Positioning of Anchor Groups in Protein Loop Prediction: The Importance of Solvent Accessibility and Secondary Structure Elements

Gerd Wohlfahrt,* Vu Hangoc, and Dietmar Schomburg
University of Cologne, Institute of Biochemistry, Köln, Germany

ABSTRACT **The prediction of loop regions in the process of protein structure prediction by homology is still an unsolved problem. In an earlier publication, we could show that the correct placement of the amino acids serving as an anchor group to be connected by a loop fragment with a predicted geometry is a highly important step and an essential requirement within the process (Lessel and Schomburg, Proteins 1999;37:56–64). In this article, we present an analysis of the quality of possible loop predictions with respect to gap length, fragment length, amino acid type, secondary structure, and solvent accessibility. For 550 insertions and 544 deletions, we test all possible positions for anchor groups with an inserted loop of a length between 3 and 12 amino acids. We could show that approximately 80% of the indel regions could be predicted within 1.5 Å RMSD from a knowledge-based loop data base if criteria for the correct localization of anchor groups could be found and the loops can be sorted correctly. From our analysis, several conclusions regarding the optimal placement of anchor groups become obvious: (1) The correct placement of anchor groups is even more important for longer gap lengths, (2) medium length fragments (length 5–8) perform better than short or long ones, (3) the placement of anchor groups at hydrophobic amino acids gives a higher chance to include the best possible loop, (4) anchor groups within secondary structure elements, in particular β-sheets are suitable, (5) amino acids with lower solvent accessibility are better anchor group. A preliminary test using a combination of the anchor group positioning criteria deduced from our analysis shows very promising results. Proteins 2002;47:370–378.**

INTRODUCTION

Large-scale gene sequencing projects generate an enormous amount of gene and protein sequence data. However, only about 1% of currently known unique amino acid sequences possess an experimentally determined corresponding 3D structure.[1]

Protein structure prediction by homology is a widely employed method whenever experimental 3D protein structures are unavailable. Comparative modeling is a tool that works well for protein cores, especially in cases where templates with high sequence identities are used. A rather challenging task has been the prediction of protein folding near insertions/deletions in sequence alignments. Insertions/deletions are typically situated at the surface of proteins where they connect secondary structure elements. These regions, often referred to as loops, can be important for protein function as they frequently take part in molecular recognition, e.g., in the hypervariable regions of antibodies, or binding of substrates or DNA (see Fetrow[30]). An important application of homology modeling is the structural analysis of the same protein from different species or tissues, e.g., in order to guide rational design of selective inhibitors. Most of these differences are located in loop regions of those proteins. The prediction of loop regions is also important in protein design, where usually only a few amino acids are modified whereas the rest of the protein stays constant.

In loop prediction, protein fragments are inserted between "anchor groups" of a template protein. Length and sequence of the inserted fragment are usually determined by sequence alignment of template and target protein. The goal is to select fragments closest in similarity to the target structure.

In general, two different approaches to the prediction problem have been developed: conformational search methods and knowledge-based methods. Basics for the conformational search methods were published by Moult and James[2] and Bruccoleri and Karplus.[3] They searched the entire conformational space for fragments matching template-defined anchor groups. In order to limit the number of basic conformations, they derived rules from known structures and developed criteria for ranking all remaining fragments.

Knowledge-based methods function by selecting fragments from data banks of known structures. This approach was first applied by Jones and Thirup[4] who employed loop prediction based on data banks with structurally known protein fragments. The authors were able to reconstruct the backbone of retinol-binding protein using the fragments of only three proteins. Meanwhile, several different fragment data banks have been constructed (e.g.,[5–8]).

*Correspondence to: Gerd Wohlfahrt, Orion Corp. ORION PHARMA, P.O. Box 65, FIN-02101 Espoo, Finland. E-mail: Gerd.Wohlfahrt@orionpharma.com

Received 8 June 2001; Accepted 13 December 2001

Classification of short- and medium-length loops has been attempted by ordering them according to the geometry and nature of anchor regions. Rooman et al.[9] generated clustered fragments in order to characterize recurrent folding motifs and to classify loops, which can be used for loop predictions. Fechteler et al.[10] applied two different clustering algorithms to build up fragment data banks, and tested several examples of insertions and deletions based on geometric criteria. Sudarsanam et al.[11] created a data bank with observed backbone angles of dimers. Using this data bank, they constructed different loops with sterically allowed conformations from which they could select those that best matched the anchor groups. In more recent works, Li et al.[12] grouped all loops in the PDB according to their framework, while Wojcik et al.[13] proposed a complete loop classification model, which was independent from the nature of flanking regions.

One of the main problems of loop prediction is the exponential increase in the number of possible conformations according to fragment length. As a result, knowledge-based methods suffer from an exponential rise in incompleteness of data banks, while conformational searches have to be limited to the prediction of shorter fragments as well as to the use of fewer basic conformations per residue in order to permit acceptable calculation times. Van Vlijmen and Karplus[14] concluded that the present Protein Data Bank[15] can be useful for the prediction of loops up to nine residues in length. Compared to conformational search methods, data banks derived from known determined structures have the advantage that they guarantee the prediction of frequently occurring, i.e., energetically favored, fragments.

Conformational search methods and knowledge-based methods have inherent problems, which increase according to loop fragment size. Not even through the combined use of both methods could these problems be eliminated (e.g., Martin et al.[16]).

