Bioinformatics Advance Access published September 7, 2007

Structural Bioinformatics

Modeling protein loops with knowledge-based prediction of sequence-structure alignment

Hung-Pin Peng and An-Suei Yang*

Genomics Research Center, Academia Sinica. 128 Academia Rd, Sec. 2, Nankang District, Taipei 115, Taiwan R.O.C.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: As protein structure database expands, protein loop modeling remains an important and yet challenging problem. Knowledge-based protein loop prediction methods have met with two challenges in methodology development: (1) Loop boundaries in protein structures are frequently problematic in constructing length-dependent loop databases for protein loop predictions; (2) knowledge-based modeling of loops of unknown structure requires both aligning a query loop sequence to loop templates and ranking the loop sequence-template matches.

Results: We developed a knowledge-based loop prediction method that circumvents the need of constructing hierarchically clustered length-dependent loop libraries. The method first predicts local structural fragments of a query loop sequence and then structurally aligns the predicted structural fragments to a set of non-redundant loop structural templates regardless of the loop length. The sequence-template alignments are then quantitatively evaluated with an artificial neural network model trained on a set of predictions with known outcomes. Prediction accuracy benchmarks indicated that the novel procedure provided an alternative approach overcoming the challenges of knowledge-based loop prediction.

Contact: A.-S. Yang yangas@gate.sinica.edu.tw

Availability: http://cmb.genomics.sinica.edu.tw

1 INTRODUCTION

Three-dimensional structures of a large portion of protein sequences can be predicted to reasonable accuracy with comparative modeling procedures (Pieper et al., 2006). For homologous protein pairs, insertions and deletions occur mostly in loop regions, each of which connects two regular secondary structure elements (Fiser et al., 2000). It is thus necessary to develop methodology to predict loop structures for which the structures of the flanking regular secondary structural elements are known.

Knowledge-based loop structure predictions have been explored as effective loop modeling methods. Moreover, recent studies have suggested increasing coverage of longer loop structure motifs due to the ever-expanding protein structure database (Fernandez-Fuentes and Fiser, 2006). The basis of the methodology is a hierarchical loop motif family library, where each of the loop motifs is composed of two stem secondary structure elements connected by a loop region (Burke et al., 2000; Donate et al., 1996; Espadaler et al., 2004; Fernandez-Fuentes and Fiser, 2006; Lessel and Schomburg, 1997; Li et al., 1999; Michalsky et al., 2003; Oliva et al., 1997a). Predicting the structure of a loop sequence is carried out by matching the loop sequence plus the bracing stem structures to a loop motif structure family in the loop motif library based on the length and the sequence of the query loop and the geometry of the known stem structures (Burke and Deane, 2001; Fernandez-Fuentes et al., 2006; Fernandez-Fuentes et al., 2005; Heuser et al., 2004; Lessel and Schomburg, 1999; Michalsky et al., 2003; Rufino et al., 1997; Wojcik et al., 1999). The premises underlying the knowledge-based procedures are the observations that protein loop structure motifs are frequently able to be sorted into a limited set of motif structure families, and that each of the motif structure families frequently embeds sequence patterns in the loop and in the bounding regions (see (Donate et al., 1996; Oliva et al., 1997b) and references therein). But since the boundaries of the bracing secondary structures are frequently ambiguous (Carter et al., 2003; Colloc'h et al., 1993) and the distribution of the bracing stem structures is not discrete in geometry, it has been difficult to construct a loop motif library with unambiguous boarders among loop motif families.

To overcome the problem, we developed a knowledge-based loop modeling procedure that predicts the structure of a query loop sequence by predicting the local structural fragments of the query sequence (Kuang et al., 2004; Yang and Wang, 2002; Yang and Wang, 2003) and by structurally aligning the predicted local structural fragments to all possible motif templates from know protein structures, circumventing the need for a hierarchical loop family library. Benchmark results indicated that this loop prediction procedure was among the best in prediction capacities in comparison with current knowledge-based loop prediction algorithms. More importantly, we have demonstrated an alternative knowledge-based loop prediction procedure without the need for a loop motif family database.

2 METHODS

Figure 1 summarizes the loop structure prediction procedure in six steps. As shown in the figure, the input is a polypeptide sequence plus two stem structures. The sequence contains the loop region and the two flanking regular secondary structures, for which the 3-D structures are known. In this work, each of the stem structures needs to be at least three residues long, and the loop regions need to contain at least two residues each.

