Chapter IV - Knowledge-based Loop Modeling: Application of *ArchDB* in Loop Modeling

4.1 Abstract

In protein structure prediction, a frequent problem is defining the structure of a loop, fitted between two secondary structures. This problem is common for homology modeling and *ab initio* structure prediction. In our previous work, we presented a classification database of structural motifs, Arch DB. The database contains 451 classes with information about ϕ–φ angles in the loops and 1492 sub-classes with cover both the ϕ–φ angles in the loop and the relative locations of the bracing secondary structures. Here, our aim is to know how useful could be the sequence information included in our database for loop structure prediction and the identification of a small subset for the inclusion in a predicted structure and subsequent evaluation of the tertiary fold. For this reason, a jack-knife test was made, removing the loops belonging same SCOP super-family, and predicting afterwards against recalculated profiles only take into account the sequence information. Two sequence profiles were used, a HMM profile and a PSSM derived from Psi-blast. If we consider the top 20 classes out of 451 the accuracy is 85,7% while if we consider the top 20 subclasses out 1492 the accuracy is 72,3%. A *Sign test* was performed to assess the significance of the prediction compared with a random prediction. Because our structural loop database discriminate between $\beta\beta_{\text{hairpins}}$ and $\beta\beta_{\text{links}}$ for β-loop-β motifs, we have found that we can distinguish between a $\beta\beta_{\text{hairpins}}$ or $\beta\beta_{\text{links}}$. The prediction of supersecondary structures formed by consecutive loops is also discussed.

4.2 Introduction

The finished projects on large-scale sequencing of genomes and many others that are still in progress have generated a vast amount of gene and protein sequence data. Besides, recent improvements in the techniques of structure determination at atomic level, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, have enhanced the quality and speed of the determination of 3D protein-structures. Nevertheless, there is a large difference between the number of known protein sequences (~ 1 million)(Boeckmann et al. 2003) and the number of available protein structures (~ 20 000)(Berman et al. 2002).

In the absence of an experimentally determined structure, *ab initio(Simons et al. 1997)* and threading methods(Domingues et al. 2000) or comparative modeling methods(Marti-Renom et al. 2000) can sometimes provide a useful 3D model and fill the gap between sequence and structure space. In general, these methods tend to correctly predict the protein core when the structure of a close homologue of the target protein is available, but not the loop regions. Up to date, the prediction of structure of a loop remains unsolved and the recent improvements of the performance of fold prediction and homology modeling methods in successive CASP experiments(Venclovas et al. 2001) have not proved to be successful in loop model building. Modeling of loop conformation is neither trivial nor unimportant.

Functional differences between the members of the same protein family are usually a consequence of the structural differences on the protein surface. In a given fold, structural variability is a result of substitutions, insertions and deletions of residues between members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold. Thus, loops often determine the functional specificity of a given protein framework, contributing to active and binding sites(Fetrow 1995).

Loop prediction can be seen as a mini protein-folding problem. The correct conformation of a given segment of a polypeptide chain has to be calculated from the sequence of the segment influenced by the core limb regions that span the loop and by the structure of the rest of the protein that cradles the loop. Many loop-modeling procedures have been described. Similarly to the prediction of whole protein structures, there are *ab initio* and threading methods (also named conformational search)(Fine et al. 1986; Moult and James 1986; Bruccoleri and Karplus 1987), database search methods (also named knowledge-based)(Jones and Thirup 1986; Chothia and Lesk 1987)and procedures that combine both approaches(Chothia et al. 1986; Wlijmen and Karplus 1997).

The *ab initio* loop prediction is based on a conformational search or enumeration of conformations in a given environment, guided by a scoring or energy function. There are many such methods, exploiting different protein representations, energy function terms, and optimization or enumeration algorithms(Fiser et al. 2000; Tosatto et al. 2002). The limitation of this approach is the accuracy of the applied scoring function that often are not accurate enough to properly rank the many alternative conformations, especially on long loops.

The database approach to loop prediction consists of finding a segment of main chain that fits between the two stem regions of a loop(Jones and Thirup 1986; Sibanda et al. 1989; Levitt 1992). The search is performed on a database of known protein structures. Usually, many different alternative segments fitting between the two secondary structures are obtained. All template segment conformations are sorted according to a geometric criteria and sequence similarity with the target loop. The selected segments are superposed and annealed between the anchoring extremities. The database search approach to loop modeling is sufficiently efficient when a specific set of loops is created to address the modeling of similar type of loops, such as β-hairpins(Sibanda et al. 1989) or the hypervariable regions in immunoglobulins(Chothia and Lesk 1987). There have been several attempts to classify loop conformations into more general categories, thus extending the possible applications of the use of key residues as an approach to additional cases(Ring et al. 1992; Oliva et al. 1997; Rufino et al. 1997; Oliva et al. 1998; Espadaler et al. 2004).

The database methods are limited by the exponential increase in the number of possible conformations in agreement with the ring closure as a function of loop length. Only segments of 7 or less residues had most of their conceivable conformations present in the database of known protein structures(Fidelis et al. 1994) lying on the correct positioning of the anchor groups(Lessel and Schomburg 1999). Therefore, the completeness of the database is the major obstacle for its use on modeling. However, a recent work published by Du et al.(Du et al. 2003) argued that there exist sufficient coverage to model even a novel fold using fragments from Protein Data Bank, as the current database of known structures has increased enormously in the last few years.

In a recent work, we developed an automatic method of classification of protein loops based on loop conformation and bracing secondary structure orientation(Oliva et al. 1997; Espadaler et al. 2004) on a non-redundant database of proteins. Besides, Blundell and coworkers(Donate et al. 1996; Burke and Deane 2001; Deane and Blundell 2001) have employed their loop database in loop structure prediction with encouraging results.

Our work is a method for loop and local structure prediction between two regular secondary structures. The loop classification clusters the loop conformation and the geometry of the super-secondary structure (or structural motif) that defines the relative positions of the flanking secondary structures. Hitherto, the prediction provides the information of the loop conformation plus the relative location between secondary structures, which can be used on a partial model building by means of loop blocks and in the evaluation of a model or a predicted fold.

A large body of evidence suggests that protein structural information is frequently encoded in local sequences, and that folds are mainly made up of a number of simple local units of super-secondary structural motifs, consisting of a few secondary structures and their connecting loops. The reports of Salem et al.(Salem et al. 1999), Wood & Pearson(Wood and Pearson 1999) and Lupas et al.(Lupas et al. 2001) suggested that folds are mainly made up of a number of simple local units of super-secondary structural motifs, formed by few secondary structures connected by loops. Two applications can be observed from this prediction: (i) in comparative modeling, offering useful information about loop conformation; and (ii) in *ab initio* fold prediction, using local structure prediction from sequence for the prediction of protein structure(Bonneau et al. 2001; Yang and Wang 2002) by means of the combination of supersecondary structures or in the final evaluation of a predicted fold.

4.3 Material and Methods

4.3.1 Loop database

The database of loops was built using a list of protein domains derived from SCOP 40(Chandonia et al. 2002) of the 1.61 release of SCOP(Conte et al. 2002). The set of loops was classified in classes and subclasses with the program ArchType(Oliva et al. 1997) forming the database ArchDB presented in http://sbi.imim.es (Espadaler et al. 2004). In descending order, ArchDB is structured in three levels of hierarchy: (i) At the top of the classification, motifs were identified according to the bracing secondary structure type ($\alpha-\alpha$, $\beta-\beta$ links, $\beta-\beta$ hairpins, $\alpha-\beta$ and $\beta-\alpha$); (ii) at *class* level, motifs are grouped according to the loop size and (ϕ,ψ) loop conformation; and (iii) at sub-class level motifs are grouped according to the loop size, (ϕ,ψ) loop conformation and orientation of secondary structures (geometry). According to these definitions two different types of prediction were made over query loops: prediction of structural class and prediction of structural sub-class.

