Chem-Bio Informatics Journal, Vol. 2, No. 1, pp.60-72(2002)

FCANAL: Structure based protein function prediction method. Application to enzymes and binding proteins

Ayumi Suzuki[1], Tadashi Ando[1], Ichiro Yamato[1*] and Satoru Miyazaki[2]

[1]*Department of Biological Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba 278-8510, Japan*
[2]*Department of Pharmaceutical Sciences, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba 278-8510, Japan*

[*]*E-mail: iyamato@rs.noda.tus.ac.jp*

(Received ; accepted ; published online)

Abstract

Structural genomics projects are beginning to produce protein structures with unknown functions, therefore, high-throughput methods for predicting functions are necessary. Although sequence based function prediction methods have been applied extensively, structure based prediction is believed to provide higher specificity and sensitivity because functions are closely related to three-dimensional structures of functional sites which are more strongly conserved during evolution than sequence.

We have developed FCANAL, a method to predict functions using the score matrix obtained from the distances between C^{α} atoms and frequencies of appearance [1]. The previous report used key residues predicted from sequence comparison (motifs). In this report, we expanded the method to enzymes and binding proteins with key residues predicted on the basis of three-dimensional structures. Using FCANAL, we constructed score matrices for 31 enzymes. When we applied them to all the structure entries deposited at the Protein Data Bank, FCANAL could detect functional sites with high accuracy. This suggests that the FCANAL will help identify the functions of newly determined structures and pinpoint their functionally important regions.

Key Words: protein function prediction, enzyme, binding protein, protein local structures, amino acid propensity, distance distribution, bioinformatics

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Thanks to the large-scale genome sequencing projects, tons of primary sequences have been determined and the number of uncharacterized sequences will continue to grow. One of the big problems of genome science is to analyze what meaning the accumulated huge genome information has for living activity.

Function prediction methods based mainly on sequence homology are generally used for genome annotation, however, is far from complete because about 40% of genome sequences have not homologous counterparts in the known sequence data base [2].

It is well known that protein structures are more strongly conserved during evolution than sequences [3]. Thus, protein function can often be predicted from its fold [4]. But the fact that two proteins have similar 3D folds does not simply imply that they have similar function. For example, an eight-stranded parallel β/α barrel structure has been detected over 60 different enzyme commission (EC) numbers [5], all of which are called as TIM barrel proteins. Thus more powerful function prediction methods which are not based on overall 3D fold similarity are needed.

In general, 3D local structures located at functional sites are the useful information to detect active sites. Currently, there are reported many 3D local structure comparison approaches based on graph-theoretic algorithm [6] or geometric hashing algorithm [7]. Because these approaches need to define 3D structural templates and fit them in all possible orientational combinations with a large amount of structural data [8], the speed of structural comparison process is usually very slow. Structural genomics projects are producing a large amount of 3D structure data of proteins, therefore, we attempted to develop a different automated approach that can be used for fast and accurate identification of active sites of proteins.

In the previous report, we have developed a method which extracts 3D structural characteristics of functions, the FCANAL (Fast Calculable protein function ANALyzer) [1]. It used protein motif sequences for identifying the key residues for extracting 3D features, which limited the number of predictable proteins. Therefore, we expanded the FCANAL to the present system which identified functionally important two to four residues in the 3D functional sites as key residues instead of using motif sequences. We chose 31 enzymes as targets to evaluate the new FCANAL, and showed that the FCANAL provided an efficient and accurate approach to discriminate these enzyme active sites from other local structures. The FCANAL accelerates the execution of function prediction because there is no necessity to fit the templates vectorially in all possible orientational combinations with a large number of structures, and predicts precisely without any apparent sequence homology.

2. Materials and methods

2.1 Data sets of protein structures

We used three sets of protein structure data obtained from the PDB. A training data set and a test data set were generated which consisted of enzyme structures with the same EC number from the Enzyme Structure Database [9]. The training data set was such that redundancy within any family was minimized by using the program CD-Hit [10] with a 40% sequence identity cutoff and that it included functionally important residues. The test data set with 90% identity cutoff of CD-Hit was used. A background data set was that in a representative list of 1771 PDB chain identifiers from the

PDB_SELECT [11] of October 2002 with a sequence identity of <25%.

2.2 Definition of functional important residues

Functionally important residues were identified as two to four residues by refering to the PROCAT [12][13] or by searches on literatures. The PROCAT database provides a library of functional site geometries that can be used as templates. The residues contained in the templates were adopted as functional residues if they were found in the training data set and verified by literatures. Figure 1 shows a close-up of a functional site of carbonate dehydratase. Glu-Thr were determined as the consensus template in the PROCAT and play an important role in the interconversion between H_2O and HCO_3^- [14]. We adopted the Glu-Thr as functional residues in the carbonate dehydratase accordingly. On the other hand, the PROCAT provides Glu-Asp template for the beta-amylases, whereas the active site of the beta-amylases has been reported to comprise residues Glu-Glu [15]. Therefore, we adopted Glu-Glu as functional residues for the beta-amylases. The functional residues used in this study are listed in Table 1.

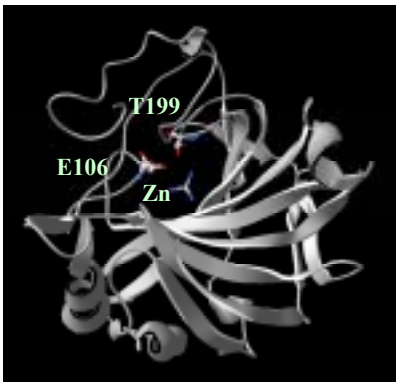

Figure 1. The functional site of the carbonate dehydratase is characterized by Thr199 and Glu 106.
The Thr 199 - Glu 106 pair aids deprotonation of the Zn^{2+} - bound water molecule. A hydrogen-bond network involving the pair plays an important role in the catalytic function of the enzyme.

Table 1. Functional residues in target proteins.