Along with the problem of database incompleteness, Lessel and Schomburg[17] were able to show the high importance of anchor group selection in loop prediction. Encouraging results were achieved in predicting loops with identical length in template and target as well as in "self-predictions" (loops are fitted into the target structure) (e.g., Deane and Blundell[18]). However, loops around insertions and deletions have still been predicted insufficiently. In order to improve structure predictions in these regions, we applied a knowledge-based method on examples of structurally aligned protein pairs with insertions and deletions. These examples resemble a real modeling situation in which loops are fitted into a partly wrong or unfinished framework. The modeling process includes the determination of structurally variable regions, the proper positioning of anchor groups, and the selection of appropriate loop fragments. Consequently, as a preceding step to fragment selection, we derived a set of rules for the positioning of anchor groups in order to improve prediction quality.

In this article, we present a systematic analysis of the effects of anchor group positioning. Anchor groups were characterized by type of amino acid, solvent accessibility, secondary structure, and fragment length (sequence distance). Additionally, conclusions were drawn about maximum prediction quality and completeness of the applied loop data bank.

The aim of this work was to improve the positioning of anchor groups. Meaningful statistical results are only obtained if the other parameters of the loop modeling process are as chosen well as possible.

The use of inaccurate sequence alignments hardly affects the results of the presented method as long as shifts of gaps are small and within the loop regions. Our algorithm has the tendency to bridge these areas by longer fragments connecting regular secondary structure elements. If gaps are placed by sequence alignments within regular secondary structure, most modeling approaches move them into loop regions of the template. If there are larger inaccuracies in the sequence alignment, all subsequent modeling steps will lead to incorrect structures anyway.

MATERIALS AND METHODS

Fragment Data Bank

The fragment data bank is based on all X-ray structures in the 2/98 release of the Protein Data Bank[15] that have resolutions of smaller or equal to 2.0 Å and sequence identities of less than 95% determined by Smith-Waterman algorithm[19] using standard gap penalties. After fitting N-, C_α and C-carbonyl atoms of two ending residues, we eliminated fragments showing a root mean square deviation (RMSD) below 0.25 Å considering all backbone atoms. The RMS fit was performed following the procedure by Diamond.[20] The limit of 0.25 Å was chosen according to the estimated standard error in X-ray analysis. As a geometric pre-filter for comparisons, the distance between anchor group atoms was determined for each fragment. This distance was defined by the distance between the middle of C_α - C = O atom bond of the N-terminal anchoring residue and the middle of N - C_α atom bond of the C-terminal anchoring residue. The fragments were considered structurally distinct when the difference in anchor group distance between two corresponding fragments exceeded 0.5 Å (Table I).[17]

Creation of the Test Data Set

All examples for insertions and deletions were derived from structurally aligned protein pairs. First, proteins from the Protein Data Bank[15] release 2/98 with less than 50% sequence identity were chosen using the algorithm by Smith and Waterman[19] with standard gap penalties. Then, selected proteins were compared with each other according to structural similarity using the method of Lessel and Schomburg.[21] All proteins with at least 35 matching C_α-atoms and at least 40% structural similarity were grouped into the same family, resulting in 132 classes where each contained more than one member. From each of these families, the protein pair with the highest structural similarity within the class was selected. Some of these families were further subdivided, since some

TABLE I. Comparison of Original Number of Fragments in Data Sets and Number of Loops in Fragment Data Bank After Elimination of Similar Fragments

Fragment length	No. of fragments in PDB files	No. of fragments in data bank
3	184,157	13,285
4	183,031	53,853
5	181,929	98,919
6	180,835	122,077
7	179,750	133,082
8	178,671	141,165
9	177,596	148,336
10	176,527	153,982
11	175,461	158,225
12	174,403	161,501

TABLE II. Composition of the Test Data Set[†]

Gap	No. of permutations		No. of permutations	
length	Insertions	Insertions	Deletions	Deletions
0	45	3,902	45	3,902
1	273	14,871	274	14,588
2	107	4,200	105	4,044
3	48	1,665	48	1,621
4	37	978	37	946
5	36	669	36	641
6	18	211	18	198
7	12	111	12	100
8	7	42	7	35
9	7	19	7	13
10	5	5	—	—
Total	595	26,673	589	26,088

[†] The number of loops to be predicted for the different gap length and the numbers resulting from all possible permutations of anchor groups.

groups of protein pairs showed higher similarity to each other than to other members or subclasses within the family. In those cases, representative protein pairs for each subclass were chosen, resulting in a total of 170 protein pairs. The selection was performed in order to avoid biases within examples, e.g., to prevent the occurrence of the same globin surface loop in several variations. The level of structural similarity of the examples is in the typical range for a real prediction problem in homology modeling.

Then, sequence alignments corresponding to the global structural fit were created for these 170 protein pairs using the method of Lessel and Schomburg.[21] These sequence alignments were systematically searched for appropriate insertions and deletions under the following condition: Blocks of at least three structurally aligned residues in a row had to be located at both ends of an insertion or deletion. Structurally aligned in this case means an RMSD for C_{α}-atoms below 1.8 Å. The number of residues between these blocks was greater for insertions than for deletions, while they were equal for loops with zero-length difference. A maximum of ten separating residues was allowed, since the length of the fragments in the loop data bank[22] was limited to 12 residues (2 anchoring plus up to 10 separating residues). Only one residue per anchor group is used since there is no advantage in using longer anchor groups.[14] By exchanging template and target protein, each example was used as an insertion as well as a deletion. This procedure resulted in 544 deletions and 550 insertions (Table II). Additionally, 45 examples with zero-length difference but folding differences in loops one to eight residues long were chosen (i.e., differences in flexible loops). These examples were included for comparison purposes.