4.3.2 Sequence profiles calculation

Two types of sequence-profiles were calculated with the multiple alignment of the sequences of the loops extracted from classes and subclasses from the ArchDB classification (see figure 3.1 for a general overview of the process): (i) a position specific scoring matrix (PSSM) obtained with the method described by Altschul et al.(Altschul et al. 1997) for PSI-BLAST and adapted on the server 3D-PSSM(Kelley et al. 2000), and (ii) a Hidden Markov Model profile using HMMER version 2.0(Eddy

1998). The profiles were calculated using the maximum common length of aligned residues of each loop plus its bracing secondary structures, or using only the residues belonging to the loop (as defined by the DSSP program(Kabsch and Sander 1983) plus two residues at each side flanking the loop. In total, four profiles were generated for each subclass and class of loops: 1) the PSSM profile with the sequences of the loops, including the flanking secondary structures (FP profiles); 2) the PSSM profile with the sequences of the loops plus only two flanking residues at each extremity (P profiles); 3) the HMM profile with the sequences of the loops, including the flanking secondary structures (FH profiles) and 4) the HMM profile with the sequences of the loops plus only two flanking residues at each extremity (H profiles).

4.3.3 Jack-knife test

A Jack-knife test was performed to asses the validity of the prediction of loops. The scheme of the test is shown in Figure 3.1, where the sequence of each loop on the database of classified loops is used as query to search on the cluster of loops (searching space). The procedure implies to remove all the sequence/structure information of the query loop being redundant or homologous (even for remote homology with the loops in clusters an use it on the calculation of profiles). Therefore, all motifs belonging to the same SCOP super-family of the query were discarded from the searching space. Additionally those remaining subclasses with less than three motifs were removed too (in order to avoid meaningless clusters). This procedure assured that no homologies (even remote) would exist between a query loop and the searching space used for the class or subclass prediction of the query. In addition, 100 motifs selected at random within the four classes of loop (100 αα loops, 100 βα loops, 100 αβ loops and 100 ββ loops) were used as queries for which homologous loops from the proteins with the same SCOP super-family of the query were not removed from the searching space, hereafter named searching space with homologs. This experiment will allow us to compare our results with similar works less restrictive on the definition of the test data set.

In order to assign the secondary structures flanking the query loop we used the program PSIPRED(Jones 1999). The secondary structure prediction of all protein chains used for this study was calculated with PSI-PRED and this was used to define the flanking regions of each query loop. The prediction of the secondary structure flanking the loop was considered correct if the assignation differed from the real location around the loop by less than 2 residues on each side (C- and N- terminal) of the loop.

The sequence of the loop was extracted from the coil region between the predicted secondary structures and aligned with the sets of profiles for subclasses (and classes) of loops with the same length ± 2 residue (i.e. if the prediction is made for a motif of length 3, the alignment will be made against profiles containing loops with lengths between 1 and 5). The scores obtained for the alignments between the query and the profiles were used to calculate a mean value (μ) and standard deviation (σ). A Zscore was calculated with the equation 1:

$$Zscore = \frac{Score - \mu}{\sigma} \qquad (1)$$

All Zscores of the query loop were ranked and the top hits were taken for the prediction of the class or subclass conformation. In order to test the specificity of the prediction the first hit, the top 5 hits and the 20 top hits were considered as probable solutions (positives).

The statistical significance of the results was evaluated using the *Sign-Test*. The positive samples obtained from the prediction were compared with a random prediction obtained by randomly selecting one, five or twenty classes of subclasses as candidates. A success is considered when is found the correct subclass of class conformation on the selection (i.e. the class or subclass from where the query loop was removed). If the correct subclass/class was predicted by the Zscore but not by random, a label "+" was given to the observation, while for the opposite case, where the prediction is successful by random but not by Zscores, a label "-" was given. We define n+ as the total of positive labels and n- as the total of negatives labels. The distribution of +/- labels was compared with the binomial distribution with probabilities 0.5 for statistical significance (which implies that sequence profiles and random approaches are equally predictive, also known as *null hypothesis*). If the sum of + and – labels is N ($N=n^{+} + n^{-}$) and k is the smallest of n+ and n-, the two-tailed probability of holding the true null hypothesis is twice the probability of finding k or less cases, and the sign-test significance is approximated to the two-tailed probability by equation 2:

$$p(two-tailed) \leq 2 * \sum_{i=0}^{k} \frac{N!}{i!(N-i)!} * 0.5^{N} \qquad (2)$$

4.3.4 Discrimination between $\beta\beta_{\mathbf{hairpins}}$ and $\beta\beta_{\mathbf{links}}$

The loops between β-strands where the PSI-PRED prediction is correct were used to test the potential discrimination between β-hairpins and β-links. A prediction of loop type (i.e. β-link or β-hairpin regardless of class or subclass) was made in order to explore the possible application to discriminate between both types.