2.1 Step 1: Search for loop motifs in known protein structural space based on the input stem structures.

First, the structure of the flanking secondary structure elements is used as a structural probe to search known protein structures for loop motifs with matched stem structure geometries regardless of the loop length. By including all loop motifs with matched stem structures, we retain the flexibility to match the query sequence to loop motifs for which the loop size might be a mismatch as defined by a secondary structure assignment program (DSSP (Kabsch and Sander, 1983) in this work).

We use the structure alignment procedure in the PrISM system (Yang and Honig, 2000a; Yang and Honig, 2000b; Yang and Honig, 2000c) to search for templates in proteins listed in PDB_SELECT25 (version Feb/2001; see Results for the rationale of selecting the template library). After the one-against-all structure alignments, a loop motif sub-database is formed to include the loop motifs that satisfied the following criteria: the stem residues of the loop motif are aligned to more than 80% of the stem residues in the probe structure and the CαRMSD for the structure align-

*To whom correspondence should be addressed.

ment is less than 2Å. These criteria have been empirically determined so as to optimize the computational efficiency and prediction capability.

2.2 Step 2: Polypeptide backbone torsion angle prediction with LSBSP1 database and a neural network model – LSBSP1+NN

The LSBSP1+NN backbone torsion angle prediction procedure has two key components: the LSBSP1 database (Yang and Wang, 2003) and a feed-forward back-propagation neural network (NN) model trained to predict backbone torsion angle with LSBSP1 database (Kuang et al., 2004).

LSBSP1 database: We only briefly describe the construction of the LSBSP1 database; more details can be found in a recently published work (Yang and Wang, 2003). The LSBSP1 database contains a total of 140106 position specific score matrices (PSSM). Each PSSM has the dimension of 9 by 20. Each of the elements in the PSSM was calculated from a sequence profile constructed with a seed 9-residue segment from a protein structure. The seed can be any 9 consecutive residue sequence segment from the non-redundant protein structures in PDB_SELECT_25 (PDB_SELECT_25 (Hobohm et al., 1992) version Feb/2001; this dataset was in agreement with the dataset in step 1).

Artificial neural network: The feed-forward back-propagation neural network (NN) model (Rumelhart et al., 1986) for backbone torsion angle prediction has been published recently (Kuang et al., 2004). In brief, the neural network takes an input of 9-residue sequence segment and torsion angle information from LSBSP1 to predict the backbone conformational state of the central residue in the sequence segment.

2.3 Step 3: Prediction of a local structural fragment of a 9-residue sequence segment

The backbone structure prediction from the LSBSP1+NN procedure described above is used to make local structure predictions for 9-residue sequence segments in the query sequence. This local structure prediction method is a slightly modified version of the previously published local structure prediction method - LSBSP1+consensus method - that has been demonstrated to predict local structure for 9-residue sequence segments with reasonable accuracy (Yang and Wang, 2003).

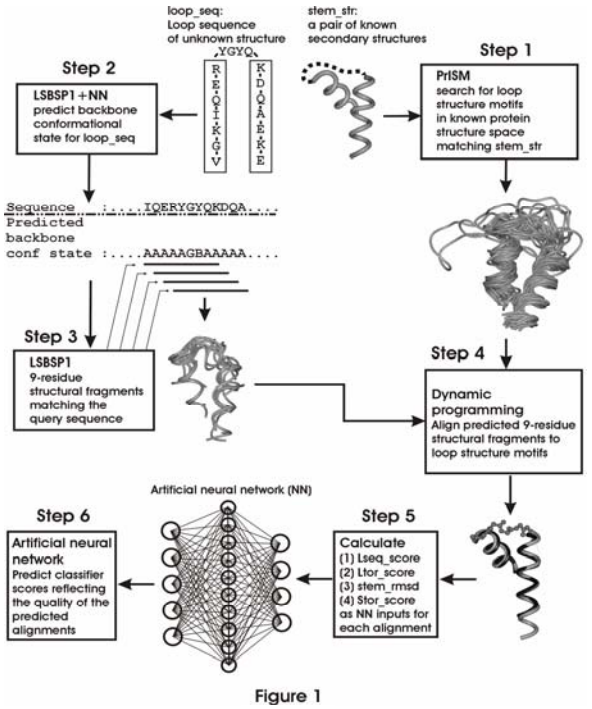

Fig. 1. The PrISM loop structure prediction flowchart. The details of the six steps of the prediction procedure are described in the Method section.

In the LSBSP1+NN method, the consensus backbone structure prediction is replaced by the output of the artificial neural network prediction. The 9-residue segment in LSBSP1 for which the PSSM in the LSBSP1 database scores against the 9-residue query sequence segment above a threshold of 16 and the backbone structure fits best to the LSBSP1+NN backbone torsion angle prediction is singled out as the predicted local structure fragment for the query sequence segment. This local structure prediction procedure is carried out for each window of 9-residue sequence segment throughout the query sequence, including the sequence regions in the flanking secondary structure elements with known 3-D structures.