Target	E. C. number	Consensus template *	Functional residues	Literature number
Phosphoribosylglycinamide formyltransferase	2. 1. 2. 2	H-D-N	**H**-D-N	
Triacylglycerol lipase	3. 1. 1. 3	S-H-D	S-**H**-[D, E]	(16)
Phospholipase A2	3. 1. 1. 4	Ca-H-D	**H**-D	
Acetylcholinesterase	3. 1. 1. 7	S-H-E	S-**H**-E	
Ribonuclease T1	3. 1. 27. 3	E-H-H	E-**H**-[H, R]	(17)
Pancreatic ribonuclease	3. 1. 27. 5	H-H-K	H-**H**-K	
Alpha-amylase	3. 2. 1. 1	E-D	D-**E**-D	(18)
Beta-amylase	3. 2. 1. 2	E-D	**E**-E	(15)
Glucan 1,4-alpha-glucosidase	3. 2. 1. 3	E-E	**E**-E	
Cellulase	3. 2. 1. 4	D-D	**E**-[D, E]	(19)
Endo-1,4-beta-xylanase	3. 2. 1. 8	E-E	**E**-E	
Lysosyme	3. 2. 1. 17	E-D	**E**-D	
Exo-alpha-sialidase	3. 2. 1. 18	E-D	**E**-D	
Beta-galactosidase	3. 2. 1. 23	E-E	**E**-E	
Glucan endo-1,3-beta-D-glucosidase	3. 2. 1. 39	E-E	**E**-E	

6-phospho-beta-galactosidase	3. 2. 1. 85	E-E	**E**-E	
Mannosyl-glycoprotein endo-beta-N-acetylglucosamidase	3. 2. 1. 96	E-E	**E**-D	(20)
Carboxypeptidase C	3. 4. 16. 5	S-H-D	S-**H**-D	
Carboxypeptidase A	3. 4. 17. 1	Zn-E-R	**E**-R	
Serine protease	3. 4. 21. -	S-H-D	S-**H**-D	
Cystein endopeptidase	3. 4. 22. -	C-H-N	C-**H**-N	
Aspartic endopeptidase	3. 4. 23. -	T-G-D-D-T-G	**D**-D	(21)
Astacin	3. 4. 24. 21	Zn-E-Y	**H**-H-H-Y	(22)
Matrilysin	3. 4. 24. 23	Zn-E-A	**E**-A	
Thermolysin	3. 4. 24. 27	Zn-E-H-D	E-**H**-D	
Adamalysin	3. 4. 24. 46	Zn-E-G	**E**-G	
Haloalkane dehalogenase	3. 8. 1. 5	D-H-D	D-**H**-[D, E]	(23)
Carbonate dehydratase	4. 2. 1. 1	E-T	**E**-T	
Mandelate racemase	5. 1. 2. 2	Mg-K-K-E-H	K-K-**E**-H	
Triosephosphate isomerase	5. 3. 1. 1	H-E-K	**H**-E-K	
Glutamine-tRNA ligase	6. 1. 1. 18	H-N-K	**H**-N-K	

*Consensus template was provided by the PROCAT.

The gray shaded regions indicate that the functional amino acid residues are different from the consensus templates.

Amino acid residues in boldface are key residues used in this study.

2.3 The FCANAL algorithm

We made a score matrix for each functional site (Figure 2) mostly according to the previous report [1]. The abstract of the flow chart of the FCANAL with slight modification is as follows:

(1) Selection of a key amino acid residue.

The key residue is such that is conserved in functional sites and especially plays an important role for the function as shown in Table 1. The previous method identified the key residue from motif sequences, thereby we could make score matrices for only target proteins which contained motif sequences. In the present method, we referred to literatures, identified the most important residue in the functional residues, and defined the residue as the key residue. For example, we adopted a His as the key residue at the active site of serine proteases (Table 1) because the His plays the most important role for the enzyme activity via proton transfer reaction.

(2) Define a local structure.

We took a sphere to be considered as the local structure with each key residue as its center, and analysed the structural information on residue arrangement in the local structure. The radius of the sphere was changed from 5 Å to 15 Å by 1 Å, and a score matrix was constructed for every radius. The radius that provided the highest *F* value and *A* value was used for function prediction. The methods of construction of a score matrix, *F* value and *A* value will be described in detail in the following sections.

(3) Extract structural information of functional sites as $P^{site}(aa, C)$ and other local structures as $P^{bg}(aa, C)$.

We counted the number of an amino acid residue *aa* within class mark *C* which is the distance mark between the center of a local structure and the cut off radius of a sphere divided by 0.5 Å interval as previously reported [1]. For example, when cut off radius is 11 Å, the class marks are 0.25, 0.75, 1.25, …, 10.25, 10.75 starting from the center, and the total mark number is 22. We estimated $P^{site}(aa, C)$, the probability of an amino acid residue *aa* locating within the class mark *C* in functional local structure in the training data set : For example, P^{site}(Gly, 1.25) represents the probability of Gly locating within the distance between 1.0 Å and 1.5 Å from the key residue in

functional sites. $P^{bg}(aa, C)$ is the probability of the amino acid residue *aa* locating within the class mark *C* in all the local structures in the background data set : $P^{bg}(aa, C)$ represents the background probability of the amino acid residue *aa* at the distance class *C*.

(4) Define a score matrix for the respective amino acids and class marks.

We defined a score matrix S as follows:

$$\mathrm{S} = S(aa, C),$$

where *aa* represents one of 20 amino acids and *C* represents a class mark. The number of score matrix elements $S(aa, C)$ changes according to different cut off radius.

We estimated $S(aa, C)$ as follows:

$$S(aa,C) = \ln\left(\frac{Q(aa,C)}{P^{bg}(aa,C)}\right).$$

$Q(aa, C)$ is a reevaluated probability of amino acid *aa* at class mark *C* at functional sites: Since the number of protein structures in the training data was much smaller than in background data, we used a pseudo-count method for estimating probability at functional sites to correct the apparent error when $P^{site}(aa, C)$ becomes 0 by reevaluating $P^{site}(aa, C)$ by using the score of $P^{bg}(aa, C)$ as below:

$$Q(aa,C) = \frac{\alpha P^{site}(aa,C) + \beta\left(P^{bg}(aa,C) \times P^{site}(aa) \Big/ P^{bg}(aa)\right)}{\alpha + \beta},$$

where α represents the weight of observed value and β represents the weight of the pseudo-count part, and we assumed α = 1 and β = 1 in our method.