Anchor Group Combinations

In order to test the effect of anchor group positioning on loop prediction, we generated all possible anchor group combinations for each example in the test set using loop fragments ranging from 3 to 12 residues. For a one-residue insertion, for instance, a total of 55 different anchor group combinations is possible (i.e., 1 position for the 3-residue fragment, 2 possible positions for a 4-residue fragment, . . . 10 possible positions for a 12-residue loop). Their positions range from 10n residues before to 10n residues behind the gap. For a 10-residue insertion, only one combination of anchor groups exists because of the fragment length limit of 12. For the 550 insertions and 544 deletions of varying loop lengths, we generated 22,771 and 22,186 anchor group combinations, respectively. The 45 zero-length examples resulted in 3,902 anchor group pairs, since in this case the "closest allowed" distance of the anchoring residues ranged here from one to eight residues (Table II). Gaps located close to either termini of the protein result in less anchor group combinations than mathematically possible.

Evaluation of Highest Possible Prediction Quality

In order to determine whether the examples of the test data set could be predicted, if an optimal selection algorithm were found, each appropriate loop from the fragment data bank was inserted into the templates and compared with the target structures for each problem in the test data set.

In the analysis presented, first of all template and target protein were fitted globally using the 3D-alignment procedure of Lessel and Schomburg.[21]

All fragments of the data bank with certain length and "intramolecular" distances similar to the template (< 0.5 Å deviation) were fitted onto the anchor groups of the template protein using the RMS fit procedure of Diamond.[20] During this process, N-, C_{α}-, and C-carbonyl atoms of both anchoring residues were considered for the overlay procedure.

Finally, the RMSD for the loop backbone atoms was derived by comparing the complete structures of template and target rather than by simply using the short loop fragments. It would not be sufficient to determine an RMSD value between solely the inserted loop and the target loop, since an incorrect orientation with respect to the target protein (of a correct loop conformation) would

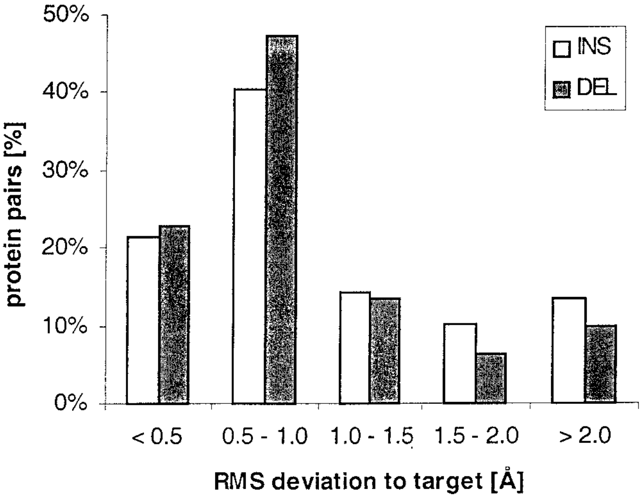

Fig. 1. Maximum prediction quality for the test set. Distribution of RMSDs between template and target protein for the best anchor group combination in each protein pair.

not be identified. In this way, the best fragment for each example, i.e., that showing the lowest RMSD to the target after insertion into the anchor groups of the template, was determined.

Properties of Anchor Groups

All loops with known RMSD were classified by gap lengths and length of inserted fragments. Secondary structure of their anchoring residues was determined by using SSTRUC by David Smith.[23] 3_{10}-, α- and π-helices were grouped into one general helix class labelled as H, β-sheets and extended conformations were classified as B and all other turns and non-regular structures were labelled as O (see Fig. 5).

The relative solvent accessibility of amino acid residues was calculated by using the method of Lee and Richards[24] implemented in the PSA program.[23] The accessibilities of the two anchoring residues at both ends of the loop fragments were averaged (see Fig. 6).