2.4 Step 4: Dynamic programming alignment for a query sequence to loop motif structural templates

The local structure prediction procedure (see Steps 2 and 3) predicts one structural fragment for each of the consecutive 9-residue windows in the query sequence. We use the Smith-Waterman dynamic programming algorithm to match the predicted 9-residue structural fragments to each of the structure templates in the loop motif sub-database (see Step 1 for the derivation of the sub-database). The goal of this one-against-all procedure is to single out a structural template that matches best to the group of the predicted structural fragments.

Unlike dynamic programming sequence alignment where the scoring matrix for all residue pairs is derived from an amino acid substitution matrix such as the BLOSUM62, the scoring matrix elements for matching 9-residue structural fragments to a structural template are calculated with a simple empirical function:

$$S(a_1,a_2)=\frac{10\times\left[B-C(\alpha,\beta)-RMSD(a_1,a_2)\right]}{B} \quad (1)$$

$RMSD(a_1,a_2)$ is the root mean square deviation of backbone Cα atoms calculated from superimposing the predicted 9-residue structural fragments (a_1 is the central residue of this fragment) to the 9-residue structural fragment excised from the structural template (a_2 is the central residue of this fragment). B is the baseline parameter and $C(\alpha,\beta)$ is the baseline adjustment for the structural fragment pairs with all alpha or all beta residues. The adjustment is designed to weight down the importance of the low RMSD calculated in superimposing regular structures such as α-helices or β-strands. The parameter set (B=5Å and $C(\alpha,\beta)$=2 Å , 1 Å , and 0 Å for all-α, all-β, and others respectively) has been determined empirically to optimize a set of fragment-based sequence alignments. After the alignment of the central residues is determined and anchored, all residues are aligned with the limitation that no indels are allowed for the alignments in the current work.

2.5 Step 5: Characterize the dynamic programming sequence-template matching results

The sequence-structure alignment procedure described in step 4 predicts alignments for the query sequence to the template loop motifs. For each alignment, the three query stem residues N-terminal to the beginning of the loop region and three query stem residues immediately following the loop region are superimposed with 6 corresponding residues in the template to calculate the stem Cα RMSD; alignments failed to map all these 6 stem resides to the template loop motif are eliminated from further consideration. To make the computational procedure more efficient, we only consider the top 99 sequence-template alignments with the smallest stem Cα RMSD.

For each of the 99 predicted sequence-template alignments, four normalized attributes are calculated to characterize the mapping of the query sequence onto the template loop motif:

(1) **Lseq_score**: this attribute is calculated by summing the amino acid substitution scores according to the predicted alignment of the query loop residues to the template residues. The calculation is limited to the residues in the loop region plus two residues into the stem at each end - the boundary residues are frequently loop structure determinant. The sum is

then normalized by dividing with the maximum score possible for the residues under consideration. The maximum score is derived by summing the self amino acid substitution scores of the residues. Structure-dependent log-odds amino acid substitution matrices (Yang, 2002) are used for this calculation.

(2) **Ltor_score**: this attribute is calculated by summing the number of correctly predicted backbone torsion angle state (see Step 2) for the query loop residues (again, the loop region plus two residues into the stem at each end), and then normalized by dividing with the number of the residues.

(3) **stem_rmsd**: this attribute is the Cα RMSD(l) calculated based on the predicted alignment of the pair of the 3-residue stems separated by l loop residues to the corresponding 6 template residues and normalized by the following equation:

$$stem\_rmsd = \frac{\overline{RMSD(l)} - RMSD(l)}{\overline{RMSD(l)}} \quad (2)$$

where $\overline{RMSD(l)}$ is the average Cα RMSD calculated with superimpositions of pairs of 3-residue stems separated by l loop residues onto the loop motif templates according the predicted alignments from the procedure described in Step 4. Supplementary Material Table I shows the list of $\overline{RMSD(l)}$ as a function of loop length l.

(4) **Stor_score**: this attribute is similar to the Ltor_score, except that the attribute is calculated for the six stem residues at the immediate ends of the loop region.

2.6 Step 6: Artificial neural network trained to evaluate the prediction accuracy of the loop motif sequence-template alignments

The four normalized attributes described in Step 5 are used to judge the fitness of sequence-template alignments. We use a neural network (NN) model to combine the four attributes to predict classifier scores that reflect the accuracy of the corresponding sequence-template alignment. A minimum feed-forward back-propagation neural network model (Rumelhart et al., 1986) with 4 nodes in the input layer, 10 nodes in the hidden layer, and 5 nodes in the output layer was adequate to be trained with reasonable predictive power.