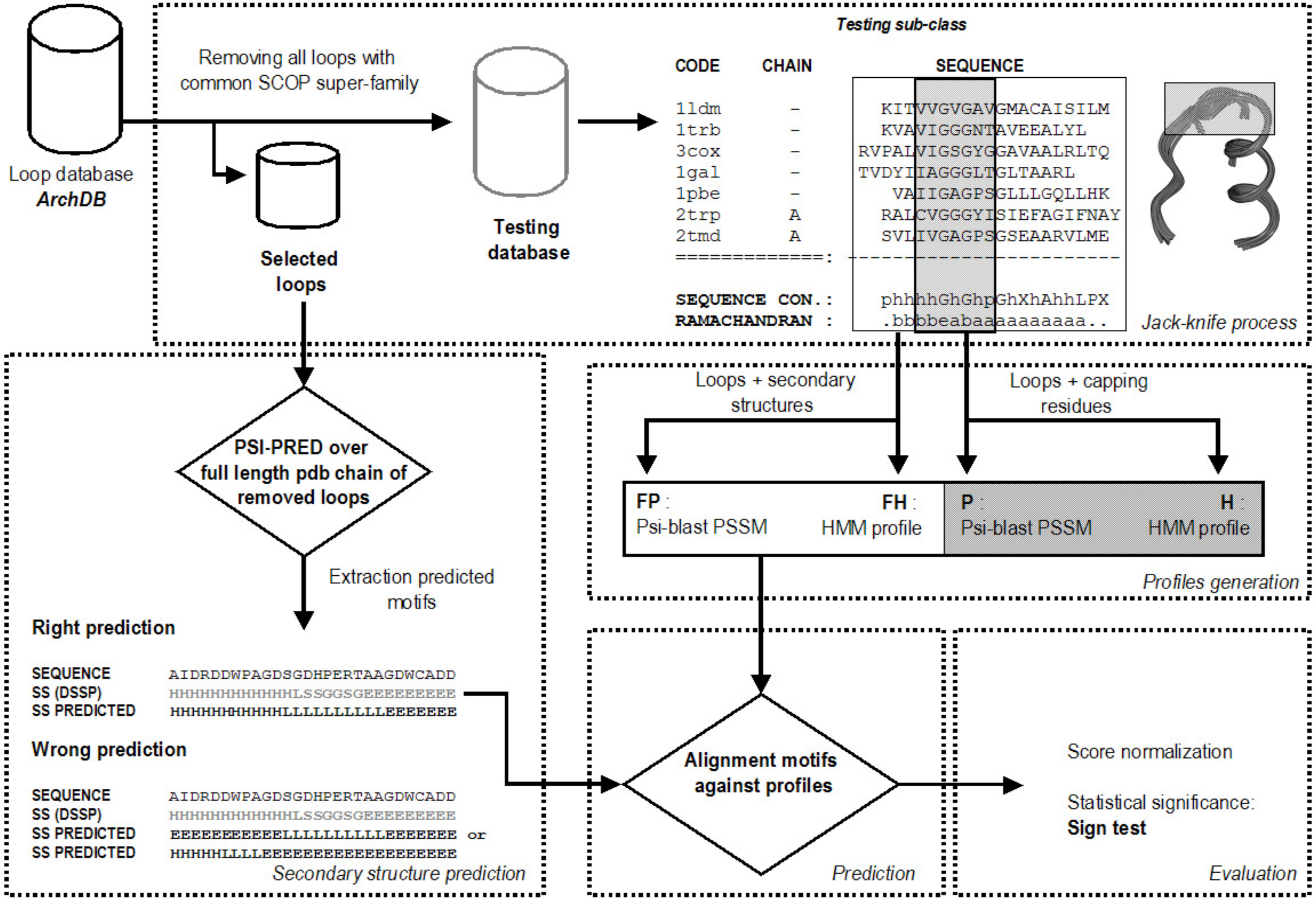

Loop database
ArchDB
Removing all loops with common SCOP super-family
Selected loops
Testing database
Testing sub-class
CODE
CHAIN
SEQUENCE
1ldm - KITVVGVGAVGMACAISILM
1trb - KVAVIGGGNTAVEEALYL
3cox - RVPALVIGSGYGGAVAALRLTQ
1gal - TVDYIIAGGGLTGLTAARL
1pbe - VAIIGAGPSGLLLGQLLHK
2trp A RALCVGGGYISIEFAGIFNAY
2tmd A SVLIVGAGPSGSEAARVLME
=============: ----------------------
SEQUENCE CON.: phhhhGhGhpGhXhAhhLPX
RAMACHANDRAN : .bbbbeabaaaaaaaaaa..
Jack-knife process
Loops + secondary structures
Loops + capping residues
FP : Psi-blast PSSM
FH : HMM profile
P : Psi-blast PSSM
H : HMM profile
Profiles generation
PSI-PRED over full length pdb chain of removed loops
Extraction predicted motifs
Right prediction
SEQUENCE AIDRDDWPAGDSGDHPERTAAGDWCADD
SS (DSSP) HHHHHHHHHHHHLSSGGSGEEEEEEEEE
SS PREDICTED HHHHHHHHHHHLLLLLLLLLLEEEEEEE
Wrong prediction
SEQUENCE AIDRDDWPAGDSGDHPERTAAGDWCADD
SS (DSSP) HHHHHHHHHHHHLSSGGSGEEEEEEEEE
SS PREDICTED EEEEEEEEEEELLLLLLLLLLEEEEEEE or
SS PREDICTED HHHHHLLLLEEEEEEEEEEEEEEEEEEE
Secondary structure prediction
Alignment motifs against profiles
Prediction
Score normalization
Statistical significance:
Sign test
Evaluation

Figure 4.1. Flowchart of the methodology for prediction and analysis.

4.4 Results

The loop structural classification, ArchDB(Espadaler et al. 2004), employed in this study, includes 12665 motifs clustered in 451 classes and 1492 sub-classes(see figures 4.2a, 4.2b for distribution of class and subclass vs loop length). The Jack-knife process generated a test set of 10492 motifs for which PSI-PRED correctly predicted the type of connected secondary structures (α or β), loop length and loop location for 6585 motifs. For 2708 motifs the prediction failed to assign the correct bracing secondary structure and for 1199 motifs the assignment of loop residues or loop location was erroneous. The set of queries (test set) was formed by the 6585 motifs for which the prediction of the secondary structure flanking the loop was correct plus the 1199 motifs where the assignation was partially correct. In figure 4.3 it is shown the distribution of loops vs length, where the $\alpha\beta$ motifs are the most abundant, especially for lengths between 3 and 4.

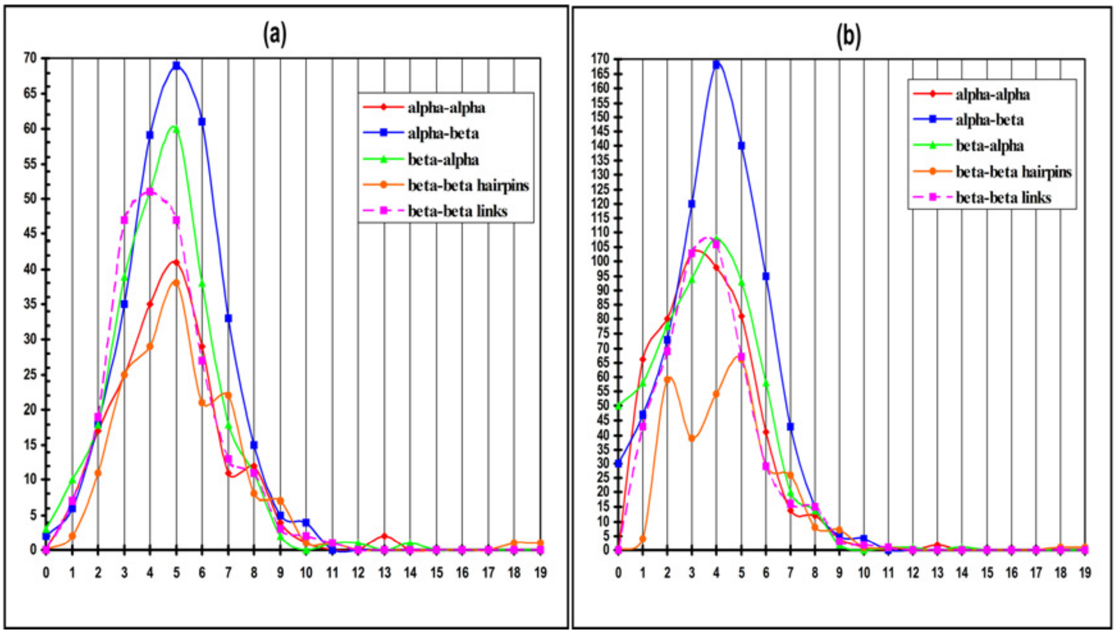


Figure 4.2(a,b). (a) Number of classes vs. loop length in ArchDB, classes are shown according to different secondary structure types. (b) Number of subclasses vs. loop length in ArchDB. Number of subclasses is shown according to different secondary structure types.

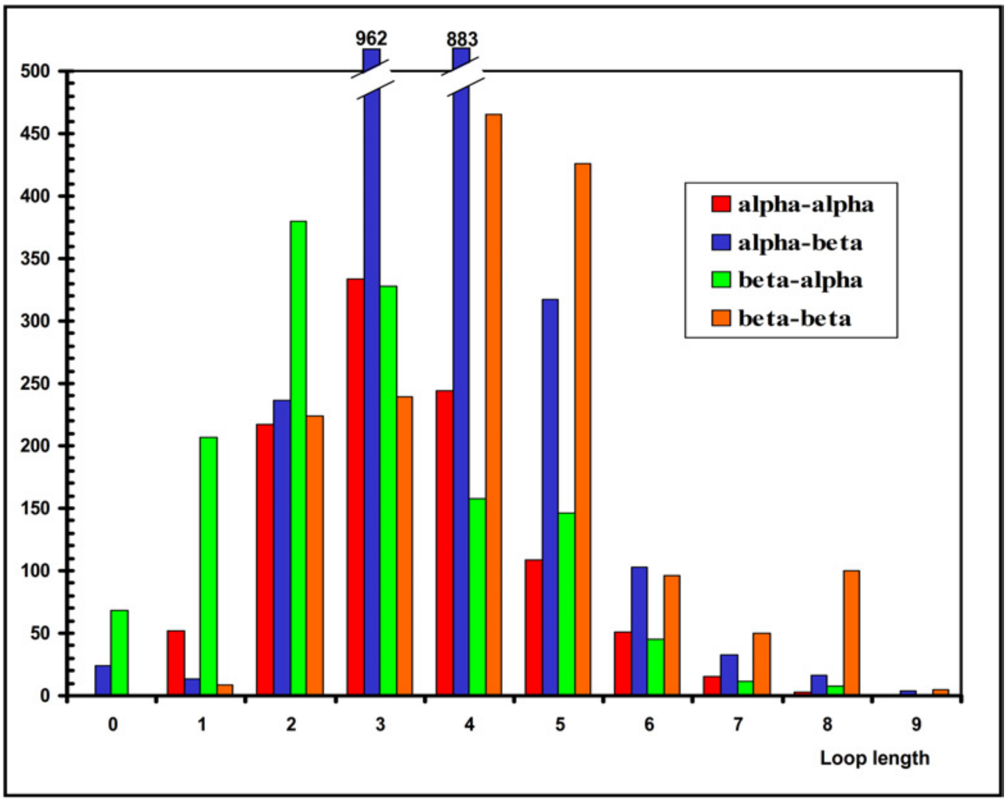


Figure 4.3. Distribution of predicted loops vs. length. Lengths of query loops are defined as the number of coil residues between flanket secondary structures predicted by PSI-PRED.

4.4.1 Overview of results

Table 4.1 shows the values of accuracy for class and sub-class prediction over the test set of queries. The results obtained with HMM profiles are better than with PSSM derived profiles. Also, the accuracy of the prediction was not highly improved by using the full sequence length of the super-secondary alignment (FP and FH profiles) rather than using the loop sequence plus two residues (P and H profiles). In addition, the values of accuracy for 400 motifs selected at random (100 $\alpha\alpha$ loops, 100 $\beta\alpha$ loops, 100 $\alpha\beta$ loops and 100 $\beta\beta$ loops) and predicted on the searching space with homologs, are also exposed. The study of accuracy of prediction was made based in the length of the loop and type of flanking secondary structures.