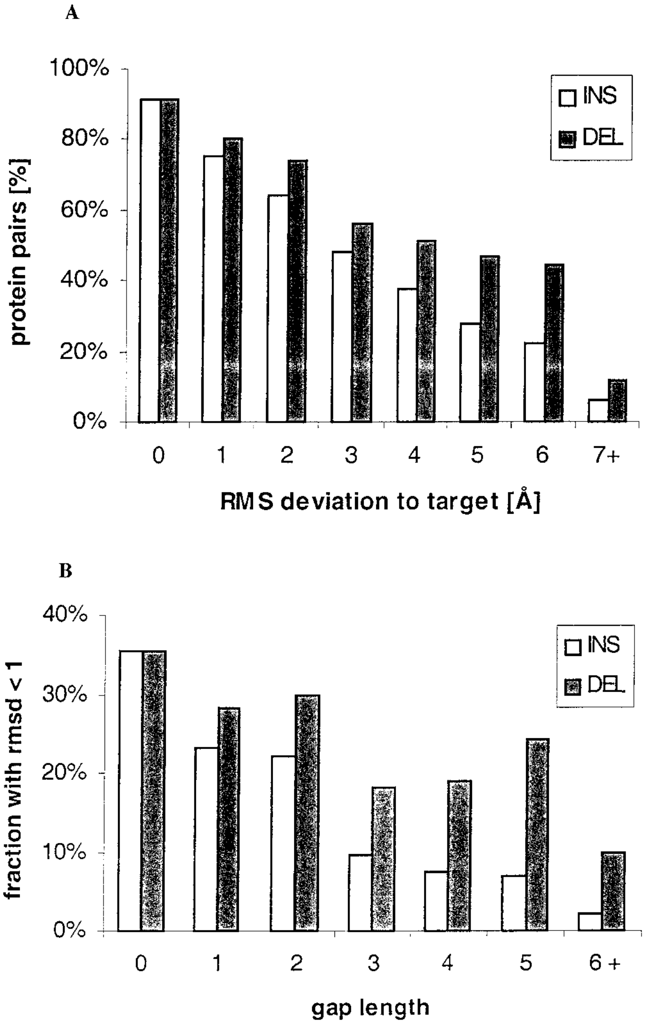

Fig. 2. Influence of gap length. **A:** Maximum prediction quality for different gap lengths. Only the anchor group combination allowing the best fit is taken into account. Bars present the fraction of hits with an RMSD < 1.0 Å. **B:** Fraction of anchor groups in the whole test set that allow the fitting of fragments with an RMSD < 1.0 Å. Anchor groups are categorized by the gap length.

RESULTS

Maximum Prediction Quality

First, the maximum prediction quality for the test data set with our fragment data bank was determined. It would be possible with 62% of the insertions and 70% of the deletions to find anchor groups that allow to insert fragments with an RMSD < 1.0 Å (Fig. 1). With a quality criteria of less than 1.5 Å this is 76 and 84%, respectively.

Out of the 22,771 anchor group combinations tested for the 550 insertions, 21% allow the fitting of at least one fragment with an RMSD of < 1.0 Å with respect to the target protein. Twenty-seven percent of the 22,186 combinations from the 544 deletions and 36% of the zero-length examples enable predictions with less than 1.0 Å RMSD.

Gap Length

For each example, the anchor group combination allowing the lowest RMSD was determined. The fraction that allows an RMSD < 1.0 Å is shown in Figure 2(a) for each separate gap length resulting from the structural alignment of template and target protein. Among deletions up to a gap length of six, sufficiently good anchor group combinations exist for 45% of the cases, assuming ideal anchor groups could be identified. On the other hand, among insertions only for gaps up to three residues a comparable prediction quality can be expected.

The total number of anchor groups that allow an RMSD < 1.0 Å shows a slightly different distribution [Fig. 2(b)]. Here, the decrease for the deletions is less pronounced, whereas for insertions a relatively low level is reached at a gap length of three.

Fragment Length

For comparative purposes, odds ratios represent a more instructive representation than percentages. Odds ratios

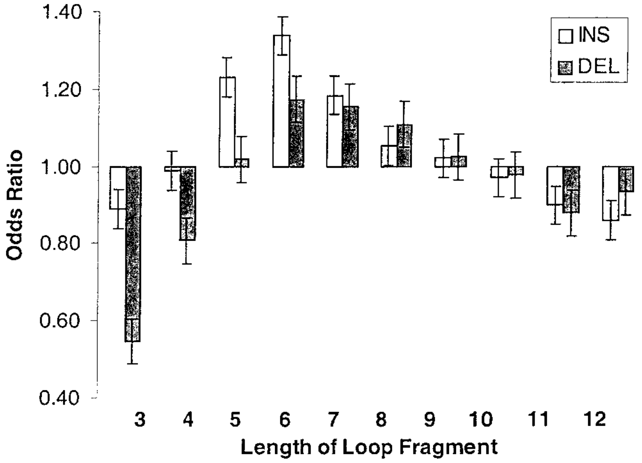

Fig. 3. Prediction quality for loop fragments of different lengths. Odds ratios > 1 represent higher likelihood of obtaining RMSDs < 1 Å. Odds ratios were calculated by dividing the ratio of fits with an RMSD < 1 Å to fits with RMSD ≥ 1 Å between residue category and total database. Error bars represent the standard error of the mean.

above 1 represent a higher likelihood of obtaining RMSDs below 1 Å. Odds ratios were calculated by dividing the ratio of fits with RMSD below 1 Å to fits with RMSD above or equal to 1 Å between residue category and total database.

In general, fragment length does not considerably influence the prediction quality. Medium-length fragments perform slightly better than short or long ones (Fig. 3). The use of three and four residue fragments (including anchor groups) for deletions leads to a comparably low probability of detecting appropriate loops.

There is no clear relationship between fragment length and gap length (Table III). In general, anchor groups very close to the gap as well as very long fragments perform quite well.

Amino Acid Type

Table IV shows the odds ratios for the performance of the 20 different amino acids as anchor groups. Cysteine and methionine lead to the best predictions, but these amino acids occur at a low frequency. Among the more frequent amino acids, tyrosine, leucine, and valine are present in well-performing anchor groups, whereas glycine and proline, which frequently occur in loop regions, show a low performance.

Differences between the different amino acids with similar physicochemical properties are minor. Therefore, they are grouped according to the nature of their side chains. Figure 4 shows good performance for anchor groups with aromatic and aliphatic residues, whereas polar uncharged amino acids perform weakly.