(5) Calculate a total score of a local structure.

Score matrix *S* represents conformational features of functional sites by probabilities of any residues at any class marks, $S(aa, C)$. Based on the score matrix S, we defined a continuous function *DB*(*distance based*) score for each amino acid, $DB_{aa}(x)$, in which *x* is the continuous distance [1], by interpolating the $S(aa, C)$ value at distance *C* and that at the distance of the next class mark as follows:

$$DB_{aa}(x) = f(S(aa,C)).$$

A total score *TS* of a local structure of a target protein is calculated as follows:

$$TS = \sum_{i=1}^{Naa} DB_{aa_i}(x).$$

where the summation is performed for all the amino acids N_{aa}, the total number of amino acids within the local structure.

(6) Perform refinement of the score matrix to improve the accuracy of prediction.

The performance of S was evaluated by its *F* and *A* values. *F* value is the sum of the sensitivity measure (P_r) and the specificity measure (P_q) as reported previously [1]. *A* value is a new criterion and is the difference between the mean of TS_{site} of functional local structures and the mean of TS_{bg} of the other local structures:

$$A = \overline{TS}_{site} - \overline{TS}_{bg}.$$

The higher the *A* value, the more separately the total scores *TS* can discriminate the local structure of functional sites from the other sites.

We performed two steps of refinements [1]; the first step was refinement of the combination of structures in the training data set. S's were calculated changing the combination of structures in

the training data set. The combination which gave the highest *F* and *A* values was used for the training data set. The second step was refinement of the combination of elements *S*(*aa*, *C*) of the score matrix S using the training data set which obtained from the first step. The combination of *S*(*aa*, *C*) which gave the highest *F* and *A* values was used for the function prediction.

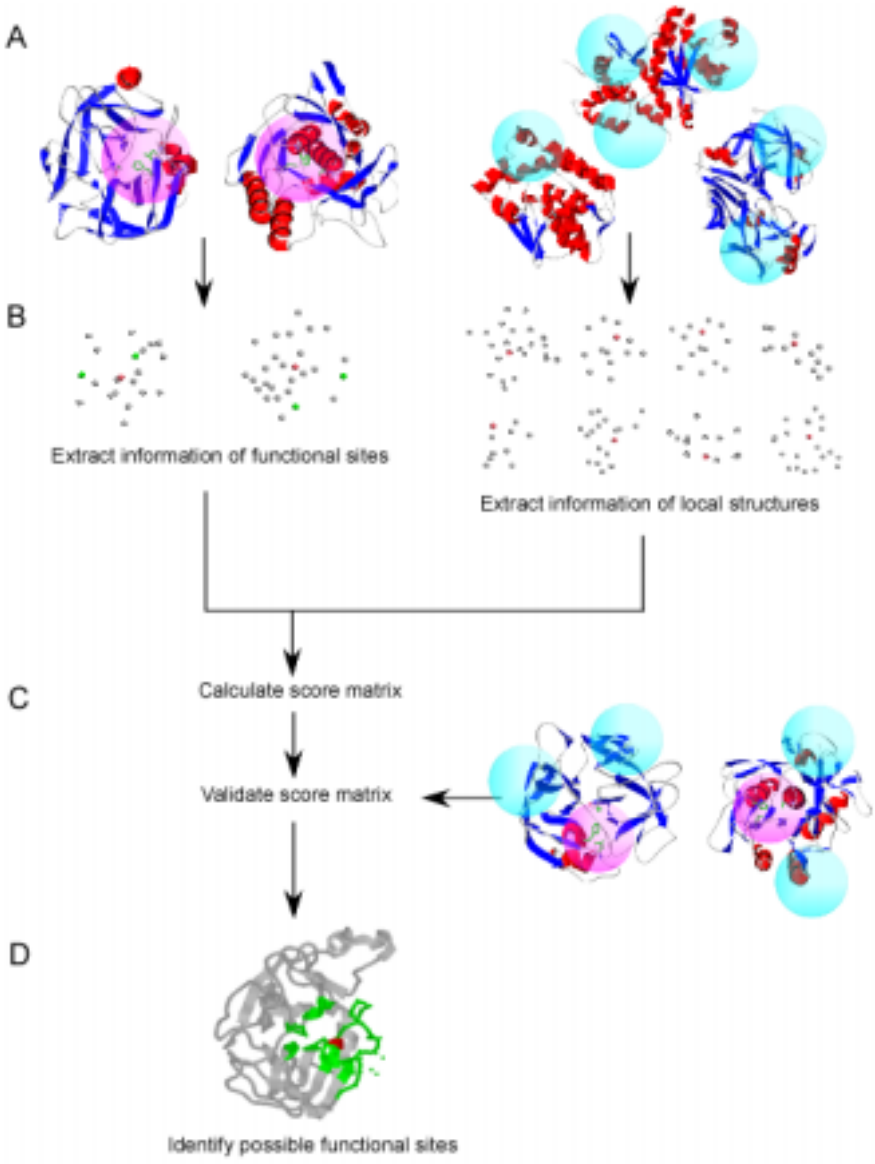

Figure 2. Scheme of our FCANAL method.

(A) Define the local structure as a sphere with each key residue as its center. The local structures containing functional key residues (green stick models) are shown as pink spheres. Sky blue spheres are the other local structures. **(B)** Extract information of local structures. Distributions of C^{α}'s in local structures with the key residue (pink balls) contained in important functional residues (green balls) in the training data and in the background data are shown. **(C)** The score matrix is calculated by using a ratio of the two probabilities as described in the text. The score matrix is validated using the test data. **(D)** Possible functional sites are identified as regions of a protein by their high scores, which are displayed by green.

2.4 Evaluation of prediction accuracy

To evaluate the performance of the function prediction method, the receiver operator characteristic (ROC) measure has been widely used [24][25]. We calculated the number of true-positive local structures (TP_{PDB}), false-positive local structures (FP_{PDB}), false-negative local structures (FN_{PDB}) and true-negative local structures (TN_{PDB}) in all the PDB entries in September 21, 2004 using different threshold values of the total score TS. The ROC curve was drawn by plotting 1-specificity ($1 - \frac{TN_{PDB}}{TN_{PDB} + FP_{PDB}}$) against sensitivity ($\frac{TP_{PDB}}{TP_{PDB} + FN_{PDB}}$). The best prediction method should yield a curve that shows two straight lines making a sharp crossing point at the upper left area of the ROC space, whereas a random method should give a straight diagonal line from bottom left to top right.

