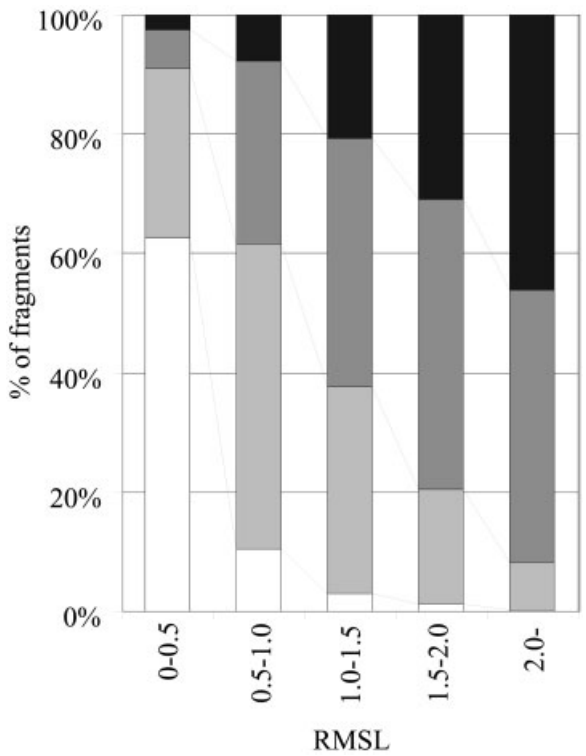

Fig. 3. Distribution of RMSA values in fractions of loops with decreasing quality. It is obvious that RMSA and RMSL values are clearly correlated (RMSA values, white bar, 0–0.4 Å; light gray 0.4–0.8 Å; dark gray, 0.8–1.2 Å; black, 1.2–1.5 Å).

insertions. However, for both data sets, a loop with a low RMSL can be placed on the first rank in 21% of all insertion and 22% of all deletion examples, and for about 37% and 42%, respectively, among the top five.

Collision Filters

A filter using the distance between the loop Cα and all template atoms was tested for its ability to differentiate between fragments making stabilizing contacts or unfavorable clashes. A test with a distance cutoff of 3.6 Å was unsuccessful, decreasing the percentage of examples with good fragments on the first rank compared to only RMSA sorting. Consequently, we reduced this distance in steps of 0.2 Å. The best results were obtained with a distance cutoff of 2.8 Å, allowing no closer distances. If one allows one contact below the cutoff distance, a slight, but statistically not significant deterioration of the results is observed.

By removing those fragments where a Cα atom comes closer than 2.8 Å to any other atom of the protein, the percentage of examples with a fragment with RMSL < 1 Å on the first rank could be increased from 20 to 27% for insertions and from 22 to 27% for deletions. For the top 10 and the top 20, the increase was from about 41 to 49% and 45 to 54%, respectively, for the insertion examples, but only 49 to 52% and 58 to 59%, respectively, for the deletions.

The application of this filter removed 77% of all fragments for insertions and 69% for deletions. That also a limited amount of good candidates were removed becomes obvious from the fact that 11 examples (3%) of protein template/target pairs did not have any fragment left.

For the Coll-2 filter, which takes all loop backbone atoms into account and allows one "clash," the best results were obtained using radii between 1.3 Å and 1.7 Å for all atoms. As the use of radial-symmetric Cβ-centered side-chains is indeed a strong simplification and could therefore lead to the removal of good fragments, we tested whether it is advantageous to allow a small number of clashes. Inaccurate positioning of loops might also lead to close contacts for good fragments. Compared to allowing one "forbidden" contact, the prediction quality was dramatically lowered by the allowance of zero or two clashes.

The Coll-2 filter with the optimized parameters was not significantly better than the improvement of the Cα-only filter for the insertions, whereas deletions could be predicted slightly better. On the other hand, the calculation time is much longer compared to the Cα-only filter, since all backbone-atoms, including the Cβ-atoms, are considered.