Each of the training sequences (see Results for the selection of the training sequences) with known structure was used as input for the procedure depicted in Figure 1. At the end of Step5, each of the training sequence was predicted to align to 99 different template loop motifs. Each of the alignments led to the four normalized attributes, which were used as training inputs for the NN model. The Cα RMSD of the loop region, which was calculated with the predicted alignment along with the known structure of the training sequence and the structure of the template, was encoded in five 0~1 classifier scores as the training outputs; the encoding scheme is explained in Figure 2. The encoding of the RMSD value into five 0~1 classifier scores is necessary to avoid abrupt switch of the RMSD classes due to small difference of the RMSD – a problem preventing convergence during the NN training.

On-line training to connect the four input attributes to the five output classifier scores from each of the predicted alignments has been carried out through optimizing the weights of the NN model with large sets of training cases (37378, detailed distributions are shown in Supplementary Material Table II) until convergence. The NN model contains 200 weights connecting nodes – hence more than 200 training cases can in theory train a NN model. One NN model has been trained for each loop length l ranging from 2 to 16 residues; all loops longer than 16 residues were lumped together to train one NN model in order to have enough training cases (i.e., >200 training cases; for numbers of training cases, see Supplementary Material Table II). For each NN model, five-fold cross validation (80% of the training cases were used in training and 20% of the training cases were used in testing; this process rotated five rounds to cover the complete training set) was carried out to ensure that the prediction error was consistently converged in both training sets and test sets of the five rounds of training.

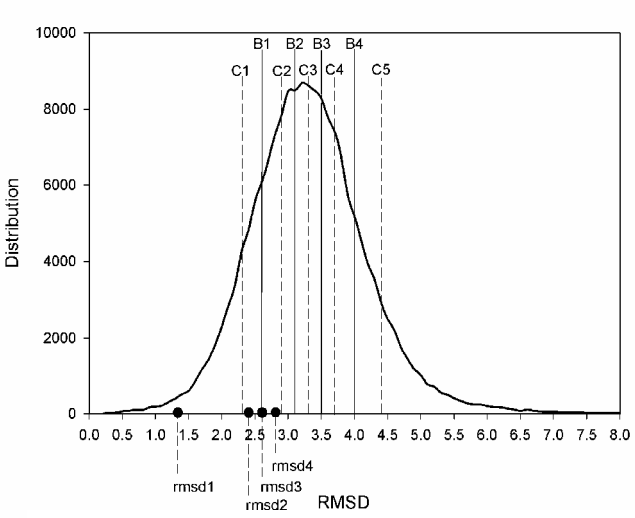

Fig.2 A scheme of encoding a RMSD value, which is calculated with a predicted alignment along with the known structure of the training sequence and the structure of the template, in five 0~1 classifier scores for NN training and prediction. The bell shape curve is the distribution of RMSD calculated from the Step 4 training sequence-template alignment predictions (99 predictions for each of the training sequences; in this case, l=8 residues, and 1973 training sequences. See Supplementary Material Table II). The RMSD population of the predictions is divided into five classes by the solid vertical lines. As shown in the figure, the boundaries of the classes (see the solid lines marked by B1, B2, B3, B4 in the figure) divide the distribution of RMSD into five equal partitions – the first class contains the lowest 20% RMSD and so on. Dashed lines marked by C1~C5 in the figure are the midpoints for the distribution of RMSD in respective classes. Supplementary Material Table I lists the values of B1~B4 and C1~C5 as functions of loop length l. For RMSD < C1, the five classifier scores are (1.0, 0.0, 0.0, 0.0, 0.0); for RMSD > C5, the five classifier scores are (0.0, 0.0, 0.0, 0.0, 1.0). Any RMSD between Ci and Ci+1 is encoded by the i^{th} and $i+1^{th}$ classifier scores; classifier scores other than these two remain zero. If RMSD=Ci, the i^{th} classifier score equals to 1 and $i+1^{th}$ classifier score equals zero; if RMSD=Bi, both i^{th} and $i+1^{th}$ classifier scores are 0.5. Linear interpolation is used to determine the i^{th} and $i+1^{th}$ classifier scores for the RMSD falls between Ci and Ci+1. A few examples of RMSD-classifier scores are shown in the figure: rmsd1 < C1 - (1.0, 0.0, 0.0, 0.0, 0.0); C1 < rmsd2 < B1- (0.8, 0.2, 0.0, 0.0, 0.0); rmsd3 = B1 - (0.5, 0.5, 0.0, 0.0, 0.0); B1 < rmsd4 < C2 - (0.2, 0.8, 0.0, 0.0, 0.0).