We also defined the accuracy value which provides an indication of specificity and sensitivity:

$$accuracy = \frac{TP_{PDB}}{TP_{PDB} + FP_{PDB}}.$$

By changing the accuracy value in the prediction by FCANAL, we can obtain results of various precision levels.

In addition, we estimated similarities among the features of score matrices by using a multidimensional scaling (MDS) method. In the MDS configuration, the distance between two points indicates the degree of being different between the features of two score matrices.

2.5 Online prediction of functional sites of proteins

Online prediction based on the FCANAL is available at http://atgc002.ps.noda.tus.ac.jp/contents/fcanal/. The FCANAL interface allows a user to choose a structure to scan by either entering the 4 letter PDB ID of a protein structure if it is available from PDB or uploading a structure from the local computer, and to select an accuracy value. Results which are displayed by an intuitive and interactive interface via the Chime plug-in are provided.

3. Results and discussion

Performance of the FCANAL

P_r, P_q, F and A values for 31 functional sites are shown in Table 2. P_r and P_q are sensitivity and specificity measures as defined in the previous report [1]. All the F values were 2.00 and almost all the A values were >20, which indicates that the score matrices quantified the possibilities of being functional sites precisely.

We take carbonate dehydratase as the example to show the high performance of the FCANAL in searching all the structure entries deposited at the PDB. The prediction results are shown in the histogram (Figure 3). The TS of local structures distributed as two completely separate peaks (Figure 3(B)). The black bars which represent the functional sites of carbonate dehydratase present in the dataset had TS of >14. Most of other local structures ranged from -23 to 13. Thus, the score matrix could distinguish clearly the functional sites from other local structures.

The ROC curve in Figure 4 shows the high sensitivity and specificity of prediction by our score matrix. Score matrices for other functional sites showed similar ROC curves.

Table 2. Results of the refinement of the score matrix for prediction of functional sites.

Target	Cut off (Å)	P_r	P_q	F	A
Phosphoribosylglycinamide formyltransferase	11	1.00	1.00	2.00	31.70
Triacylglycerol lipase	10	1.00	1.00	2.00	13.42
Phospholipase A2	10	1.00	1.00	2.00	38.36
Acetylcholinesterase	10	1.00	1.00	2.00	29.29
Ribonuclease T1	10	1.00	1.00	2.00	8.02
Pancreatic ribonuclease	10	1.00	1.00	2.00	33.81
Alpha-amylase	10	1.00	1.00	2.00	18.50
Beta-amylase	11	1.00	1.00	2.00	30.392
Glucan 1,4-alpha-glucosidase	11	1.00	1.00	2.00	34.25
Cellulase	10	1.00	1.00	2.00	18.46
Endo-1,4-beta-xylanase	10	1.00	1.00	2.00	28.80
Lysosyme	10	1.00	1.00	2.00	8.32
Exo-alpha-sialidase	10	1.00	1.00	2.00	19.92
Beta-galactosidase	9	1.00	1.00	2.00	30.71
Glucan endo-1,3-beta-D-glucosidase	11	1.00	1.00	2.00	33.98
6-phospho-beta-galactosidase	10	1.00	1.00	2.00	35.09
Mannosyl-glycoprotein endo-beta-N-acetylglucosamidase	12	1.00	1.00	2.00	36.03
Carboxypeptidase C	9	1.00	1.00	2.00	30.31
Carboxypeptidase A	10	1.00	1.00	2.00	32.63
Serine protease	10	1.00	1.00	2.00	16.42
Cystein endopeptidase	10	1.00	1.00	2.00	11.01
Aspartic endopeptidase	12	1.00	1.00	2.00	23.58
Astacin	10	1.00	1.00	2.00	37.42
Matrilysin	10	1.00	1.00	2.00	37.67
Thermolysin	11	1.00	1.00	2.00	40.36
Adamalysin	10	1.00	1.00	2.00	32.32
Haloalkane dehalogenase	12	1.00	1.00	2.00	41.38
Carbonate dehydratase	10	1.00	1.00	2.00	27.09
Mandelate racemase	10	1.00	1.00	2.00	32.31
Triosephosphate isomerase	10	1.00	1.00	2.00	26.33
Glutamine-tRNA ligase	10	1.00	1.00	2.00	33.52

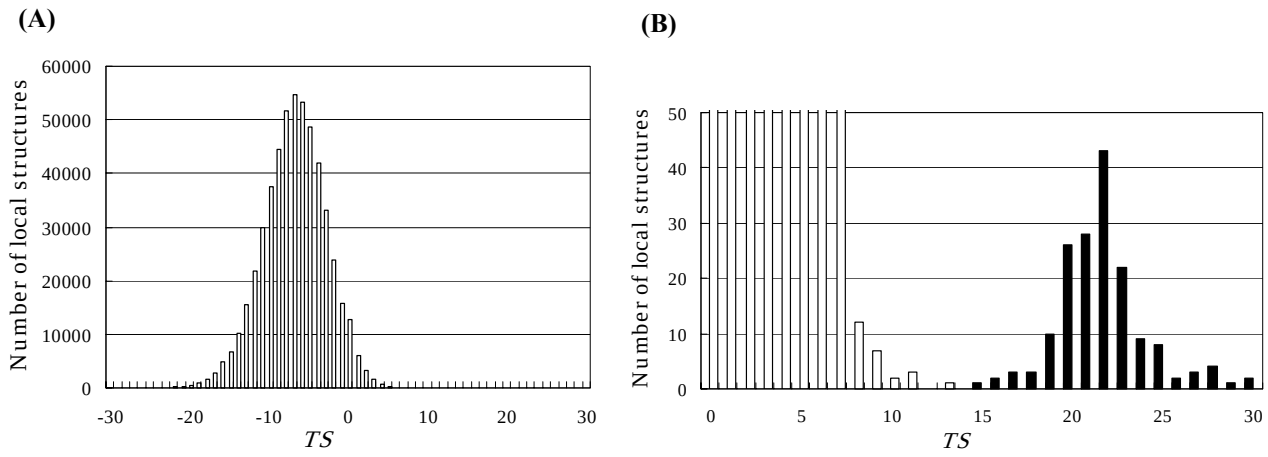

Figure 3. Score distribution for carbonate dehydratase of local structures in all the PDB entries.