The results for the Coll-3b filter, which takes into account all loop backbone atoms with an explicit individual atomic radius and Cβ with a radius of 2.6 Å, are shown in Table III. Also, for this filter, the best results were obtained when one clash with the rest of the protein was allowed. After applying this filter, 30% of the tested examples had a good fragment on the first place, which corresponds to an increase by 9% for insertions and 8% for deletions.

It is noteworthy that this filter works better for longer fragments. An increase from 33 to 50% is observed for examples with a 12-amino acid-long fragment with an RMSL < 1 Å prediction on the first place. (The poor result for short loops with a length of 3 or 4 amino acids is misleading, since there are only very few examples; compare Table I). This filter removed 83% of all fragments for insertion examples (deletions 78%). For 5 of the insertion (1.5%) and 8 of the deletion (2%) examples, no fragments were left that satisfy the preselection criteria Δ-anchordist and RMSA.

Torsion Angle Potential Filters

The torsion angle potential was tested to get an estimation of how well the loop geometry from the candidate loop fits to the given target sequence. The attempt to use the knowledge-based potentials to rank fragments was unsuccessful and led to lower quality prediction results (data not shown). Therefore, we decided to use it as a filter instead and to define a cutoff value beyond which fragments are not considered to be fitting candidates because of their low sequence–structure compatibility. In order to determine a threshold for the torsion-potential filter, we analyzed the fragments according to their RMSL and their energy derived from the knowledge-based potential. Figure 4 shows that most fragments with low RMSL values do have energy values below 1 kcal/mol per amino acid. Consequently, we tested 3, 2, 1, 0, and −1 kcal/mol as

TABLE III. Improvement of Predictions by Application of the Coll-3b-Filter to (A) Insertion and (B) Deletion Examples for Different Fragment Lengths

(A)	Top			Top 5			Top 10			Top 20		
Length	No.	%	+	No.	%	+	No.	%	+	No.	%	+
All	96	30	9%	140	44	6%	157	49	7%	176	55	10%
3	1	17	17%	2	33	−33%	3	50	−17%	4	67	0%
4	2	18	0%	3	27	0%	3	27	0%	4	36	0%
5	3	18	0%	4	24	6%	4	24	−6%	5	29	0%
6	2	9	1%	6	26	7%	7	30	4%	7	30	0%
7	6	15	5%	12	29	2%	14	34	7%	17	41	10%
8	10	59	12%	11	65	0%	11	65	0%	12	71	6%
9	5	15	6%	10	29	9%	13	38	15%	15	44	15%
10	11	25	8%	20	45	17%	22	50	17%	26	59	24%
11	9	31	10%	13	45	3%	17	59	14%	19	66	21%
12	47	50	17%	59	63	9%	63	67	7%	67	71	7%

(B)	Top			Top 5			Top 10			Top 20		
Length	No.	%	+	No.	%	+	No.	%	+	No.	%	+
All	108	30	8%	192	54	12%	207	58	9%	228	64	6%
3	1	6	0%	7	44	6%	9	56	19%	11	69	13%
4	6	25	8%	10	42	8%	12	50	0%	12	50	−4%
5	7	17	0%	16	40	16%	19	47	16%	21	52	9%
6	15	29	10%	31	60	8%	32	62	4%	35	67	6%
7	14	27	14%	23	44	8%	26	50	3%	29	56	1%
8	9	23	13%	16	41	16%	17	44	11%	23	59	11%
9	13	38	8%	22	65	17%	23	68	15%	23	68	9%
10	18	39	5%	29	63	8%	31	67	4%	35	76	10%
11	25	54	15%	38	83	28%	38	83	24%	39	85	14%