Secondary Structure

3_{10}-, α- and π-helices were grouped into one general helix class H. β-sheets and extended conformations were grouped in class B and all other turns and non-regular structures in class O (Fig. 5). About 22% of the 105,522 anchoring residues used here are in helices; 30% belong to class B and 48% to class O. The probability for a sufficient prediction using different combinations of these three classes is shown in Figure 5. Loops connecting two β-sheets have the highest probability of being predicted correctly, whereas if both anchor groups are situated in non-regular structure, the lowest performance is expected.

Solvent Accessibility

The relative solvent accessibility averaged for both anchoring residues is shown in Figure 6. A very clear relationship between solvent accessibility and suitable anchor groups can be seen. The odds ratios for the anchor groups with 0% accessibility are even higher with 2.18 for insertions and 2.32 for deletions. About one third of all possible anchor groups in our test set has a solvent accessibility of less than 20%, which still results in a higher than average probability of finding a good match. The few examples of anchor groups that are nearly completely accessible (> 70%) result in odds ratios of about 0.2 and less.

Combination of Criteria

The odds ratios described in this article were combined in the following manner: For each anchor group combination, the class of secondary structure, loops length, relative solvent accessibility, and amino acid type was determined and the corresponding odds ratios assigned as scores. Combined odds ratios were calculated by multiplying the individual odds ratios for each pair of anchor groups. Finally, the anchor groups with the highest combined total odds ratio score were selected for a particular protein pair.

The distribution of the maximum prediction quality for the highest ranked anchor groups is shown in Figure 7(a). In 27% of the insertions and 35% of the deletions in the test data set, an anchor group combination that allows predictions with less than 1 Å RMSD was found with these criteria. The values increase to 46 and 62% if an RMSD of less than 1.5 Å is considered as sufficient. Summation of odds ratios leads to results very similar to multiplication (data not shown). Compared to the prediction quality with anchor groups positioned by chance, the highest improvement is seen for gaps of medium length (three to five residues) [Fig. 7(b)].

DISCUSSION

The prediction of protein structures around insertions and deletions represents one of the most difficult challenges in comparative modeling. For regions without gaps in the sequence alignments, in most modeling processes simply the backbone of the template is used, even if they are probably structurally different. These are possible targets for loop predictions as well.

Knowledge-based loop predictions depend mainly on the following: identification of structurally variable regions, choice of the anchor groups, the algorithm for selecting appropriate fragments, fitting/optimization procedure, and completeness of the fragment data bank.

TABLE IIIA. Correlation Between Fragment Length and Gap Length for Deletions[†]

Fragment	Gap length (%)									
length (%)	0	1	2	3	4	5	6	7	8	9
3	**100.0**	**100.0**								
4	**100.0**	90.0	66.7							
5	**100.0**	**100.0**	83.3	**100.0**						
6	**100.0**	86.5	87.5	33.3	50.0					
7	**100.0**	71.4	80.0	75.0						
8	**100.0**	75.9	**93.8**	40.0	50.0		0.0			
9	75.0	63.3	66.7	50.0	33.3	33.3				
10	83.3	75.0	64.3	60.0	**72.7**	**71.4**	**100.0**	**25.0**		
11	**100.0**	79.7	66.7	57.9	40.0	42.3	35.7	12.5	**14.3**	0.0

[†] The fragments with the highest relative prediction possibility for each gap length are marked in bold.

TABLE IIIB. Correlation Between Fragment Length and Gap Length for Insertions[†]

Fragment	Gap length (%)									
length (%)	0	1	2	3	4	5	6	7	8	9
3	**100.0**	**100.0**								
4	**100.0**	90.9	**100.0**							
5	**100.0**	83.3	**100.0**							
6	**100.0**	77.8	**100.0**	0.0						
7	**100.0**	73.7	58.3	**60.0**	50.0	**100.0**				
8	**100.0**	63.2	40.0	0.0	**100.0**	0.0				
9	75.0	75.0	80.0	22.2	50.0					
10	83.3	75.0	71.4	55.6	44.4	20.0	0.0			
11	**100.0**	64.3	30.8	50.0	16.7	22.2	0.0	0.0		
12	77.8	77.6	70.0	**60.0**	33.3	27.8	**33.3**	**20.0**	0.0	0.0

[†] The fragments with the highest relative prediction possibility for each gap length are marked in bold.

Identification of structurally variable regions is mainly performed via sequence alignments, and if structural information is available, the identification process should be guided by structural comparison.[25]

The problem of database incompleteness is slowly decreasing with the growing number of structures in the PDB.[17] Another approach is the systematic ab initio generation of fragments with all relevant conformations.[19] For fragments of up to nine residues there is already sufficient coverage of the conformational space based on the PDB.[14]

The methods for selecting/ranking the loop fragments still suffer from low efficiency.[26] Therefore, in our analysis only a simple geometric criterion for fragment selection was used in order to determine rules for anchor group positioning. Before improved ranking algorithms can be tested in real prediction problems, anchor groups have to be selected that allow sufficient prediction quality. For this reason, the present article focuses on the positioning of anchor groups based on criteria that do not use information from the protein structure to be predicted.