(A) The score distribution for all the local structures. **(B)** The score distribution for the local structures with scores more than 0. The bars show the number of the local structures containing functional sites (black bars) and the number of the other local structures (open bars) within each relative score range.

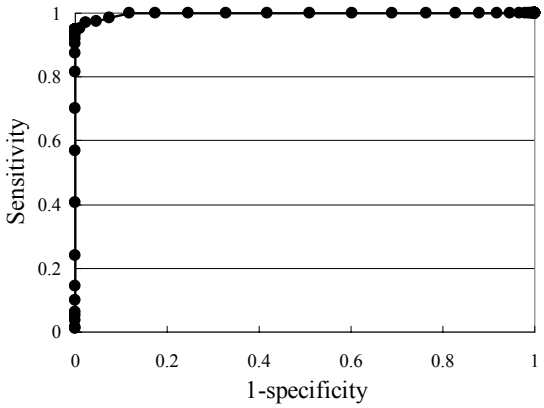

Figure 4. ROC plot showing the performance of FCANAL in case of the score matrix for carbonate dehydratase.

Figure 5 shows the performance of our FCANAL depending on the choice of the accuracy value. The higher the accuracy value, the higher the specificity as judged from the increase of TN_{PDB} and the decrease of FP_{PDB}. The lower the accuracy value, the higher the sensitivity as judged from the increase of TP_{PDB} and the decrease of FN_{PDB}. We recommend that the choice of the accuracy value is between 0.5 and 0.8. When we choose the accuracy value 1.0, the predicted functional sites should be all true positive. Higher accuracy value (~0.8) would be better for target proteins with a large number of 3D structures determined having strictly conserved 3D functional sites such as carbonate dehydratase. On the other hand, lower accuracy value (~0.5) would be recommended for target proteins with only a few structures determined.

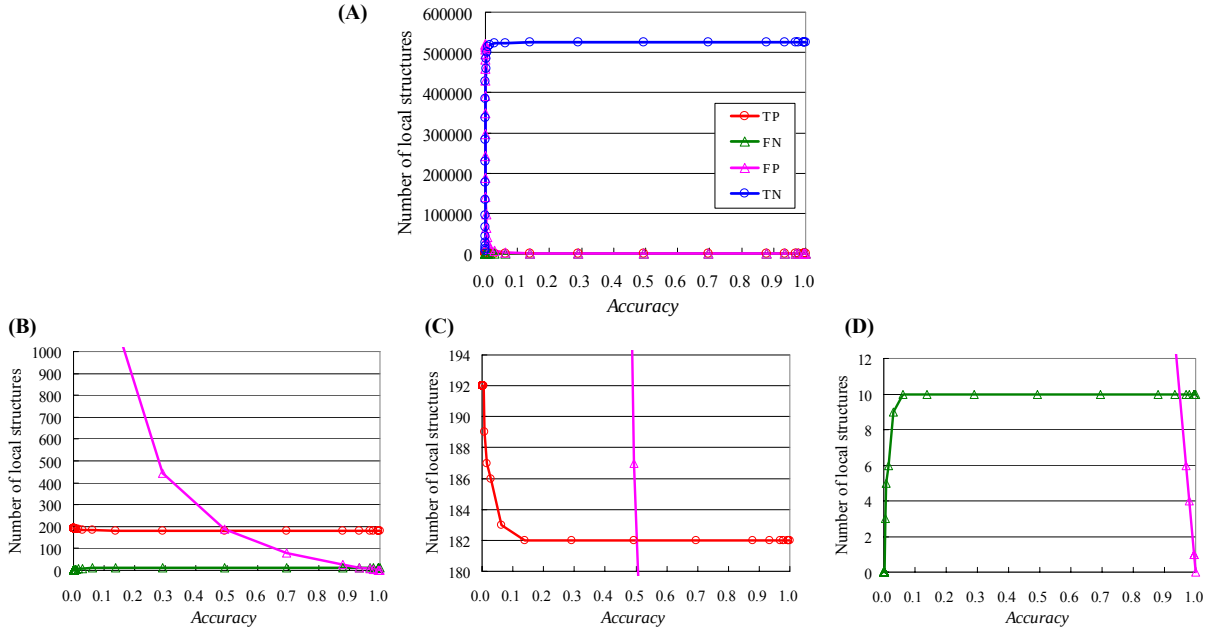

Figure 5. Efficiency of FCANAL discriminating TP_{PDB} (⊖), FN_{PDB} (△), FP_{PDB} (△) and TN_{PDB} (⊖) depending on the choice of accuracy value.
The ordinates are drawn in different scales from **(A)** to **(D)**.

The features of functional sites predicted by the score matrix

The score matrices can quantify the kinds of residues present at any distances from the key residues in functional sites. Figure 6 shows a part of score matrix for the triacylglycerol lipase. There are some notable features in the score matrix. First, hydrophobic or positively charged residues (e.g. Met, Lys and Arg) tend to have negative $S(aa, C)$ at almost all the class marks. This suggests that a large number of these hydrophobic residues are in buried regions, and positively charged residues are rarely found in the functional sites. Second, the functional residues Ser and Asp or Glu are present at approximately 8.0 Å and 4.5 Å from the key residue, which form the actual functional structure. Third, the tendency of $S(aa, C)$ of Cys and His at various class marks suggests that these residues specifically influence the conformation of functional sites. These results indicate that the score matrix captures well the conformational features of functional sites.