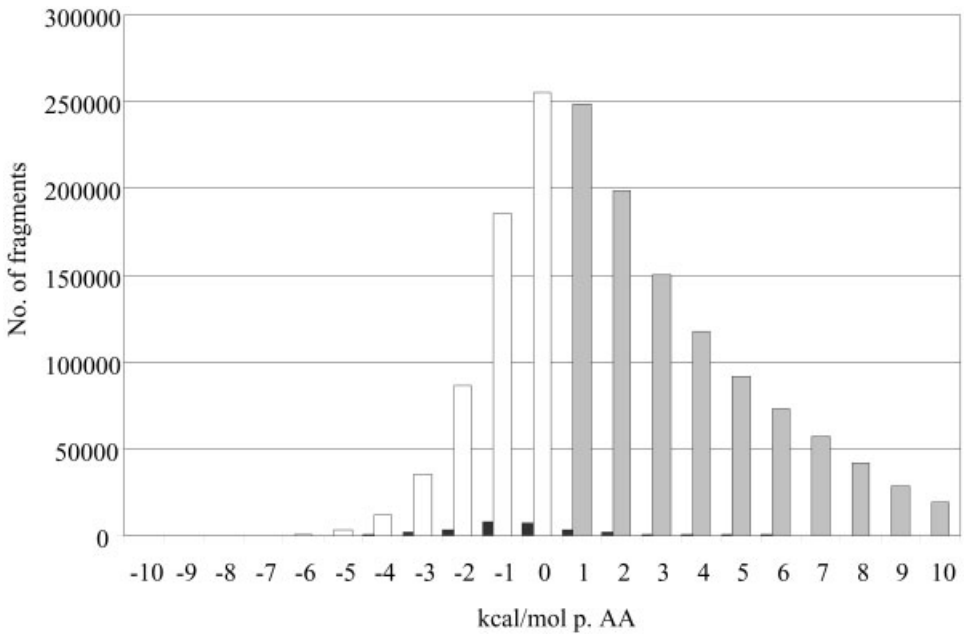

Fig. 4. Distribution of fragments with RMSL < 1 Å (black) and with RMSL > 1 Å (white and gray) according to their torsion potential energy values in kcal/mol per amino acid.

possible cutoff values. There are very few fragments having an average energy value below −2 kcal/mol per amino acid.

The improvements achieved by thresholds of 1, 0, and −1 kcal/mol per amino acid are rather similar. All of them increase the percentage of examples with a good fragment (RMSL < 1 Å) on the first place by about 9% for the insertion and 8% points for the deletion examples. Table IV shows the results with two different cutoff values per amino acid. About 80% of all insertion fragments are removed by using this cutoff, in contrast only 64% of all fragments removed for the deletions.

TABLE IV. Improvement of the Prediction Quality by the Torsion-Potential Filter Using A Cut/off of −1 kcal/mol or +1 kcal/mol per Amino Acid (AA)

	−1kcal/mol per AA			1kcal/mol per AA		
	No.	%	+	No.	%	+
Insertions						
Top	96	30	9%	91	28	7%
Top 5	139	43	6%	137	43	5%
Top 10	155	48	7%	155	48	7%
Top 20	169	53	8%	171	53	8%
Deletions						
Top	109	31	9%	97	27	5%
Top 5	170	48	6%	175	49	8%
Top 10	192	54	5%	197	55	6%
Top 20	209	59	1%	235	66	9%

For the top 5 fragments, the results obtained with a threshold of 1 kcal/mol per amino acid are slightly better than for the other cutoff values (Table IV). This value was chosen for further examinations as a compromise between enrichment of good fragments and keeping many good fragments in order to predict as many cases as possible.

Improvements are relatively small for loop lengths 3–10, but rather large (>10 percentage points) for loops with 11 and 12 amino acids, which represent a large part of our test cases (data not shown).

The results described above were obtained using the torsion potential derived from all amino acids. The use of the loop-only torsion potential produced similar results. The differences are about 1%, with none of the torsion potentials performing significantly better in any certain group of the results.

Combination of the Filters

The dark bars in Figure 5 show the improvement of the accuracy after the application of both postfilters. For each filter-type, collision- and torsion-potential filter, we selected the previously determined optimal parameters (see above; i.e., the collision-filter coll-3b and the torsion-potential filter with a cutoff of 1 kcal/mol per amino acid). It is obvious that both filters eliminate different unsuitable fragments. There is almost no overlap, as the increase obtained by the combination nearly equals the sum of the improvements of each single filter.