Maximum Prediction Quality

With the present completeness of the fragment database used here, about two thirds of the loops in our test data set could be predicted in principle with an RMSD < 1.0 Å if anchor groups are well positioned and an appropriate algorithm for the fragment selection is used (Fig. 1). An

TABLE IV. Prediction Quality of Anchor Groups With Different Amino Acids[†]

	Insertion	Deletion	Frequency in PDB
ALA	0.98	1.02	8.4
ARG	0.94	0.99	4.9
ASN	0.78	0.84	4.4
ASP	1.04	0.86	5.8
CYS	1.38	1.39	2.1
GLN	0.93	0.84	3.7
GLU	1.05	0.99	6.8
GLY	0.84	0.88	7.6
HIS	1.13	1.09	2.2
ILE	1.11	1.10	5.5
LEU	1.25	1.07	8.1
LYS	0.92	1.00	6.8
MET	1.11	1.35	2.2
PHE	1.09	1.17	3.8
PRO	0.84	0.74	4.5
SER	0.97	0.83	5.8
THR	0.88	0.96	5.7
TRP	1.06	1.11	1.4
TYR	1.10	1.41	3.5
VAL	1.15	1.18	6.9

[†] Odds ratios > 1 represent higher likelihood of obtaining RMS deviations of < 1 Å. Odds ratios were calculated by dividing the ratio of fits with RMSD < 1 Å to fits with RMSD ≥ 1 Å between residue category and total database. The relative occupancy proportions of the 20 amino acids in the whole PDB are taken from Wojcik et al.[13]

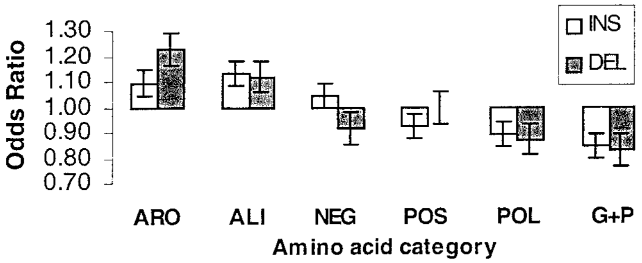

Fig. 4. Influence of amino acid types in anchor groups on possible prediction quality. Odds ratios with error bars representing the standard error of the mean are shown. Anchoring residues are grouped according to the nature of their side chains: aromatic (**ARO**: F, H, W, Y); aliphatic (**ALI**: A, C, I, L, M, V); negative (**NEG**: D, E), positive (**POS**: K, R), polar **POL**: (N, Q, S, T) and an extra group (**G+P**: G, P).

RMSD of < 1.0 Å compared to the target structure can be regarded as a successful prediction with respect to the error in subsequent optimizations with force field methods (1.8 Å RMSD was used as similarity criterion in the test set generation). This criterion is quite strict, as one should see high RMSD values for loop regions in the context of their mobility as shown by NMR structures and X-ray temperature factors. Due to our determination method of RMSD values for the loop fragments, the RMSD between overall template and target structure has an influence. This leads to an apparently lower maximum prediction quality in the case of protein pairs with low overall similarity. In this respect, more advanced fitting methods, e.g., by minimizing interaction energies, could further decrease the RMSD values. This implies that anchor groups allowing RMSD values of less than 1.5 Å are probably useful for several applications. In this case, our database based on the PDB from 2/98 would provide good fragments for about 80% of our test cases. Short gaps are over-represented in the test set, but this reflects the observed distribution of insertions/deletions in evolution where an exponential decrease of examples with longer gaps is observed.[27]

Influence of Gap Length

The prediction quality rapidly decreases with increasing length of gaps in the sequence or structural alignment [Fig. 2(a)]. On the one hand, this reflects the incompleteness of the fragment database for longer fragments. On the other hand, since the decrease in prediction quality is stronger than it would be solely due to database incompleteness, increasing structural diversity of the loop regions in template and target plays a significant role as well. The longer the gap is, the more the diversity of its structural environment with respect to the target is expected. In general, longer gaps cause stronger distortions of the local fold. In these cases, the anchoring regions do not provide enough information about the structure of the whole loop region.

Anchor groups, which are not too close to the structurally different region, perform quite well (Table III). This means that longer fragments are better in many cases, but at the same time, the prediction quality is reduced for the known reason of database incompleteness.

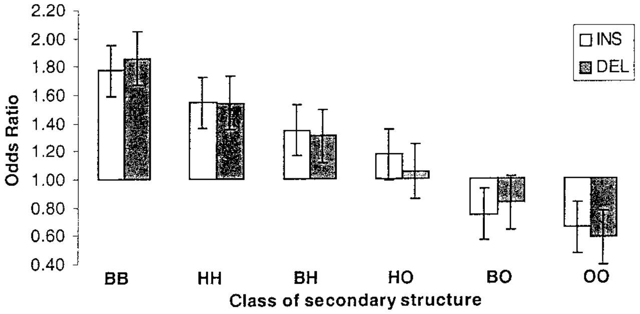

Fig. 5. Likelihood for anchor groups to obtain a prediction with RMSD < 1 Å, categorized by secondary structure (H helical, B extended, O turns and non-regular structure). Odds ratios were calculated by dividing the ratio of fits with RMSD < 1 Å to fits with RMSD < 1 Å between residue category and total database.