Table 4.1. Percentages of accuracy of class and sub-class prediction using the four types of sequences profiles, FP, P, FH and H.

Profile	RANK	% accuracy of prediction on searching space without homologs[a]		% accuracy of prediction on searching space with homologs (only for 400 loops)[b]	
		Class	***Sub-class***	***Class***	***Sub-class***
FP	1	15,9 (13,5)	5,9 (3,4)	43,1	14,5
	5	47,7 (42,3)	23,1 (19,5)	75,1	42,3
	20	81,1 (79,4)	53,8 (51,2)	98,8	85,2
P	1	15,3 (12,9)	4,4 (3,1)	40,2	12,2
	5	45,8 (40,8)	19,5 (16,9)	73,2	39,8
	20	82,7 (80,1)	55,5 (53,2)	97,2	87,3
FH	1	21,2 (19,9)	10,5 (8,2)	51,8	26,4
	5	54,8 (52,1)	37,1 (35,4)	85,4	57,6
	20	85,7 (81,2)	72,3 (70,7)	100,0	95,8
H	1	18,6 (16,6)	6,4 (4,3)	48,7	17,2
	5	54,1 (51,8)	30,3 (28,6)	81,4	50,2
	20	81,7 (80,3)	66,3 (64,5)	100,0	92,5

[a] Values of accuracy obtained in a searching space without homologs and [b] searching space with homologs are shown. Between parenthesis values of accuracy including these motifs where the assignation of secondary structure boundaries by PSI-PRED was in part correct (i.e. location of loop).

4.4.2 Class prediction

Although ArchDB database contains classes with loops up to nineteen residues long (see figure 4.2a), theses classes are poorly populated, therefore these classes were discarded. Thus, class prediction was made for $\alpha\alpha$ motifs from length 1 to 8, length 0 to 10 for $\alpha\beta$ motifs, length 0 to 8 for βα motifs and 1 to 9 for ββ motifs. The information provided by the structural class could be informative since a structural class implies different sub-classes that share a common loop conformation.

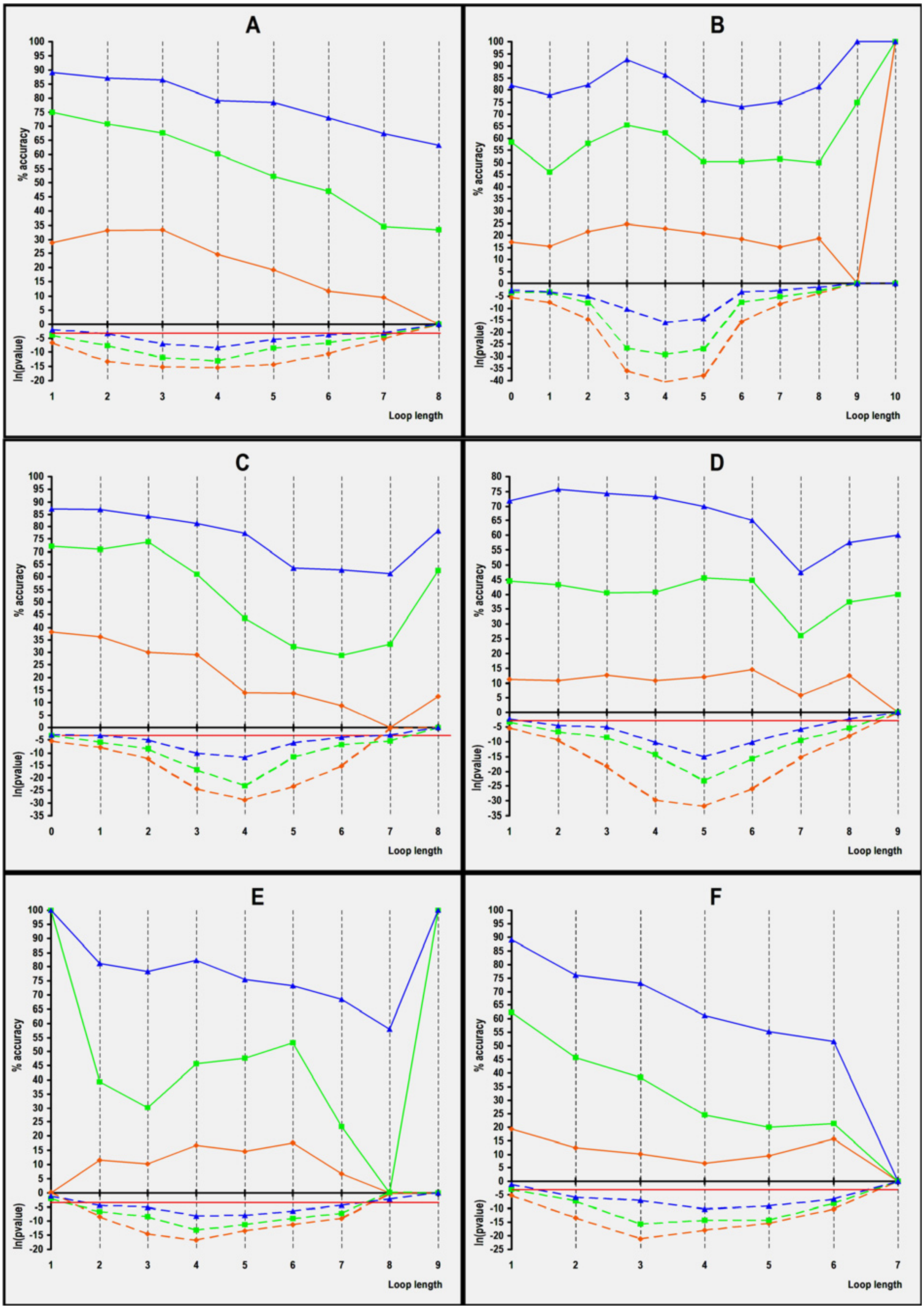
There is a general tendency in which the accuracy decreases with loop length (see figures 4.4A, 4.4B, 4.4C, 4.4D, 4.4E and 4.4F). This is general for loop structure prediction methods, either *ab initio* methods or database search methods. The increase in length brings an increase on the number of possible loop conformations and the larger variability of the structural classes.

In general, the *Sign test* shows a minimum p-value for loops with length of about 4 residues length. This reason is because the major number of classes is in this region (see figure 4.2a). The values of accuracy shows different behavior with respect to the loop length depending in the type of loop and in the number of hits (1, 5 or 20) used for prediction (see figures 4.4A, 4.4B, 4.4C, 4.4D, 4.4E and 4.4F).

We have obtained values of accuracies between 80-90% when the 20 first hits were taken into account. The accuracy decreased to a range between 60% and 70% when only the top five hits where considered and to a range between 10% and 40% for the first hit. The classes of the αβ motifs were predicted with higher significance.

The advantage of using the twenty first hits instead of using the first hit or top five hits is a larger range of correct prediction. Nevertheless, the significance is worse, especially for loops with either short or long length. The p_{value} was around above 0.05 for lengths larger than 7 residues or the shortest length of 0, 1 or 2 residues (see figures 4.4A, 4.4B, 4.4C, 4.4D, 4.4E and 4.4F).

Figure 4.4A to 4.4F. Loops class prediction. Variation of the accuracy of the prediction and ln(p_{value}) vs loop length and split according to the type (A for $\alpha\alpha$ motifs, B for $\alpha\beta$ motifs, C for $\beta\alpha$ motifs, D for $\beta\beta$ motifs, E for $\beta\beta_{hairpins}$ motifs and F for $\beta\beta_{links}$ motifs). Accuracy of prediction for first hit (solid square (■) and continuous line) and for ln(p_{value}) (solid square (■) and broken line). Accuracy of prediction for the top five hits (solid triangle (▲) and continuous line) and for ln(p_{value}) (solid triangle (▲) and broken line). Accuracy of prediction for the top twenty hits (solid diamond (♦) and continuous line) and for ln(p_{value}) (solid diamont (♦) and broken line). $\beta\beta$ motifs include both $\beta\beta_{hairpins}$ and $\beta\beta_{links}$. The values of accuracy have been obtained with FH profiles. The values of accuracy have been obtained with FH profiles and the values of significance have been obtained comparing each motif prediction against 100 independent randomly selected loop subclass on the searching space (see material and methods). The red line marks the limit of significance (p = 0.05).