3 RESULTS

Protein structures from PDB_SELECT_95 (July, 2005) constituted our initial non-redundant set of protein structures. The

protein sequences in the non-redundant protein set were compared with the protein sequences in the PDB_SELECT_25 Feb/2001 (used for the data libraries of the loop prediction method in Step 1 ~ Step 4 described in the Methods section) with the Smith-Waterman sequence alignment algorithm. Protein sequences in the PDB_SELECT_95 July/2005 that are similar to at least one protein sequence in the PDB_SELECT_25 Feb/2001 by sequence identity greater than 95% were removed from the non-redundant protein set. Protein sequences in the non-redundant protein set that are similar to any of the protein sequences in the PDB_SELECT_25 Feb/2001 by no more than 35% in sequence identity were collected and half of these proteins were randomly selected as the training set for the NN models described in Step 6 of the Methods section. These training proteins were removed from the non-redundant protein set. The remaining proteins in the non-redundant protein set are the loop structure prediction benchmarking proteins, which are neither overlapped with the proteins used in the data libraries of the loop prediction method (i.e. PDB_SELECT_25 Feb/2001 used in Step1~Step4 in Figure 1), nor overlapped with the training set for the NN models (Step 6 in Figure 1).

These benchmarking proteins were divided into four sets: proteins in the 0~25% set are the benchmarking proteins that are similar to any of the protein sequences in the PDB_SELECT_25 Feb/2001 by no more than 25% in sequence identity. Proteins in the 25~50% set are the benchmarking proteins that are similar to any of the protein sequences in the PDB_SELECT_25 Feb/2001 by no more than 50% in sequence identity and that do not belong to the 0~25% test set. Similar grouping definition applies to the50~75% benchmarking set and the 75~95% benchmarking set. The numbers of benchmarking cases are 31655, 16653, 4209, 1068 for the four sets respectively with increasing sequence relationship to the knowledge basis. The distributions of the benchmarking cases as functions of loop length *l* can be found in the Supplementary Material Table II.

After training one NN model for each of the loop length with the corresponding set of training cases shown in Supplementary Material Table II, we applied the prediction procedure shown in Figure 1 to predict only one structure for each of the benchmarking cases. The prediction with the predicted classifier scores encoding the smallest RMSD was selected among the predicted sequence-template alignments.

Before benchmarking the prediction procedure in Figure 1, we benchmarked the prediction capacities of the NN models. For each of the sequence-template matches from Step 4 of the prediction procedure in Figure 1, the structures of the query sequence and the template were known and thus the RMSD class could be assigned based on the predicted alignment and the scheme shown in Figure 2. Comparing the RMSD class with the NN predicted RMSD class on the basis of the classifier scores, we were able to calculate the Matthews correlation coefficient:

$$C = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \qquad (3)$$

,where *tp, tn, fp, fn* are true positive, true negative, false positive, and false negative respectively. For each of the benchmarking cases, 99 predictions contributed to the *tp, tn, fp, fn* calculation. The correlation coefficients for the 0~25% benchmarking set are 0.467, 0.154, 0.123, 0.230, 0.389 for RMSD class 1 to 5 respectively. This result indicated that the best sequence-template alignments (predicted to be in the first RMSD class) were able to be identified with much higher prediction accuracy. Supplementary Material Table III shows all the correlation coefficients derived from the training set and the benchmarking sets shown in Supplementary Material Table II. The comparable results between the training set and the 0~25% benchmarking set indicate that the trained NN models were converged but not over-trained. As expected, the prediction accuracy increases with increasing sequence relationship to the data libraries used in the predictions. The correlation coefficients shown in Supplementary Material Table III also indicate that predictions in the lowest RMSD class (class 1) were best correlated with the true RMSD class (with the correlation coefficients 0.47, 0.55, 0.61. 0.62 respectively for the four benchmarking groups with increasing sequence relationship to the knowledge basis). Thus the first predicted classifier score from the first output node of the NN models can be used as a quantitative indicator for the prediction confidence level as shown below.

To benchmark the overall prediction accuracy of the loop structure prediction procedure in Figure 1, we use two types of accuracy measure to characterize the predicted sequence-template alignment: (1) the backbone heavy atom NCαCO-RMSD based on the superimposition of the N, Cα, C, O atoms for the residues in the query loop region to the corresponding atoms in the template; (2) the percentage of the query residues for which the backbone torsion angle states (a total of 17 states in Ramanchandran plot as defined by Oliva et al (Oliva et al., 1997a)) are in agreement with those of the corresponding residues in the template.