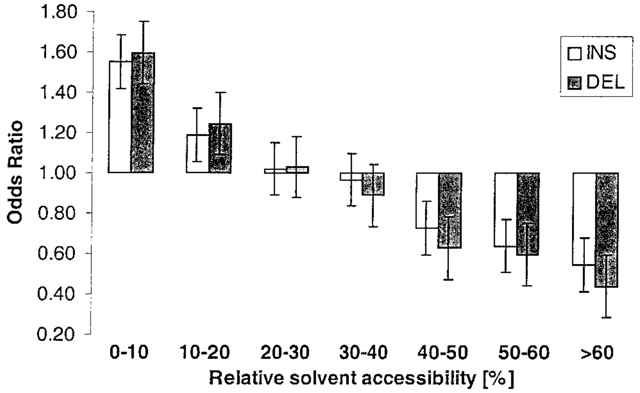

Fig. 6. Likelihood for anchor groups to achieve a prediction with RMSD < 1 Å categorized by relative solvent accessibility. Relative accessibility values were averaged for the two anchoring residues. Odds ratios are displayed with error bars representing the standard error of the mean.

Influence of Fragment Length

Interestingly, there is only a weak correlation between the length of the inserted fragments and the maximum prediction quality. Many long fragments allow good predictions even for gaps of one-residue length (Table III). This means that database incompleteness for longer fragments is compensated by their higher ability to fit. This is particularly obvious for short fragments (3 and 4 residues) in the prediction of short deletions where the steric strain is high (Fig. 3). The PETRA algorithm by Deane and Blundell is also less successful for shorter loops.[18] In general, fragments with a length of six to seven residues give slightly better results than the average. In this range, the database is quite complete and the fragments can readily adjust to "steric strain."

Amino Acid Type

The classification of the amino acids of the anchor groups according to the chemical nature of their side chains leads to a very clear distribution (Fig. 4). All amino acids with similar properties exhibit similar probabilities in the loop prediction quality. Hydrophobic anchor groups, which are quite often buried, give better predictions than

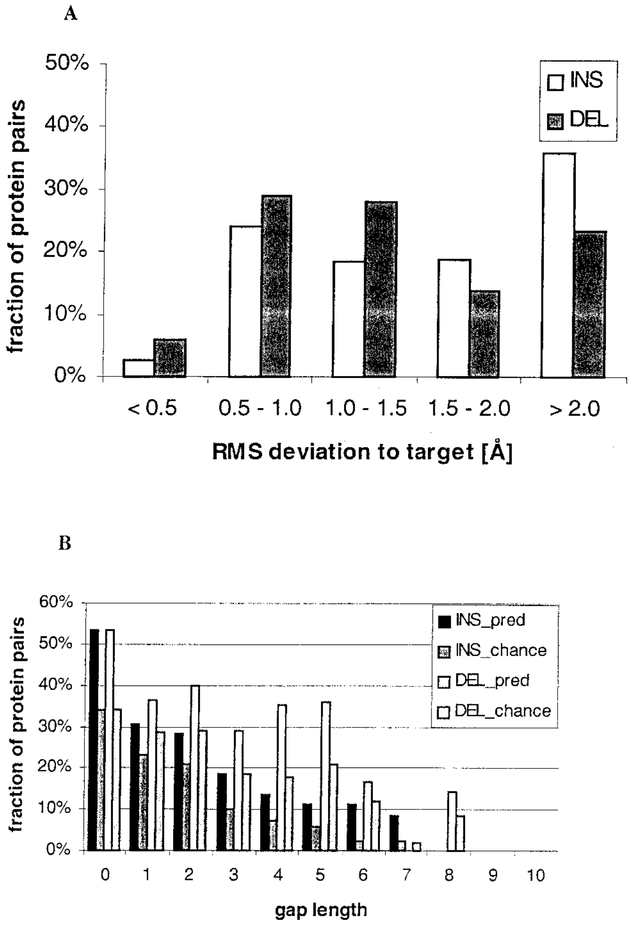

Fig. 7. Prediction quality by application of the combined anchor group positioning criteria. **A:** Data set was established by selecting the anchor group combination with the highest combined odds ratio for each protein pair. Combined odds ratios were calculated by establishing the classes of secondary structure, loop length, relative solvent accessibility, and amino acid type for each anchor group combination and assigning the corresponding odds ratios as scores. Scores were then multiplied for each pair of anchor groups, and the anchor groups with the highest combined total odds ratio score were selected for each protein pair. **B:** The fraction of possible predictions with an RMSD < 1.0 Å determined by the combined odds ratios is compared with possible predictions with an RMSD < 1.0 Å obtained by random choice of anchor groups.

charged residues. The lowest prediction quality can be expected for polar uncharged side chains, glycine and proline. In addition to the nature of the side chains, the backbone flexibility and tendency to occur in loop regions seem to play a role. Glycine and serine, which occur in flexible hinge regions and proline, which interrupts regular secondary structure, give poor results (Table IV).

Badly performing amino acids are found more frequently in loop regions.[13] However, the correlation of amino acid type and prediction quality is not as strong as for the other criteria (secondary structure, solvent accessibility), except for a positive correlation for cysteine and for deletions with methionine and tyrosine.