Figures 3.4A to 3.4F (legend opposite)

4.4.3 Sub-class prediction

Sub-class structural prediction is more informative because it gives information about the conformation of the loop plus the relative position of the secondary structures of the motif, this being a characteristic feature of the 3D packing useful on fold prediction and *ab initio* folding.

The accuracy values for sub-class prediction are not as good as for class prediction but the values of significance were improved (see figures 4.5A, 4.5B, 4.5C, 4.5D, 4.5E and 4.5F). Structural sub-class prediction is more difficult than structural class prediction because of the larger number of clusters (see figure 4.2b) and further reduction of sequences aligned on the profiles. Nevertheless, the larger number of possible loop clusters also implies that a random prediction will decrease the specificity and this produces the improvement of the p_{values} .

Figure 4.5A to 4.5F. Loops subclass prediction. Variation of the accuracy of the prediction and $\ln(p_{value})$ vs loop length and split according to the type (A for $\alpha\alpha$ motifs, B for $\alpha\beta$ motifs, C for $\beta\alpha$ motifs, D for $\beta\beta$ motifs, E for $\beta\beta_{hairpins}$ motifs and F for $\beta\beta_{links}$ motifs). Accuracy of prediction for first hit (solid square (■) and continuous line) and for $\ln(p_{value})$ (solid square (■) and broken line). Accuracy of prediction for the top five hits (solid triangle (▲) and continuous line) and for $\ln(p_{value})$ (solid triangle (▲) and broken line). Accuracy of prediction for the top twenty hits (solid diamond (♦) and continuous line) and for $\ln(p_{value})$ (solid diamont (♦) and broken line). $\beta\beta$ motifs include both $\beta\beta_{hairpins}$ and $\beta\beta_{links}$. The values of accuracy have been obtained with FH profiles. The values of accuracy have been obtained with FH profiles and the values of significance have been obtained comparing each motif prediction against 100 independent randomly selected loop subclass on the searching space (see material and methods). The red line marks the limit of significance (p = 0.05).

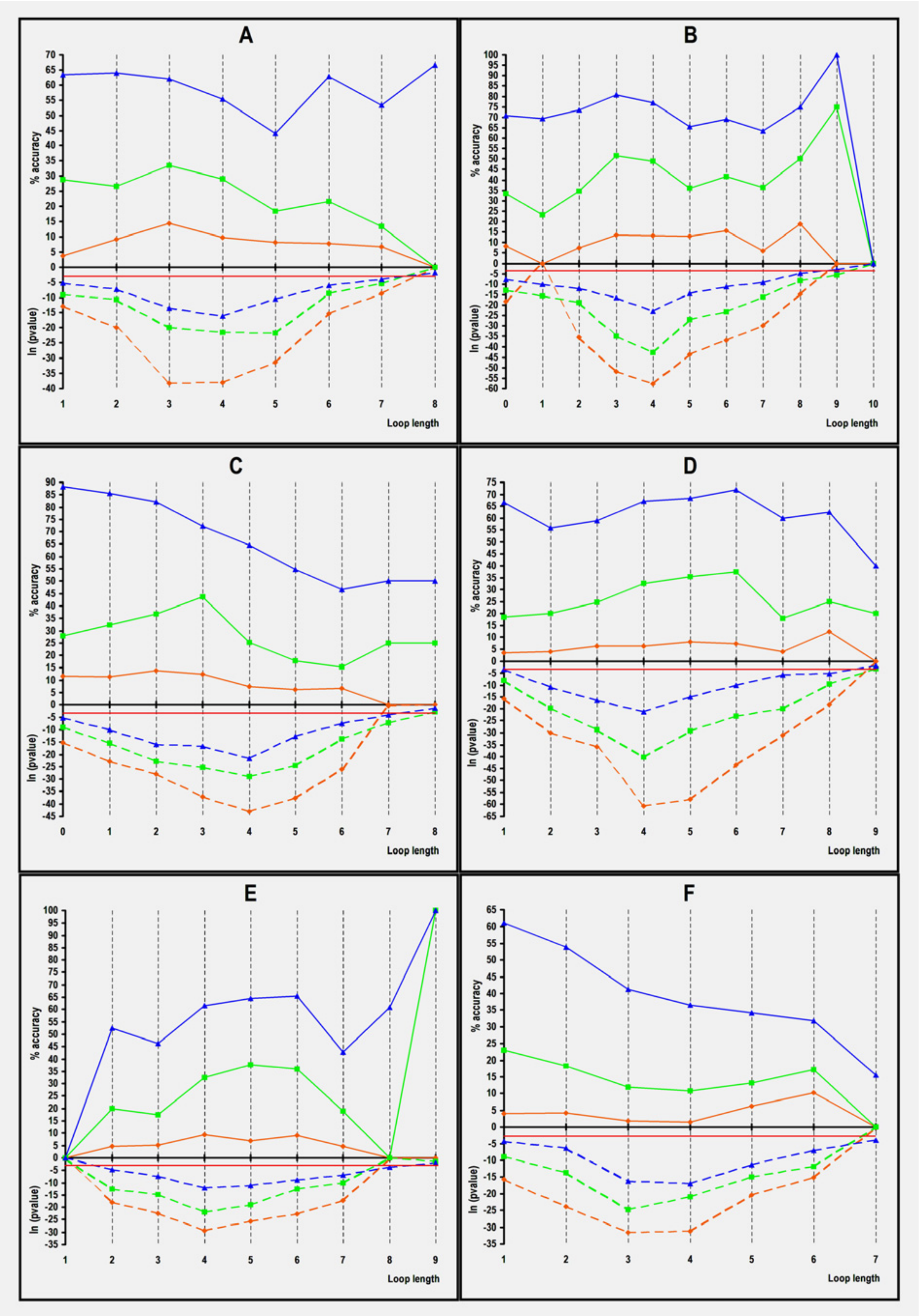


Figure 3.5A to 3.5F (legend opposite)

4.4.4 Discrimination between $\beta\beta_{links}$ and $\beta\beta_{hairpins}$

An important result derived from our study was the discrimination between $\beta\beta_{links}$ and $\beta\beta_{hairpins}$ on the prediction of loops braced by two β strands. A loop braced by two beta strands can be either $\beta\beta_{link}$ or a $\beta\beta_{hairpin}$. Topologically, the $\beta\beta_{hairpin}$ implies the formation of main-chain hydrogen bonds between both beta strands and usually a tight turn in the backbone, while strands of a $\beta\beta_{link}$ motif do not have a network of main-chain hydrogen bonds. An association of $\beta\beta_{hairpins}$ forms a beta sheet called β-meander, while a combination of $\beta\beta_{links}$ and $\beta\beta_{hairpins}$ produces different types of Greek key motifs.

The correct discrimination between $\beta\beta_{hairpins}$ and $\beta\beta_{links}$ can be very useful to predict different types of these complex super-secondary structures. There are two basic approaches attempting to predict $\beta\beta_{hairpins}$, by means of learning algorithms as neural networks(Cruz et al. 2002) and by means of statistical analyses (see reviews (Chou 2000; Kaur and Raghava 2002)). The present work relies on the statistical basis of sequence profiles for the distinction between $\beta\beta_{link}$ and $\beta\beta_{hairpins}$.

Table 4.2 shows the prediction of $\beta\beta_{hairpins}$ or $\beta\beta_{links}$ assignations among the top one, top five and top twenty scores within the profiles. The results exposed on table 2 reveals that 87.5% of loops are $\beta\beta_{hairpins}$ type prediction for first hit, 89.5% considering the top five hits and 82.2% for the top twenty, using FH profiles. With this, we contemplate the possibility of reconstructing a complex super-secondary structure formed by several continuous β-strands by combination of $\beta\beta_{links}$ and $\beta\beta_{hairpins}$

Table4.2 Percentages of $\beta\beta_{links}$ or $\beta\beta_{hairpins}$ type on the first hit, the top five hits and top twenty hits.