Percentage of successful predictions

After application of both postfilters, 36% of all insertion examples have a fragment with RMSL < 1 Å on the first place, and almost 50% have one in the top 5 ranks. Finally, 94% of all fragments initially chosen by Δ-anchordist and RMSA are removed. The average RMSL of all remaining fragments decreases hereby from 5.0 to 3.4 Å.

Remarkably, the enrichment of fragments with low RMSL is much higher than for other fragments. Of fragments with RMSL < 0.5 Å, 79% are preserved, whereas 95% of the more than 10 million fragments with RMSL > 1.5 Å are rejected by the postfilters (Table V). For about 20% of the insertion examples, all good fragments that would allow a prediction with RMSL < 1 Å, were removed by one of the filters, but only 2% of the examples became completely unpredictable because all fragments were removed.

For the first rank of the deletion examples, the combination increases the percentage of examples with good fragments even more than the sum of the results for each filter would suggest. While the sum of the improvement by both filters is 13%, their combination leads to an improvement by 15%. Thus, finally, 37% of all deletion examples do have a fragment with RMSL < 1 Å on the first rank, and 57% have a good fragment among the top 5. Only 9 examples (2.5%) are left without any candidate loops, whereas 92% of all RMSA-selected fragments are removed by the two filters.

Quality of template loops on first position

The average RMSL of fragments ranked on position one in the insertion test set decreases from 2.8 Å after sorting by RMSA to 1.9Å after application of both postfilters. After sorting by RMSA only, a clear dependency of the accuracy on the fragment length is seen. Application of the postfilters, which work better for longer fragments, removes the loop-length dependency for the RMSL of the loop on position one, but not for the RMSL of the loops on position 1–5. This shows that there are only very few good fragments for long loops in the database, but after postfiltering, they are mostly found on the first rank.

Applying the combination of the best filters results in an average calculation time per loop of about 4.5 min on an Ultra Sparc III 750Mhz processor.

Comparison with Other Methods

A data set by van Vlijmen and Karplus[18] is used in several publications to compare the prediction quality of different approaches. It contains 14 protein loops, which—unlike our examples—represent "self-prediction" examples, as template and target are the same here and the gap length is consequently zero. To maintain the comparability with the other publications, we used the same anchor groups as specified in that article.[18] These "self-predictions" are of (significantly) lower complexity than the examples discussed in this paper so far.

Table VI shows the results of loop-prediction methods from literature[17,18,21,22] in comparison to our method applying all filters. For 7 of the 14 examples, the original loop was found by the RMSA criterion, which results in an RMSL of 0 Å, and for four examples, a fragment from a protein with more than 50% sequence identity to the target was found on the first rank. These trivial solutions have been removed and the best fragment from the first five ranks was taken instead (examples marked with *). In one case, the application of the torsion-potential filter would have removed the original loop. The RMSL values for our prediction (Table VI) are calculated as described in the Methods section, which means that they are global RMSD values. It is possible to predict 10 out of 14