Four panels in Figure 3 (Figure 3(a)~3(d)) show the results for the four benchmarking sets with increasing sequence similarity (0~25%, 25~50%, 50~75%, 75~95%) to the knowledge basis of the prediction method. In each of the panels, the bottom X-Y plot shows the average NCαCO-RMSD; the middle histogram shows the average percentage of residues with correctly predicted backbone torsion angle state; the top histogram shows the percentage coverage of the four prediction confidence level (see below). In each of the panels, four sets of data show the prediction accuracy with increasing prediction confidence level. As described above, the confidence level is determined by the classifier scores from the five output nodes of the NN models. The first set of data contains the best predictions regardless of the predicted RMSD class – confidence level 1; the second set of data limits to the predicted first RMSD class – confidence level 2; the third set of data limits to the predicted first RMSD class and with the first node output classifier score greater than 0.5 – confidence level 3; the fourth set of data limits to the predicted first RMSD class and with the first node output classifier score greater than 0.7 – confidence level 4.

As shown in Figure 3, the prediction accuracy invariably increases with increasing confidence level, suggesting that the NN models succeeded, to certain extend, in optimally combining the four normalized scores in Step 5 into a quantitative confidence level for the prediction results. The Figures also show, as expected again, that the prediction accuracy increases with increasing sequence relationship to the knowledge basis of the prediction method.

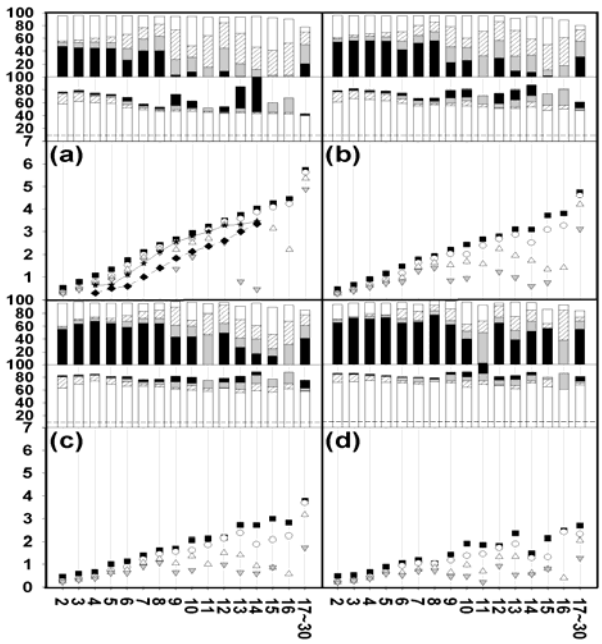

Fig. 3 Average NCαCO-RMSD in Å (X-Y plot in each panel), average percentage of residues with correctly predicted backbone torsion angle state (middle histogram in each panel), and the percentage coverage of the four prediction confidence level (top histogram in each panel) plotted as functions of loop length *l*. Panels (a) to (d) show the results from benchmarking sets of 0%~25%, 25%~50%, 50%~75%, and 75%~95% respectively. For the X-Y plot of each panel, the solid squares are the data set predicted with confidence level 1; the empty circles with confidence level 2; the empty triangles with confidence level 3; the grey triangles with confidence level 4. For the two histograms of each panel, white/lined/grey/dark bars represent data corresponding to the prediction confidence level 1/2/3/4 respectively. The solid line (solid stars) and the dashed line (solid diamonds) in panel (a) are reproduced from Fernandez-Fluentes et al. 2006 for comparison with the predictions in this work. The solid line is the results from ArchPred – a knowledge based prediction method; the dashed line is the results from ModLoop *ab initio* predictions. The dashed line in the middle histograms shows the accuracy baseline from random prediction of the backbone torsion angle state.

As shown in the top histograms in the panels of Figure 3, the coverage of the predictions decreases with increasing prediction confidence level. For loop structures of size > 9 residues, predictions with confidence level 4 were useful in practical structure prediction tasks only for the query sequences that are related to the knowledgebase with >50% sequence identity. For loop sequences with sequence similarity <25% to the knowledgebase, the structure predictions with confidence level 4 were rare for loop size > 9 residues. Two factors accounted for the lack of prediction confidence here: first, common loop motifs shared by unrelated structures were increasingly rare for loops with increasing size, and consequently, even with the best alignment algorithm, the sequence-template matches deteriorated in longer loops (>9 residues) judging by the stem structure constrains; second, as the correct predictions became difficult, the training of the NN models was biased toward the majority of the predictions with inferior confidence levels for longer loops. For the 0~25% group, predictions with confidence level 2 were better balanced with prediction accuracy, which generally reached the average accuracy marked by the C1 values in Supplementary Material Table I, and the level of predictability.