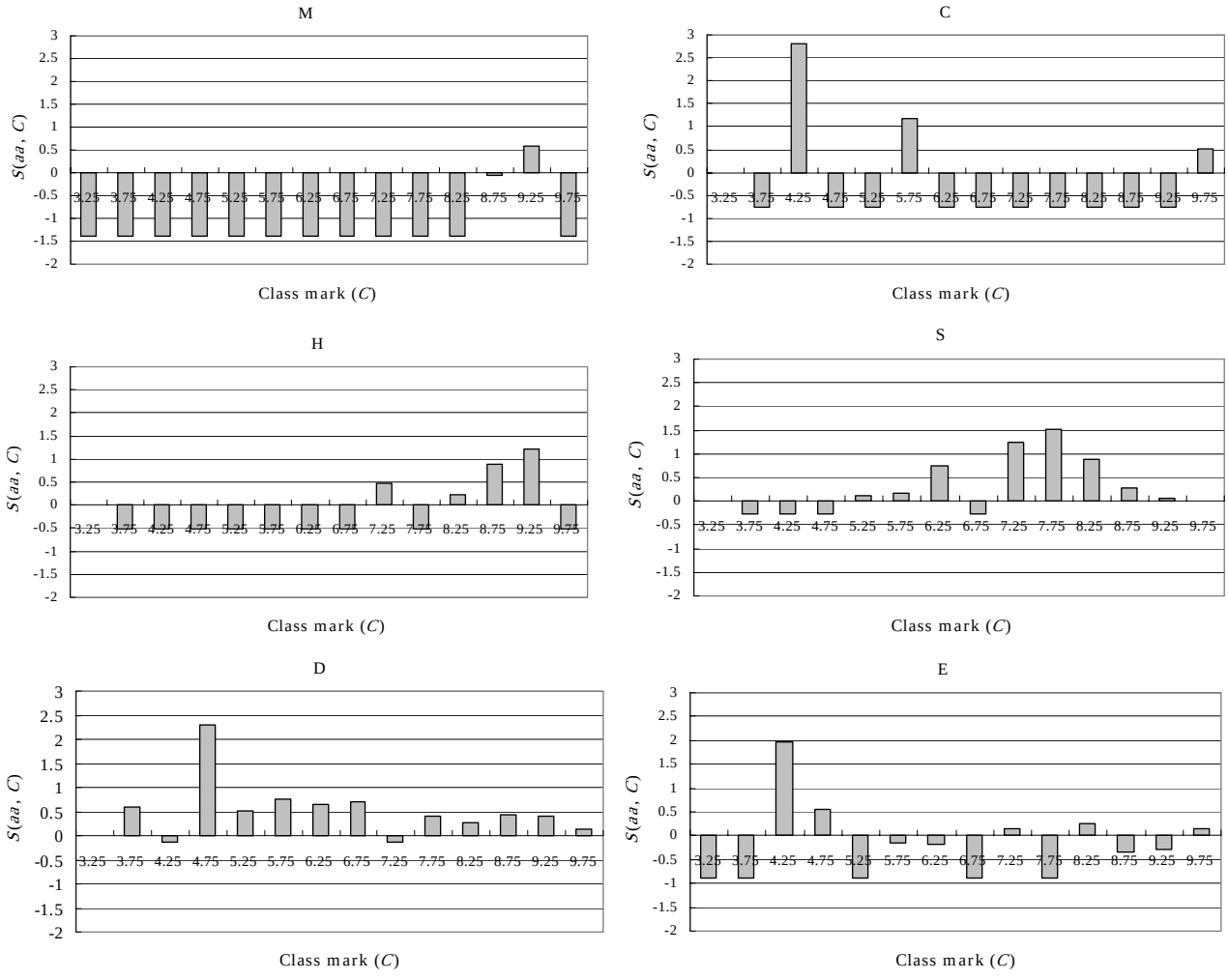

Figure 6. Histogram of $S(aa, C)$ for typical residues of triacylglycerol lipase at various class marks (C).

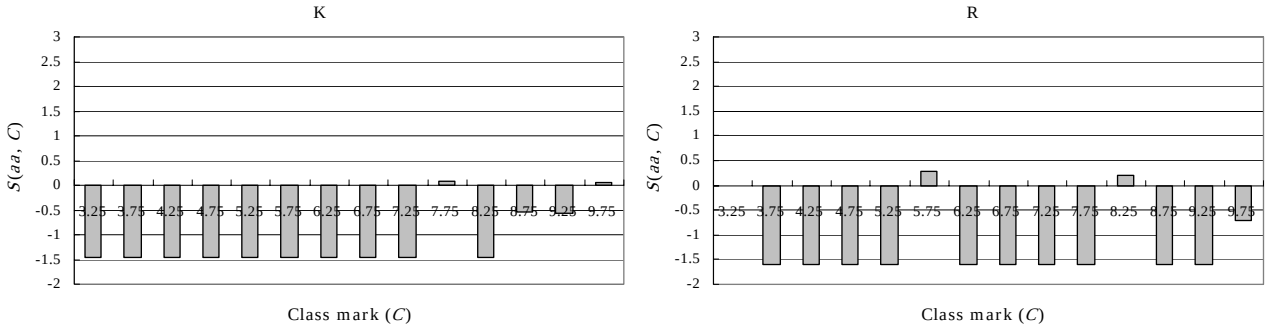

Figure 6. (Continued)

The similarities of score matrices of five enzymes classified under EC 3.1 (triacylglycerol lipase, phospholipase A2, acetylcholinesterase, ribonuclease T1 and pancreatic ribonuclease) are interesting. The two-dimensional configuration analyzed by MDS is shown in Figure 7. We analyzed along four dimension and we showed the two dimensional data in Figure 7. The two points for both triacylglycerol lipase and acetylcholinesterase are close to each other, suggesting that these two enzymes contain the same functional residues, His, Ser and Asp with similar 3D structure and thus these residues would perform similar functions in the respective enzyme activities. Furthermore, points for two ribonucleases classified under EC 3.1.27 (ribonuclease T1 and pancreatic ribonuclease) are also close to each other. This indicates that the score matrices of similar functions tend to show similar distributions of $S(aa, C)$. In the above sections, we have shown that the FCANAL can identity different functional sites with high specificity and sensitivity. Therefore, we can conclude that the FCANAL quantifies the features of difference of local structures as well as their similarity.