Secondary Structure

Especially good predictions of loop regions are possible by placing anchor groups in β-sheets and helices, while anchor groups in non-regular structural elements are less suitable (Fig. 5). A defined secondary structure of the anchor groups leads to a defined orientation of the neighboring loop residues. This makes template and target more comparable by the RMSD of the anchor groups. In previous works, the connection of regular secondary structure[28] or the classification of conserved frameworks around the loops[12,29] is used in the prediction process. From our study, it becomes obvious that it is advantageous to position the anchor groups in secondary structure elements; on the other hand, fragment length should not become too long. As an additional problem, the definition of secondary structure is not exact and depends on the method of calculation. Some authors even consider the difference between non-regular structural elements and secondary structure as a continuous geometric transition.[13]

Solvent Accessibility

Relative solvent accessibility of the anchor groups and probability for a good prediction are strongly correlated. Buried anchor groups perform much better than surface groups (Fig. 6). This is probably due to the fact that buried parts of a protein are less flexible.

The significance of the solvent accessibility is even slightly higher if not averaged between both anchoring residues (data not shown).

Combination of Rules

Choosing anchor groups by pure chance would lead to a maximal prediction rate of 18% for insertions and 26% for deletions with the 1 Å RMSD criteria (42 and 52% for < 1.5 Å). Already, the application of simply multiplied odds ratios increases these values to 27% for insertions and 35% for deletions, respectively [Fig. 7(a)]. This is in the range one can reach with pre-knowledge derived from the target structure like the RMSD value of the anchor groups of template and target in a global fit,[17] but this type of information is not available in real prediction problems.

Additional criteria, such as flexibility defined via temperature factors and improved rules for weighting all criteria, should lead to a sufficient choice of anchor groups. Additional attention should be paid to the redundancy of our criteria, which cannot be assumed to be independent.

CONCLUSIONS

The prediction of loop regions in protein structure prediction projects is still an unsolved problem. The process can be split into several steps: (1) the identification of anchor groups that belong to structurally conserved regions of the proteins and are located as close to the variable region as possible; (2) the selection of possible loops connecting the anchor groups either by conformational search or from loop data bases; and (3) the identification of the best candidate from the list of possible loops. These essential steps could be followed by an energy optimization step.

In an earlier publication, we could show that the correct placement of anchor groups is a highly important step and an essential requirement within the whole process.[17] In this paper, we analyzed the quality of possible loop predictions with respect to gap length, fragment length, amino acid type, secondary structure, and solvent accessibility. For 550 insertions and 544 deletions it could be shown that 80% could be predicted within 1.5 Å RMSD from a knowledge-based loop data base if the correct anchor groups could be found and the loops ranked correctly.

From our analysis, several conclusions concerning the best placement of anchor groups become obvious: (1) The correct placement of anchor groups becomes even more important for longer gap lengths, (2) medium-length fragments (length 5–8) perform better than short or long ones, (3) the placement of anchor groups at hydrophobic amino acids gives a higher chance of including the best possible loop, (4) anchor groups within secondary structure elements, in particular β-sheets are suitable, (5) amino acids with lower solvent accessibility are a better anchor group.

A preliminary test to use a combination of the anchor group positioning criteria deduced from the analysis shows very promising results. For a successful loop prediction, an improved scoring of a combination of the criteria in conjunction with a good ranking algorithm seems to be a promising approach.

ACKNOWLEDGMENTS

The authors thank Jürgen Dönitz for supporting the preparation of the test data set of structurally aligned protein pairs.

REFERENCES

1. Moult J. Predicting protein three-dimensional structure. Curr Opin Biotechnol 1999;10:583–588.
2. Moult J, James MNG. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
3. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
4. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
5. Sibanda BL, Blundell TL, Thornton JM. 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 1989;206:759–777.
6. Claessens M, Van Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with 'spare parts' from known protein structures. Protein Eng 1989;2:335–345.
7. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355–373.
8. Levitt M. Accurate modelling of protein conformation by automatic segment matching. J Mol Biol 1992;226:507–533.
9. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990;213: 327–336.
10. Fechteler T, Dengler U, Schomburg D. Prediction of protein three-dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. J Mol Biol 1995;253: 114–131.
11. Sudarsanam S, DuBose RF, March CJ, Srinivasan S. Modeling protein loops using a f_{i+1}, y_i dimer database. Protein Sci 1995;4: 1412–1420.
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. van Vlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
16. Martin ACR, Cheetham JC, Rees AR. Modeling antibody hypervariable loops: a combined algorithm. Proc Natl Acad Sci USA 1989;86:9268–9272.
17. Lessel U, Schomburg D. Importance of anchor group positioning in protein loop prediction. Proteins 1999;37:56–64.
18. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
19. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
20. Diamond R. A note on the rotational superposition problem. Acta Cryst A 1988;44:211–216.
21. Lessel U, Schomburg D. Similarities between protein 3D structures. Protein Eng 1994;7:1175–1187.
22. Lessel U, Schomburg D. Creation and characterization of a new, non-redundant fragment data bank. Protein Eng 1997;10:659–664.
23. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. JOY: protein sequence-structure representation and analysis. Bioinformatics 1998;14:617–623.
24. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
25. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1987;1:377–384.
26. Jones TA, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins 1999;37:30–46.
27. Qian B, Goldstein RA. Distribution of indel length. Proteins 2001;45:102–104.
28. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 1997;267:352–367.
29. Donate LE, Rufino SD, Canard LHJ, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. Protein Sci 1996;5:2600–2616.
30. Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. FASEB J 1995;9:708–717.