	Predicting $\beta\beta_{links}$			Predicting $\beta\beta_{hairpins}$		
Profile	% of a $\beta\beta_{link}$ loop type with highest score	% of a $\beta\beta_{link}$ loop type among top five scores	% of a $\beta\beta_{link}$ loop type among top twenty scores	% of a $\beta\beta_{hairpin}$ loop type with highest score	% of a $\beta\beta_{hairpin}$ loop type among top five scores	% of a $\beta\beta_{hairpin}$ loop type among top twenty scores
FP	61,2	60,2	59,5	83,1	75,2	65,8
P	64,7	63,9	63,1	87,2	80,1	73,1
FH	71,7	79,4	69,5	87,5	89,4	82,2
H	68,5	69,8	64,8	80,6	85,6	75,7

The prediction of loop type was made for $\beta\beta_{hairpins}$ and $\beta\beta_{link}$ over all ββ profiles.

4.4.5 Prediction of supersecondary structures

Table 4.3 shows the results of consecutive loops correctly predicted using the top five hits and with high significance ($p_{value} < 0.05$). By the prediction of loop class we are capable to find 231 pairs of consecutive motifs and 26 triads of consecutive loops. In subclass prediction, we predicted 79 pair of consecutive motifs and 7 triads.

If we increase the rank up to 20 top hits we were able to find a larger set of consecutive loops in class prediction (subclass prediction between parenthesis): 420 (377) with 2 loops, 90 (61) with 3 loops, 9 (8) with 4 loops, 2 (0) with 5 loops and 1 (1) with 6 loops.

Table 4.3 Types of predicted supersecondary structures.

	Profile	**Number of contiguous motifs predicted with Zscores between rank 1 and 5 and high significance ($p_{value} < 0.05$)**			
		2		**3**	
Class prediction	**FP**	**231**	80 βαβ 66 αβα 29 ααα 16 βββ 15 ααβ 9 αββ 9 βαα 7 ββα	**26**	12 αβαβ 8 βαβα 2 αβαα 2 αααα 1 ββαβ 1 ααββ
Sub-class prediction	**FP**	**79**	31 βαβ 23 αβα 13 ααα 5 βββ 3 αββ 2 ααβ 1 ββα 1 βαα	**7**	3 αβαβ 3 βαβα 1 αααα

Supersecondary structures with consecutive motifs that loop class or sub-class have been predicted among the top five hits were considered.

Interestingly, the prediction is better for alternate β-strands and α-helices (see table 4.3), 2 super motifs; 80 of βαβ and 66 of αβα motifs (12 and 8 using the sub-class prediction respectively).The rest of combinations of motifs (i.e. βββ, αββ, ββα, ...) are less represented. This being a result of the classical βαβ and αβα super-secondary structures in proteins of class α/β.

As an example for class prediction of three adjacent motifs, the region between residue 104 and 158 of structure 1i2o chain A was predicted by: (i) a βα motif (104-132) in top hit, (ii) a βα motif (128-147) in the third rank and (iii) a αβ motif (136-158) also in the top hit.

On the prediction of supersecondary structures formed by consecutive β-hairpins or intercalated β-hairpins and β-links we have found that 13 out 16 predicted motifs are two consecutive $\beta\beta_{\text{hairpins}}$, forming a β-meander, 2 out 16 are two consecutive $\beta\beta_{\text{links}}$ and 1 out of 16 is a $\beta\beta_{\text{hairpin}}$ following by a $\beta\beta_{\text{link}}$, likely forming a β-greek key. For instance, a β-meander was predicted in the region between residue 2 and 13 of structure 1a7f chain A by: (i) a $\beta\beta_{\text{hairpin}}$ motif (2-13) in top hit and (ii) a $\beta\beta_{\text{hairpin}}$ motif (8-23) in top hit. A supersecondary structure of $\beta\beta_{\text{hairpins}}$ followed by a $\beta\beta_{\text{link}}$ was correctly predicted in the region between residues 236 and 254 of structure 1qhd chain A by: (i) a $\beta\beta_{\text{hairpin}}$ motif (236-249) in the third rank and (ii) a $\beta\beta_{\text{link}}$ motif (245-254) in top hit.

Additionally, in the case of a β-meander we do not need to know the structural class or subclass of the $\beta\beta_{\text{harpins}}$. In the database used for prediction, ArchDB, we have a total of 78 continuous pairs of $\beta\beta_{\text{hairpins}}$, 24 triplets and we not have any further combination. We accurately predicted 54 β-meanders formed by 3 strands(2 consecutive $\beta\beta_{\text{hairpins}}$) and 8 β-meanders of 4 strands (3 consecutive $\beta\beta_{\text{hairpins}}$) by means of distinguishing between $\beta\beta_{\text{links}}$ and $\beta\beta_{\text{hairpins}}$.

4.4.6 Examples of predicted classes and sub-classes of loops

There are 58 motifs in the test that belong to the αβ-4.1 class, 30 out of them belong to sub-class αβ4.1.1, 16 to sub-class αβ4.1.2 and 12 to sub-class αβ4.1.3. A correct prediction of the class was obtained with the top five hits for 75% of these motifs. For sub-class prediction among the five first hits, 67% motifs were correctly predicted for αβ4.1.1 subclass, 62.5% motifs for αβ4.1.2 subclass and 60% motifs for αβ4.1.3 sub-class. The consensus sequence of this class is XXXPpXXh (where X: any residue; p: for polar residue; h: for hydrophobic residue; and one code letter for conserved residue). The common structural feature of this class is a Pro in position L2 (see figure 4.6, motifs belonging to sub-class αβ-4.1.1 are shown). The examination of the structure of the motif indicates that the conserved Pro bends the capping of the helix. In addition, the conserved hydrogen bond between the carboxylic oxygen of the residue in position L1 and the nitrogen atom of the residue in position L4 may contribute to maintain the structure of the loop.

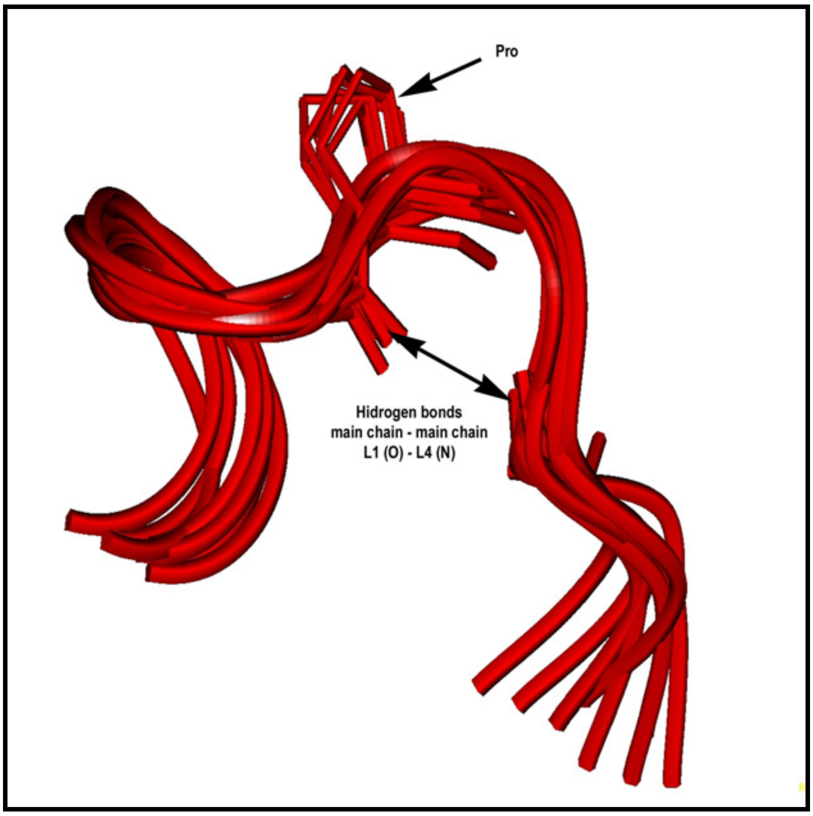


Figure 4.6 Coiled coil representation of αβ4.1.1 subclass. Conserved Pro and conserved hydrogen bonds are represented in the picture. Residue positions within clusters are denoted by the secondary structure: B for β-sheet, A for α–helix, and L for coil; and counting backwards from the C-terminal residue of the first secondary structure and forwards from the N-terminal residue of the loop. Figures 4.6 to 4.8 were generated using PREPI v.0.9 (SA Islam and MJE Sternberg; available upon request at: http://www.sbi.bio.ic.ac.uk/prepi/index.html).