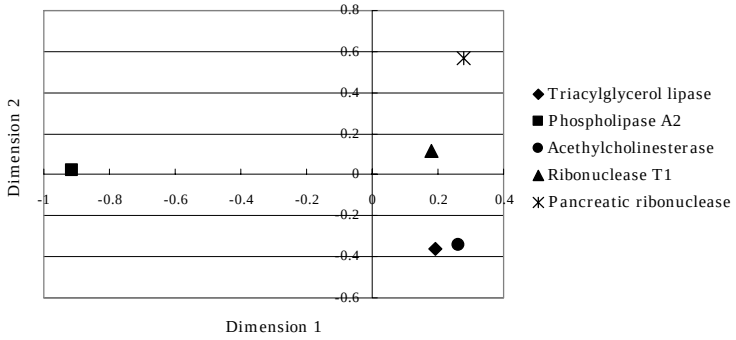

Figure 7. Similarity measures among the score matrices analyzed by MDS.

Results of structure-based function assignment

By using the score matrix of pancreatic ribonuclease, a function prediction was carried out against the non-redundant data set which was generated from the PDB on December 12, 2004, containing 2286 entries by using online search tools in the PDB-REPRDB (http://mbs.cbrc.jp/pdbreprdb-cgi//reprdb_query.pl). In this data set, the sequence similarity is less

than 30% and fragments and mutants are excluded. Three local structures which were identified to have functional sites of pancreatic ribonuclease when we used the accuracy value of 0.8, were true positives (Figure 8): Only these three structures have function of pancreatic ribonuclease in the data set. The two structures with a relatively high *TS* were angiogenin (PDB code: 1agi, *TS* = 24.3 [26]) and pancreatic ribonuclease (PDB code: 1ssa, *TS* = 23.1 [27]), and were followed by eosinophil-derived neurotoxin (PDB code: 1gqv, *TS* = 10.3 [28]). The difference of the *TS* was derived from the difference of the position of two helices, although the topological relationship of the functional residues in 1gqv is similar to 1agi and 1ssa (Figure 9). This difference may be due to the difference of species; 1agi and 1ssa are from bovine and 1gqv is from human. Thus, we can conclude that the FCANAL can predict functional sites flexibly and specifically, even in the case where there are some differences in 3D structures of functional sites due to various factors such as different species and the existence of ligands.

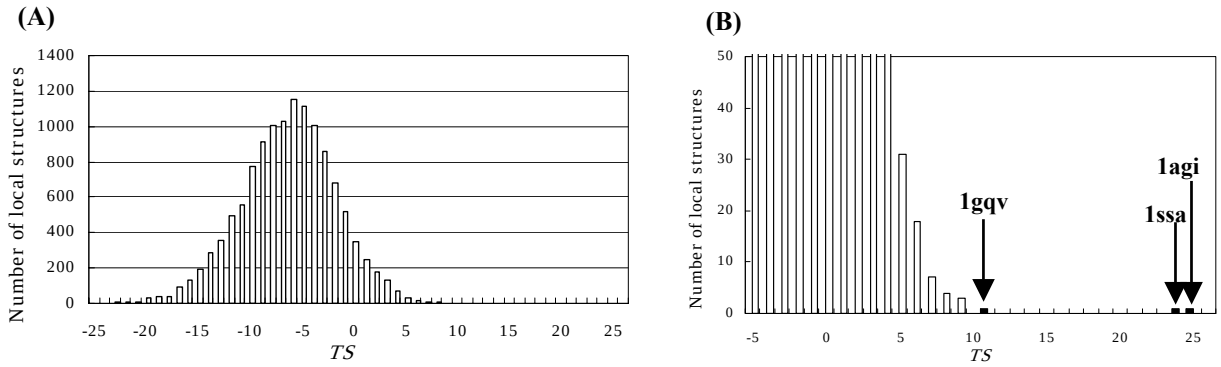

Figure 8. Search results for pancreatic ribonuclease against the non-redundant data set.
TS distributions were shown in the same way as in Figure 3. **(A)** The TS distribution for all the local structures. **(B)** The TS distribution for the local structures with TS more than -5. The three local structures were identified as functional sites.

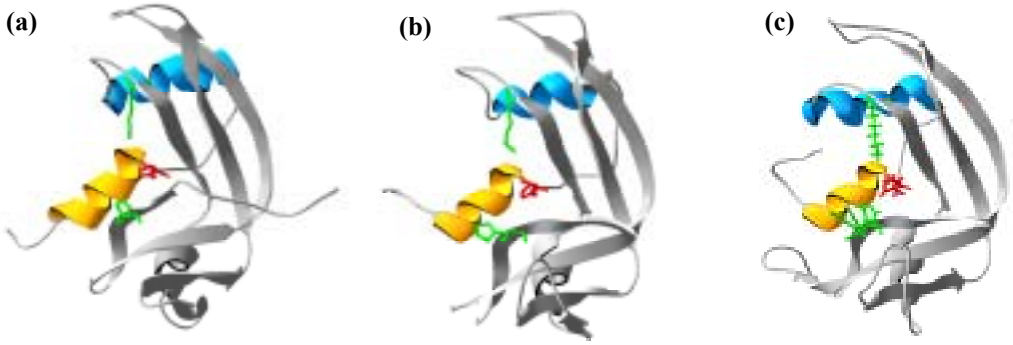

Figure 9. The predicted functional sites of pancreatic ribonuclease are shown on the target proteins, **(a)** angiogenin (PDB code 1agi), **(b)** pancreatic ribonuclease (PDB code 1ssa) and **(c)** eosinophil-derived neurotoxin (PDB code 1gqv).
Functional residues are shown in stick models, key residues His are in red, and the others, Lys and His, are in yellowish green. The predicted functional sites are in a sphere with a radius of 10 Å centered on each key residue. The positional relationship between the upper helix (shown in sky blue) and the lower helix (shown in orange) involved in the functional site of 1gqv is different from those of 1agi and 1ssa.