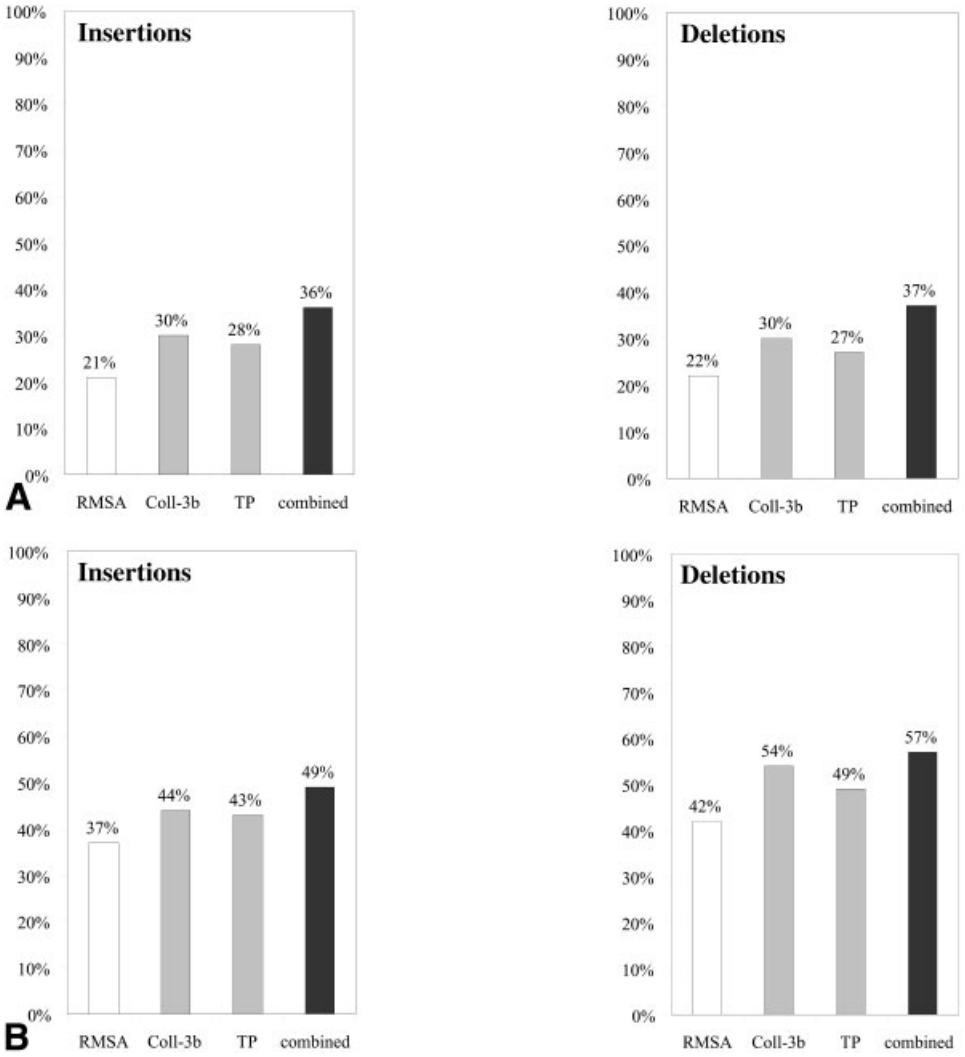

Fig. 5. (**A**) Fraction of insertion and deletion examples having fragments with RMSL < 1 Å on the first place. The first bar shows the results after RMSA sorting, the second additional application of the coll-3b-filter, the third after application of the 1 kcal/mol torsion-potential filter, and the last after application of all filters. (**B**) Fraction of insertion and deletion samples having fragments with RMSL < 1 Å on the first five places. The first bar shows the results after RMSA sorting, the second additional application of the coll-3b-filter, the third after application of the 1 kcal/mol torsion-potential filter, and the last after application of all filters.

TABLE V. Total Number, Fraction, and Enrichment of Good Fragments for Different RMSL Values Before and After Application of the Postfilters (Insertions)

	Before postfilter		After postfilter		
RMSL	No.	%	No.	%	Enrichment
< 0.5	353	0.003	279	0.050	16.7
< 1.0	27,270	0.260	10,921	2.000	7.7
< 1.5	128,745	1.200	29,097	5.300	4.4
> 1.5	10,285,383	98.500	509,281	92.600	0.94
Total	10,441,751	100	548,948	5	

examples with RMSL values below 1 Å. Only one example (2fbj, an IgA Fab fragment) could not be predicted satisfactorily after the original loop has been discarded as a solution.

For the 61 examples of loops with the length 8 from Xiang et al.,[23] we obtained comparable results. For 46 cases, the original loop was found in our fragment database. After removing those, an average RMSL of 1.51 Å using the second-best fragments has been obtained, which is comparable to 1.46 Å obtained by Xiang et al. using colony energies. For this test set, only in two cases could no solution be found after application of our postfilters.