Overall, the four attributes as the input for the NN models contributed unequally to the output predictions of the confidence level. The most decisive attributes were the *stem_rmsd* and the *Stor_score*. The former reflects the fitness of the stem structures, which inevitably affect the prediction accuracy of the loop structures connecting the stems. The later indicates the accuracy of the predicted sequence-template alignments: a mismatch of the query stem torsion angles with the corresponding torsion angles in the template is a sign for misalignment, which most likely results in mismatch in overall structures of the query and the template. The *Lseq_score* attribute was important only for query-template matches with high sequence relationship. The *Ltor_score* attribute contributed the least among the four attributes to the decisions on the prediction quality.

The average CPU time on a 3GHz processor for benchmarking cases is on the order of 60 seconds regardless of the loop length. The computational time is expected to increase linearly with increasing size of the databases.

4 DISCUSSION

Panel (a) in Figure 3 compares the prediction accuracies of the prediction method in Figure 1 with prediction accuracies from previous published methods. In general, the *ab initio* method ModLoop predicted more accurate loop structures with limited test cases (50 for each loop length). Only the predictions with confidence level 4 are comparable with the ModLoop prediction accuracy, although the prediction coverage of this confidence level is very low for longer loops ($l > 8$, see Figure 3). The ArchPred is the most recent knowledge-based loop prediction method and perhaps is the most accurate prediction procedure of its kind in the public domain according to the published work (Fernandez-Fuentes et al., 2006). Overall, our predictions with confidence level 3 match with the performance of the ArchPred predictions.

Although the comparison of the current work with the ArchPred prediction is relevant in terms of methodology, there are discrepancies between the two predictions: First, the ArchPred results shown in Figure 3 were averaged over 50 predictions for each loop length, while our prediction procedure was benchmarked with at least hundreds, and in shorter loop length ($l < 11$ residues), thousands benchmarking cases (see Supplementary Material Table II). Second, the prediction confidence level (and the coverage level) of the ArchPred prediction results reproduced in Figure 3 are not available from the published work; consequently, the prediction accuracies are difficult to be compared at the same coverage level. Third, The ArchPred loop predictions have an additional constrain: using the whole protein structure that contains the query loop sequence to eliminate the predicted loop structures that clash with the

rest of the parent structure. In our loop structure prediction, the parent structure constrain is not used due to the consideration that in realistic comparative modeling procedures, the parent structures could be less well defined and thus could provide misleading models. Fourth, the knowledge basis of our prediction method is derived from a non-redundant protein set published before Feb/2001 with pairwise sequence similarity less than 25%, so that we could form large non-overlapping benchmarking sets, in particular for longer loops (see Supplementary Material Table II). On the other hand, the ArchPred prediction used entries in PDB released in 2005 as its knowledge basis for the predictions. It is conceivable that updating our data libraries with the most recent PDB release would improve the prediction capabilities. A website server with the updated data libraries will be available to the public domain (publication for this web server is currently in preparation).

Loop prediction becomes increasingly difficult with increasing loop length. Still, as shown in Figure 3, we are able to predict loop structures up to ~30 residues in length. We examined the accurate predictions (NCαCO-RMSD<2Å) in the 0~25% test set with loop length greater than 11 residues. For loop length between 11 and 13 residues, more than half of the pairs of the test sequences and the loop motif templates came from proteins with different SCOP super families (Murzin et al., 1995). For loops with more than 14 residues, correctly predicted sequence-template pairs were all from proteins in the same SCOP super families, although the sequence similarity were usually very low. These results indicated that conserved loop structures from remotely related homologous proteins were useful templates for loop modeling if properly identified, and that stem structures from distantly related proteins are occasionally well-conserved. The findings also suggested that structures of longer loops were so divergent that only stem structure constrains and conserved sequence patterns among homologous proteins provided reliable hints for loop structure similarity. It is difficult to conceive that loop structures beyond certain size (for example, >14 residues) would ever be sorted into limited structural families, even with the upcoming structural data in the foreseeable future. The reason is that the requirement of the size of the structural database is simply too huge to cover all the possible structural patterns of loops of increasing size. Thus, the structures of long loops could only be reliably modeled when homologous protein structures with similar loop structures were identified.