On the queries of $\beta\alpha$ type of motifs there are 38 motifs of the class $\beta\alpha$-3.1, 27 out of them belong to sub-class $\beta\alpha$-3.1.1 and 11 to sub-class $\beta\alpha$-3.1.2. We have predicted 48% of motifs with the first hit and 92% with the five top hits. The prediction of sub-class was correct for 40% of motifs of subclass $\beta\alpha$-3.1.1 with the first hit and for 90% of motifs when considering the five top hits. For sub-class $\beta\alpha$-3.1.2 the assignation of the correct sub-class on the top hit was correct for 55% of motifs, and for 91% of motifs when considering the five top hits. The consensus sequence of this class is hhGXGXh and the analysis of the structure of the motifs illustrates that the two gly residues in L1 and L3 could allow a severe turn in the chain without steric impediments. A conserved hydrogen bond between the carboxylic oxygen of the residue in position L3 and the nitrogen in position A3 probably stabilize the N terminus capping of the α-helix (see figure 4.7; motifs belonging to sub-class $\beta\alpha$-3.1.1 are shown).

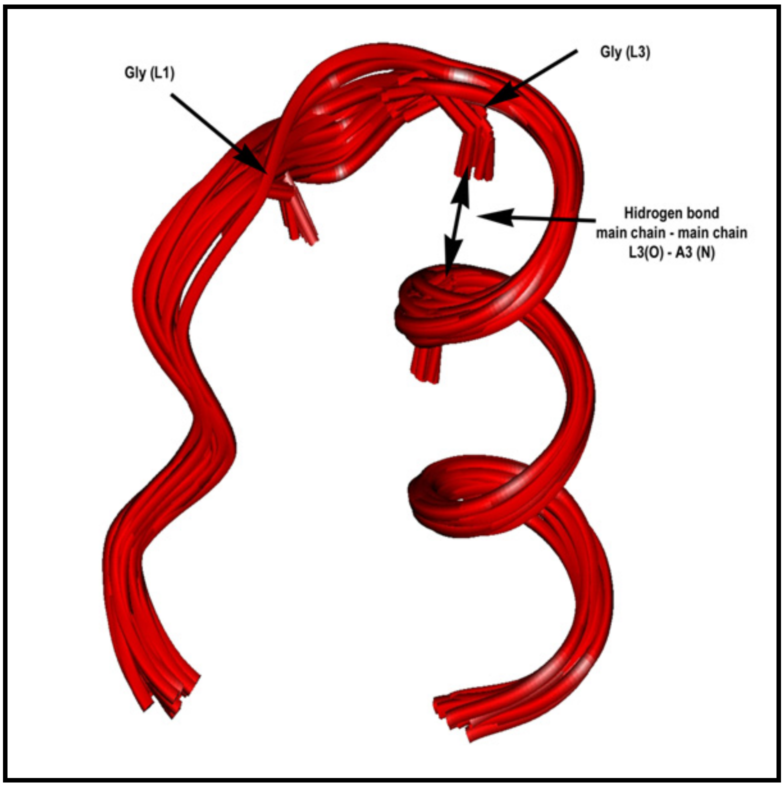


Figure 4.7. Trace plate of the sub-class $\beta\alpha$3.1.1 subclass. Conserved gly residues and conserved hydrogen bonds are shown.

Finally, on the ββ queries we have predicted 68 motifs of the class ββhairpin-5.1. Taking into account the five top hits, the subclass prediction was correct for 72% motifs of subclass $\beta\beta_{hairpin}$5.1.1, 69% motifs for $\beta\beta_{hairpin}$5.1.4 subclass, 59% motif of subclass $\beta\beta_{hairpin}$5.1.5, 82% motifs for $\beta\beta_{hairpin}$5.1.6 subclass, 75% motifs of subclass$\beta\beta_{hairpin}$5.1.7, 66% motifs of subclass $\beta\beta_{hairpin}$5.1.11, 50% motif of subclass $\beta\beta_{hairpin}$5.1.12 and 66% motifs of subclass $\beta\beta_{hairpin}$5.1.15. The consensus sequence for these motifs is XXXppGpXX. A gly in position L4 is well conserved and the loop conformation is maintained by the main-chain hydrogen bond between -B1 (carboxylic oxygen) and L4 (nitrogen) that hold the structure of the loop (see figure 4.8, motifs of sub-class $\beta\beta_{hairpin}$5.1.1 are shown). A careful study of the structure of the motifs explains that the conservation of the gly is due to avoid side-chain hindrances.

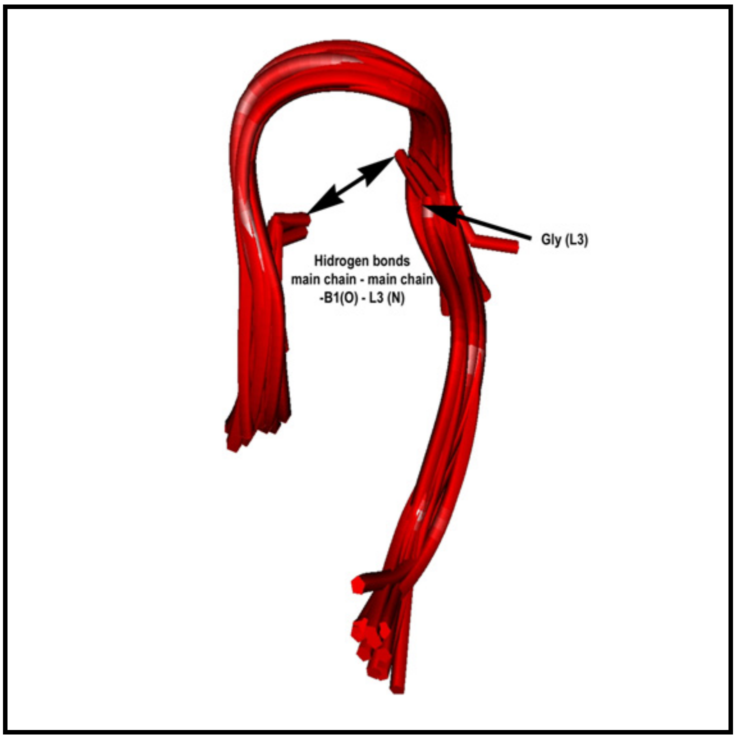


Figure 4.8. Coiled coil representation of the $\beta\beta_{hairpin}$5.1.1 subclass. Conserved gly residues and conserved hydrogen bonds are described.

4.5 Discussion and Conclusions.

4.5.1 Applications

We have shown by an extensive analysis that it is possible to obtain good prediction results using our loop structure database, ArchDB. Values of accuracy and significances of predictions justify the usefulness of our database in loop structure prediction, both for loop class prediction and sub-class prediction.

The loop structure prediction can be applied on: (i) comparative modeling, (ii) fold recognition and (iii) *ab initio* prediction. The relative high values of accuracy in class prediction validate the use of loop structure prediction in homology modeling of proteins where errors in loops are the dominant problem. In ranges of identity larger than 35%, loops from homolog proteins vary while in the core regions are still conserved and accurately aligned(Fiser et al. 2002). On the other hand, the sub-class prediction adds a topological value (as it is the orientation between secondary structures) worth for *fold recognition*.

Since the prediction is made using only sequence information, knowledge of the local structural environment of the loop can be used to identify incorrect predictions. In addition, geometry definitions of loop sub-classes can be considered as a validation for loops built on to homology models, discarding predictions based on the fitting to a structure or suggesting changes to the orientation of the bounding secondary structures.

Furthermore, for each query loop a rank of possible conformations is given. This set can be combined in a combinatorial fashion to construct several candidates in an *ab initio* fold building.

The success obtained in the absence of an exact prediction of the boundaries of the secondary structure suggests that the alignment between the motifs and the profiles might help in the refinement of secondary structure prediction.

Further positive aspects of the presented method are the short computing times compared with *ab initio* approaches and the fact that the database grows in a natural manner with the PDB, which will lead to better prediction accuracies continuously.