DISCUSSION

Test Data Set

All loops in our test set are bridging gaps in a correct, structure-based alignment, and the fragments are fitted onto a template and then compared to the target structures. It has been shown earlier[19,20] that the correct RMS-fit of fragments into the target structure is relatively easy to perform. These "self-predictions" in fact do not

TABLE VI. Comparison of Average RMSL for Loops on the First Rank for Different Methods (Self-Predictions) [Å]

							Present method	
PDB			Vlijmen/Karplus (1997)	Fiser et al. (2000)[a]	CODA (2001)	PETRA (2000)	RMSL	From PDB (Seq. ID)
3dfr	20–23	4	2.6	0.8	0.4	1.7	0.4*	1DYR (23.1%)
3dfr	89–93	5	1.6	0.9	0.6	1.2	0.8(*)[b]	3BTO (3.1%)
3dfr	120–124	5	0.5	0.3	0.7	1.2	0.4*	2CHS (3.6%)
3blm	164–168	5	0.8	1.2	0.2	1.8	1.3	8DFR (28.5)
8abp	203–208	6	0.3	0.4	0.8	2.5	0.4	1NSY (6.8%)
3grs	83–89	7	4.6	2.0	1.4	2.0	1.0*	2CYP (5.6%)
5cpa	231–237	7	2.1	0.4	0.2	1.3	0.4*	1CCA (7.5%)
2fb4	H26–H32	7	1.6	1.0	0.4	1.9	0.2	8FAB (50.0%)
2fbj	H100–H106	7	0.5	4.2	1.4	3.2	3.4*	1WBA (7.8%)
8tln	E32–E38	7	3.7	1.0	1.9	—	2.4*	1OSP (1.8%)
2apr	76–83	8	5.2	1.3	2.2	2.6	0.4*	1STE (9.9%)
2act	198–205	8	1.6	2.0	3.1	1.5	0.6*	1MEM (41%)
8tln	E248–E255	8	1.8	0.3	1.8	—	0.9*	1VID (2.8%)
3sgb	E199–E211	9	1.8	0.2	—	—	1.1*	1DAN (4.2%)
Average			2.0	1.1	1.2	1.9	1.0	

[a]Lowest energy prediction (global*).
[b]Original fragment removed by torsion potential filter.
*Instead of the original fragment or a fragment from a protein with more than 50% sequence identity to the target protein, which has been found in our database, the best prediction from the first five ranks is used here.

represent realistic homology-modeling tasks. Method development based on those test sets might even be misleading (e.g., a nonoptimal backbone of the template or wrong side-chain orientations have to be considered when selecting a loop in a real case). Therefore, we propose the use of test sets with pairs of corresponding template and target structures.

In order to allow this analysis, we had to take only those template/target protein pairs into consideration where at least one good candidate loop was in the database. As this was the case only for about two thirds of all templates, the absolute values of the positive predictions have to be multiplied by this factor if one is interested in the general probability for successful prediction of a loop. On the other hand, our criterion for a good loop (RMSL $\leq$ 1 Å) is very strict, and the derived criteria for loop selection are certainly not affected by this preselection of template-target examples.

Anchor Group Distance

The distance between anchor groups of the loops is a fast criterion, as it can be precalculated for the whole fragment database, and it is independent of any coordinate transformation. The number of fragments to be fitted on the target proteins can be substantially reduced prior to application of more CPU-time-consuming filters. Interestingly, even fragments with anchor group distance deviations of about 3 Å compared to the distance in the template anchor groups can still give overall RMSL values below 1 Å. Therefore, cutoffs at -2 Å and $+3$ Å represent a good compromise between reduction of the number of fragments and preservation of suitable fragments for a good prediction.

RMSA

According to our analysis, a cutoff of 1.5 Å seems to be the optimal RMSA value for our strict quality criterion (RMSL $<$ 1 Å). The RMSA cutoff should certainly be above 1 Å in order to obtain a sufficient coverage.