The major limitation of the knowledge-based loop structure prediction method is the coverage of the loop structures from known protein structures. Using structural alignment algorithm PrISM (which only considers structural similarity), we found that 33%, 42%, 47% and 54% of the 8-residue loop motifs in the 0~25%, 25~50%, 50~75%, and 75~95% respective benchmarking sets could find templates with identical Oliva et al. torsion angle states throughout the loop residues in the proteins listed in PDB_SELECT_25 Feb/2001. For 16-residue loop motifs, the fractions drop to 1%, 6%, 15%, 36% for the 0~25%, 25~50%, 50~75%, and 75~95% respective benchmarking sets. If we demand the most stringent criterion for sequence-template match (i.e., 100% correctly predicted torsion angle states), these fractions would be the upper limits of our prediction method due to the limited coverage of the protein data set used as the loop structure templates. The coverage limitation would be partially alleviated after updating our data libraries. Moreover, the coverage of the loop motifs in the known protein space will continue to expand as the protein structures are determined due to high throughput experiments. Nevertheless, the coverage of the loop structure in proteins of unknown structure remains the bottleneck for the current knowledge-based loop structure prediction methods.

ACKNOWLEDGEMENTS

A.-S. Yang would like to thank National Health Research Institutes for the grant: NHRI-EX95-9525EI, and for financial support from Genomics Research Center, Academia Sinica.

REFERENCES

Burke, D. F., and Deane, C. M. (2001). Improved protein loop prediction from sequence alone. Protein Eng *14*, 473-478.

Burke, D. F., Deane, C. M., and Blundell, T. L. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. Bioinformatics *16*, 513-519.

Carter, P., Andersen, C. A., and Rost, B. (2003). DSSPcont: Continuous secondary structure assignments for proteins. Nucleic Acids Res *31*, 3293-3295.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng *6*, 377-382.

Donate, L. E., Rufino, S. D., Canard, L. H., and Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. Protein Sci *5*, 2600-2616.

Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J., and Oliva, B. (2004). ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res *32*, D185-188.

Fernandez-Fuentes, N., and Fiser, A. (2006). Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol *6*, 15.

Fernandez-Fuentes, N., Oliva, B., and Fiser, A. (2006). A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res *34*, 2085-2097.

Fernandez-Fuentes, N., Querol, E., Aviles, F. X., Sternberg, M. J., and Oliva, B. (2005). Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. Proteins *60*, 746-757.

Fiser, A., Do, R. K., and Sali, A. (2000). Modeling of loops in protein structures. Protein Science *9*, 1753-1773.

Heuser, P., Wohlfahrt, G., and Schomburg, D. (2004). Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. Proteins *54*, 583-595.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). Selection of representative protein data sets. Protein Science *1*, 409-417.

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

Kuang, R., Leslie, C. S., and Yang, A. S. (2004). Protein backbone angle prediction with machine learning approaches. Bioinformatics *20*, 1612-1621. Epub 2004 Feb 1626.

Lessel, U., and Schomburg, D. (1997). Creation and characterization of a new, non-redundant fragment data bank. Protein Eng *10*, 659-664.

Lessel, U., and Schomburg, D. (1999). Importance of anchor group positioning in protein loop prediction. Proteins *37*, 56-64.

Li, W., Liang, S., Wang, R., Lai, L., and Han, Y. (1999). Exploring the conformational diversity of loops on conserved frameworks. Protein Eng *12*, 1075-1086.

Michalsky, E., Goede, A., and Preissner, R. (2003). Loops In Proteins (LIP)--a comprehensive loop database for homology modelling. Protein Eng *16*, 979-985.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology *247*, 536-540.

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997a). An automated classification of the structure of protein loops. J Mol Biol *266*, 814-830.

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997b). An automated classification of the structure of protein loops. Journal of Molecular Biology *266*, 814-830.

Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B. M., Eramian, D.*, et al.* (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res *34*, D291-295.

Rufino, S. D., Donate, L. E., Canard, L. H., and Blundell, T. L. (1997). Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol *267*, 352-367.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature *323*, 533-536.

Wojcik, J., Mornon, J. P., and Chomilier, J. (1999). New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. Journal of Molecular Biology *289*, 1469-1490.

Yang, A. S. (2002). Structure-dependent sequence alignment for remotely related proteins. Bioinformatics *18*, 1658-1665.

Yang, A. S., and Honig, B. (2000a). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structure alignment and quantitative measure for protein structural distance. J Mol Biol *301*, 665-678.

Yang, A. S., and Honig, B. (2000b). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol *301*, 679-690.

Yang, A. S., and Honig, B. (2000c). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J Mol Biol *301*, 691-712.

Yang, A. S., and Wang, L. (2002). Local structure-based sequence profile database for local and global protein structure predictions. Bioinformatics *18*, 1650-1657.

Yang, A. S., and Wang, L. (2003). Local structure prediction with local structure-based sequence profiles. Bioinformatics *19*, 1267-1274.