4.5.2 Discrimination between $\beta\beta_{\text{hairpins}}$ and $\beta\beta_{\text{links}}$

A secondary structure prediction of a beta-beta motif does not give information about the structural arrangement of two consecutive β-strands, this information is important in fold recognition because of its structural implications. As discussed before, the structural arrangement of a $\beta\beta_{\text{hairpin}}$ is different from a $\beta\beta_{\text{link}}$ and indicates topological characteristics to distinguish one fold and the 3D contact map. We have correctly predicted 54 β-sheets forming a β-meander by continuous $\beta\beta_{\text{hairpins}}$ that would help on fold prediction. The discrimination between $\beta\beta_{\text{hairpins}}$ and $\beta\beta_{\text{links}}$ also helps on folding *ab initio* as it may describe β-sheet cores

4.5.3 Comparison with other previously published works

The obtained results compare very favorably with other protein loop prediction methods. The closer related work to test the performance of a loop database in loop prediction was done by Burke & Deane(Burke and Deane 2001) using Sloop database(Burke et al. 2000). They got an accuracy of 58%, 78% and 85% considering the highest score, the three highest scores and the five highest scores respectively. The percentages of accuracy are higher that those obtained in our work. We can argue different reasons to explain the differences in the accuracy values. First, the size of the database. Burke and Deane used a database of 560 classes (equivalent to sub-classes in our database format) while our database has 1492 sub-classes. Second, for cross-validation they used a 7-fold cross validation splitting randomly the loops in seven sets of ~1700 loops each set where homology may exist between the searching space and the test set. In our case, the jack-knife test was made by removing all possible homology between the query and its searching space.

To measure whether or not the homology could affect the accuracy of the prediction, the prediction was made on the searching space with homologs. The results are shown in table 1 indicate that accuracy of prediction is higher when only the query loop is removed than when the whole super-family of the query is removed. In class prediction, the gain in prediction is around 20 %, considering the first hit and in subclass prediction, the gain was around 15-20% when considering the first hit.

4.6 References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "*Gapped BLAST and PSI-BLAST a new generation of protein database search programs.*" Nucleic Acids Res. **25**: 3389-3402.

Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain and e. al. (2002). "*The Protein Data Bank.*" Acta Crystallogr. D. Biol. Crystallogr. **58**: 899-907.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). "*The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.*" Nucleic Acids Res **31**(1): 365-70.

Bonneau, R., J. Tsai, J. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss and D. Baker (2001). "*Rosetta in CASP4: progress in ab initio protein structure prediction.*" Proteins **Suppl. 5**: 119-126.

Bruccoleri, R. E. and M. Karplus (1987). "*Prediction of the folding of short polypeptide segments by uniform conformational sampling.*" Biopolymers **26**: 137-168.

Burke, D., C. Deane and T. Blundell (2000). "*Browsing the Sloop database of structurally classified loops connecting elements of protein secondary structure.*" Bioinformatics **16**: 513-516.

Burke, D. F. and C. M. Deane (2001). "*Improved protein loop prediction from sequence alone.*" Protein Eng **14**(7): 473-8.

Chandonia, J. M., N. S. Walker, L. L. Conte, P. Koehl, M. Levitt and S. E. Brenner (2002). "*ASTRAL compendium enhancements.*" Nucl Acid Res **30**: 260-263.

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

Chothia, C., A. M. Lesk, M. Levitt, A. G. Amit, R. A. Mariuzza, S. E. Phillips and R. J. Poljak (1986). "*The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure.*" Science **233**: 755-758.

Chou, K. C. (2000). "*Prediction of tight turns and their types in proteins.*" Analyticla Biochem. **286**: 1-16.

Conte, L. L., S. E. Brenner, T. J. P. Hubbard, C. Chothia and A. Murzin (2002). "*SCOP database in 2002: refinements accommodate structural genomics.*" Nucl Acid Res **30**: 264-267.

Cruz, X. d. l., E. G. Hutchinson, A. Shepherd and J.M.Thornton (2002). "*Toward predicting protein topology: An approach to identifying beta hairpins*." PNAS **00**(17): 11157-11162.

Deane, C. M. and T. L. Blundell (2001). "*CODA: a combined algorithm for predicting the structurally variable regions of protein models*." Protein Sci **10**(3): 599-612.

Domingues, F. S., P. Lackner, A. Andreeva and M. J. Sippl (2000). "*Structure-based evaluation of sequence comparison and fold recognition alignment accuracy.*" J. Mol. Biol. **297**: 1003-1013.

Donate, L. E., S. D. Rufino, L. H. Canard and T. L. Blundell (1996). "*Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction*." Protein Sci **5**(12): 2600-16.

Du, P., M. Andrec and R. M. Levy (2003). "*Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update*." Protein Eng **16**(6): 407-14.

Eddy, S. (1998). "*Profile Hidden Markov Models.*" Bioinformatics **14**: 755-763.

Espadaler, J., N. Fernandez-Fuentes, A. Hermoso, E. Querol, F. X. Aviles, M. J. Sternberg and B. Oliva (2004). "*ArchDB: Automated protein loop classification as a tool for Structural Genomics.*" Nucleic Acids Res **32**: D185-D188.

Fetrow, J. S. (1995). "*Omega loops: nonregular secondary structure significant in protein function and stability.*" FASEB **9**: 708-717.

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

Fine, R. M., H. Wang, P. S. Shekin, D. L. Yarmush and C. Levinthal (1986). "*Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MPCPC603 from many randomly generated loop conformations.*" Proteins **1**: 342-362.

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

Fiser, A., M. Feig, C. L. Brooks, 3rd and A. Sali (2002). "*Evolution and physics in comparative protein structure modeling*." Acc Chem Res **35**(6): 413-21.

Jones, D. T. (1999). "*Protein secondary structure prediction based on position-specific scoring matrices.*" J Mol Biol **292**: 195-202.

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

Kabsch, W. and C. Sander (1983). "*Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*." Biopolymers **22**(12): 2577-637.

Kaur, H. and G. P. S. Raghava (2002). "*An evaluation of beta turn prediction methods*." Bioinformatics **18**(11): 1508-1514.

Kelley, L., R. MacCallum and M. Sternberg (2000). "*Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM.*" J. Mol. Biol. **299**: 499-520.

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

Levitt, M. (1992). "*Accurate modelling of protein conformation by automatic segment matching.*" J Mol Biol **226**: 507-533.

Lupas, A. N., C. P. Ponting and R. B. Russell (2001). "*On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?*" J Struct Biol **134**(2-3): 191-203.

Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali (2000). "*Comparative protein structure modeling of genes and genomes.*" Annu. Rev. Biophys. Biomol. Struct. **29**: 291-325.

Moult, J. and M. N. James (1986). "*An algorithm for determining the conformation of polypeptide segments in proteins by systematic search.*" Proteins **1**: 146-163.

Oliva, B., P. A. Bates, E. Querol, F. X. Aviles and M. J. Sternberg (1997). "*An automated classification of the structure of protein loops*." J Mol Biol **266**(4): 814-30.

Oliva, B., P. A. Bates, E. Querol, F. X. Aviles and M. J. Sternberg (1998). "*Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction*." J Mol Biol **279**(5): 1193-210.

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

Rufino, S., L. Donate, L. Canard and T. Blundell (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.

Salem, G. M., E. G. Hutchinson, C. A. Orengo and J. M. Thornton (1999). "*Correlation of observed fold frequency with the occurrence of local structural motifs*." J Mol Biol **287**(5): 969-81.

Sibanda, B. L., T. L. Blundell and J. M. Thornton (1989). "*Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering*." J Mol Biol **206**(4): 759-77.

Simons, K. T., C. Kooperberg, E. Huang and D. Baker (1997). "*Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions*." J Mol Biol **268**(1): 209-25.

Tosatto, S. C., E. Bindewald, J. Hesser and R. Manner (2002). "*A divide and conquer approach to fast loop modeling*." Protein Eng **15**(4): 279-86.

Venclovas, C., A. Zemla, K. Fidelis and J. Moult (2001). "*Comparison of Performance in Successive CASP Experiments*." PROTEINS: Structure, Function and Genetics **Suppl 5**: 163-170.

Wlijmen, H. W. V. and M. Karplus (1997). "*PDB-based protein loop prediction: parameters for selection and methods for optimization.*" J Mol Biol **267**: 975-1001.

Wood, T. and W. Pearson (1999). "*Evolution of protein sequences and structures.*" J Mol Biol **291**: 977-995.

Yang, A. S. and L. Y. Wang (2002). "*Local structure-based sequence profile database for local and global protein structure predictions*." Bioinformatics **18**(12): 1650-7.