Compared to other filter criteria, the RMSA shows the highest correlation with the RMSL (Fig. 3), which makes it useful not only as a cutoff but also as a sorting criterion. Ranking by RMSA works especially well for short fragments. In these cases, the anchor group geometry has an especially strong influence on the overall structure because of lower conformational freedom compared to longer fragments. For longer fragments, additional postfilters become more important. The combination of Δ-anchordist and RMSA works better for deletions than for insertions.

Collision Filters

A number of additional factors are available that give indications whether a particular peptide loop from the database is a good candidate for the target protein. In homology modeling, neither the side-chains of the candidate loops nor the target protein are wholly defined at this stage. Therefore, we evaluated simple methods, which are supposed to be robust enough to overcome this problem. A filter using the distance between loop Cα and all template atoms was tested for its ability to differentiate between favorable or repulsive contacts between loop and target protein. This Cα-only filter is comparable in performance to more elaborate filters using explicit descriptions of all loop atoms. This result can be explained by the lower sensitivity of the Cα-only filter for wrongly positioned side-chains in the target, which makes it useful for template-target pairs with low homology.

Whereas the improvement of prediction quality after application of the Coll-3b filter compared to the Cα-only filter is not significant, this filter provides a higher enrichment of good fragments, as the number of remaining fragment candidates is about 50% smaller (66% for deletions) and at the same time, less fragment test cases are left without any solution. This should provide a better prediction quality at higher computational cost.

Predictions for longer fragments, whose ranking by RMSA is less accurate, benefit from the application of collision filters to a larger extent (Table III). Interestingly, the improvement of the predictions is higher for insertions than for deletions. A possible explanation is that a newly introduced insertion fragment occupies more space than the original loop in the template, which more frequently generates steric clashes, whereas unfavorable contacts with the rest of the protein are less likely for deletions.

Knowledge-Based Torsion Potential

This type of torsion potential is only useful as a cutoff filter for the elimination of fragments with an unfavorable sequence/conformation fit, but not as a ranking criterion. Internal energies are only weakly correlated to RMSL, as the conformation of the loop is obviously determined by the interaction with the rest of the protein, and the energy cost for conformational distortions is small. Therefore, we decided to define a cutoff value beyond which fragments are not considered because of their low sequence–structure compatibility. This improves the accuracy of prediction for long fragments significantly.

In contrast to other methods described in the literature, which usually define some preferred regions in the Ramachandran plot (e.g., Deane and Blundell[21]), we use a much finer classification of torsion angles applying a 1° grid. The advantage of this approach is a higher sensibility of this energy function.[28]

We could not see a significant difference between the prediction quality using a torsion potential derived from all amino acids or using one derived from loop regions. More refined cutoff values and larger test sets might reveal the benefits of different torsion potentials.

CONCLUSIONS

The method presented here consists of relatively fast algorithms, which do not depend on the prediction of amino acid side-chains, as the advantage of more accurate methods might be compensated by the inaccurate environment of the template structure.

RMSA and Δ-anchordist take the geometry of the anchor groups into account, which determines the rest of the loop structure, at least to a certain extent. Collision filters, on the other hand, consider the protein environment and torsion potentials the internal energy of a fragment. From this difference one would expect that those criteria are rather orthogonal in their influence on the results. Indeed, we could show that they eliminate different fragments and that their positive effects on the prediction quality is additive. The only useful ranking criterion out of those tested is the RMSA value. The other criteria are useful as cutoff filters due to their weaker correlation with the RMSL value or because of their more qualitative nature (e.g., the number of collisions with the template). The enrichment of good fragments by combination of different postfilters and decrease of the average RMSL of the remaining fragments indicate that especially poorly performing fragments are removed by the filters.

Several reports in the literature[17,21,22] describe a strong dependency of the prediction quality on the loop length. We also observed this tendency after application of the RMSA criterion alone, but the use of additional postfilters, exhibiting the opposite fragment-length dependency, largely compensates for this trend if one compares the success rate of the first rank prediction. In contrast, for the average RMSL value of the first five ranks, this tendency of lower quality for long loops is still seen after application of postfilters. This reflects the lower number of good fragments for long loops in the database, but on the other hand, it shows the ability of the postfilters to promote those to the first rank.

The RMSL values obtained for self-predictions (target and template identical) are always lower than those using homologous template–target pairs, as an additional difference is produced by the global fit of template and target structures. Thus, for template–target pairs with high overall similarity also, the RMSL values are generally expected to be lower. For self-prediction examples, we reached at least the same or higher accuracy as described in the literature—in many cases even after rejection of the original loop.

The results obtained so far are in fact very encouraging. Further improvement of the prediction quality can be expected by better positioning of the introduced fragments (e.g., by minimization of the anchor region). An even better performance of the collision filters is expected if the amino acid side-chains of the target protein are converted prior to the loop selection. A more complete fragment database will increase the number of good solutions for longer fragments. Additional filter functions (e.g., solvation terms) that take the loop environment into account could improve the prediction. On the other hand, positioning of anchor groups is very important.[20]

Based on the results presented here and in previous articles, and the potential for improvements, a successful prediction of the structures of many loop regions, at least for those proteins with moderate dissimilarity between template and target, seems to become probable in the near future. The loop prediction process in protein structure projects consists of (1) optimal alignment, (2) optimal placement of anchor groups positions, and (3) prediction of the loop conformation. Including this work, we were recently able to demonstrate significant progress for (2) and (3).[19,20]

ACKNOWLEDGMENTS

We wish to thank Jürgen Dönitz for his contribution to the programming of the torsion angle potential.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Henford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Meslrov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardls ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predkl P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazler M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapatl RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Huttorl M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totokl Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottler P, Bruls T, Pelletler E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubols J, Rosenthal A, Platzer M, Nyakatura G, Taudion S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McComble WR, de la Bastide M, Dedhla N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smith AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenballi S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabriellan AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelan B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalusch F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Purl V, Qureshl H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaverl J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karluk B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrall S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kashu J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wung M, Wen M, Wu D, Wu M, Xia A, Zandich A, Zhu X. The sequence of the human genome. Science 2001;291:1304–1351.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–562.
4. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001; 45(Suppl 5):98–118.
5. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;45(Suppl 5):55–67.
6. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001; 45(Suppl 5):22–38.
7. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
8. Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. FASEB J 1995;9:708–717.
9. Bajorath J, Sheriff S. Comparison of an antibody model with an X-ray structure: the variable fragment of BR96. Proteins 1996;24: 152–157.
10. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein–DNA interactions: A structural analysis. J Mol Biol 1999;287:877–896.
11. Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol 1993;229:1065–1082.
12. Li W, Liang S, Wang R, Lai L, Han Y. Exploring the conformational diversity of loops on conserved frameworks. Protein Eng 1999;12:1075–1086.
13. Wojcik J, Mornon JP, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289:1469–1490.
14. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng 2001;14:473–478.
15. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
16. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
17. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
18. van Vlijmen HW, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
19. Lessel U, Schomburg D. Importance of anchor group positioning in protein loop prediction. Proteins 1999;37:56–64.
20. Wohlfahrt G, Hangoc V, Schomburg D. Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. Proteins 2002;47: 370–378.
21. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
22. Deane CM, Blundell TL. A novel exhaustive search algorithm for

predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
23. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
24. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
25. Lessel U, Schomburg D. Similarities between protein 3-D structures. Protein Eng 1994;7:1175–1187.
26. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
27. Diamond R. A note on the rotational superposition problem. Acta Crystallogr A 1988;44:211–216.
28. Niefind K, Schomburg D. Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. J Mol Biol 1991;219:481–497.
29. Sippl MJ. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
30. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding: an approach to the computational determination of protein structures. J Comput Aid Mol Des 1993;7:473–501.
31. Kocher JP, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
32. Hooft RWW, Sander C, Vriend G. Verification of protein structures: Side-chain planarity. J Appl Crystallogr 1996;29:714–716.
33. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
34. Qian B, Goldstein RA. Distribution of Indel lengths. Proteins 2001;45:102–104.