We applied score matrices for 31 enzymes to 114 hypothetical proteins of unknown function in the non-redundant dataset. Two local structures in ybgI (PDB code 1nmp [29]) and AQ_1354 (PDB code 1oz9 [30]) from *Escherichia coli* were identified to have predictable functional sites even when we used the accuracy value 0.8. The ybgI belongs to the NIF3 (NGG1p interacting factor 3)-like family and its biological function is not known. Experiments [31] showed that it binds metal ions at the putative active site comprised of four His, two Glu, Asp, Asn, Tyr, Cys and Trp. Furthermore, the experimental data implied that the ybgI may catalyze the hydrolysis or oxidation or involved in DNA metabolism [29]. The FCANAL assigned the local structure at a sphere with Glu 194 as its center, which is one of the Glu in the putative active site, to a functional site of beta-amylase (Figure 10 (A)). Of course, this prediction needs to be experimentally confirmed. If this is true-positive, we are proud of the FCANAL, and even if our prediction was incorrect, we think that the FCANAL can help experimental design and performance at this genome era. Expansion of the entries of predictable functional sites by the FCANAL should increase its accuracy and performance.

The AQ_1354 from *Aquifex aeolicus* belongs to the UPF0054 family and is predicted as a metal-dependent proteinases by structure-based homology analysis, however, is not substantiated by experimental evidence [30]. The amino acid sequence alignment between the AQ_1354 and three proteinases which were shown to have the highest structural similarity showed that His 115, His 119 and His 125 were conserved in the binding site of Zn ion [30]. The FCANAL assigned the local structure with a radius of 10 Å with His 103 as its center to the functional site of triosephosphate isomerase (Figure 10 (B)). By the CASTp [31] analysis, the location of this predicted site was at the second large pocket next to the Zn binding site, which suggests that the predicted site may bind any substrate or co-factor. Further experimental work is awaited to reveal the true function of this protein.

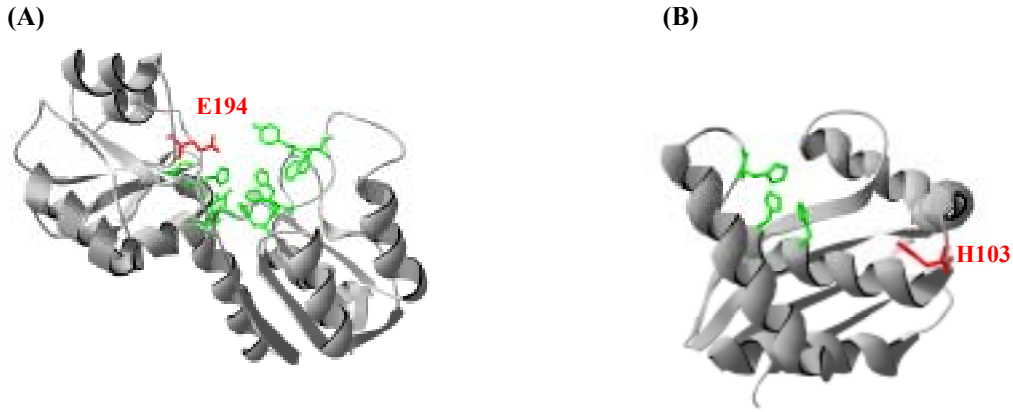

Figure 10. Function prediction of hypothetical proteins by FCANAL.
(A) YbgI protein (PDB code 1nmp). The most likely residues contained in the functional site (H63, H64, H97, H215, E194, E219, D101, N108, C171, Y22 and W68) are shown in yellowish green stick models. The predicted functional site of the alpha-amylase is at a sphere with a radius of 11 Å centered on E194 (a red stick model). **(B)** AQ_1354 (PDB code 1oz9). The residues which bind Zn ion are shown in yellowish green stick models. The predicted functional site of the triosephosphate isomerase is at a sphere with a radius of 10 Å centered on H103 (a red stick model).

4. Conclusion

Although the FCANAL uses only the C^{α}-C^{α} distances and frequency of appearance of amino acid residues in the local structures in proteins, sensitivity and specificity are high enough to predict the functions. Furthermore, our method does not require either motif sequence or sequence homology. Therefore, our method can predict functions of proteins with low fold and low sequence homology.

Our method will benefit in both accuracy and coverage in parallel with the expansion of structural databases. Accordingly, we are going to increase the number of functions predictable by our method.

References

[1] T. Asaoka, T. Ando, T. Meguro and I. Yamato, *CBIJ*, **3**, 96-113 (2003).
[2] D. Devos and A. Valencia, *Proteins: Struct. Funct. Genet.*, **41**, 98-107 (2000).
[3] C. Chothia and A. M. Lesk, *EMBO. J.*, **5**, 823-826 (1986).
[4] L. Holm and C. Sander, *Nucleic Acids Res.*, **24**, 206-209 (1996).
[5] N. Nagano, C. A. Orengo and J. M. Thornton, *J. Mol. Biol.*, **321**, 741-765 (2002).
[6] P. J. Artymiuk, A. R. Pirrette, H. M. Grindley, D. W. Rice and P. A. Willett, *J. Mol. Biol.*, **243**, 327-344 (1994).
[7] R. Nussinov and H. J. Wolfson, *Proc. Natl. Acad. Sci.* USA, **88**, 10495-10499 (1991).
[8] D. Fischer, H. Wolfson, S. L. Lin and R. Nussiv, *Protein Sci.*, **3**, 769-778 (1994).
[9] R. A. Laskowski, V. V. Chistyakov and J. M. Thornton, *Nucl. Acids Res.*, **33**, D266-D268 (2005).
[10] W. Li, L. Jaroszewski and A. Godzik, *Bioinformatics*, **18**, 77-82 (2002).
[11] U. Hobohm and C. Sander, *Protein Sci.*, **3**, 522-524 (1994).
[12] A. C. Wallace, R. A. Laskowski and J. M. Thornton, *Protein Sci.*, **5**, 1001-1013 (1996).
[13] A. C. Wallace, N. Borkakoti and J. M. Thornton, *Protein Sci.*, **6**, 2308-2323 (1997).
[14] Y. Xue, A. Liljas, B. H. Jonsson and S. Lidskog, *Proteins: Struct. Funct. Genet.*, **17**, 93-106 (1993).
[15] B. Mikami, H. J. Yoon and N. Yoshigi, *J. Mol. Biol.*, **285**, 1235-1243 (1999).
[16] D. Ghosh, Z. Wawrzak, V. Z. Pletnev, N. Li, R. Kaiser, W. Pangborn, H. Jornvall, M. Erman and W. L. Duax, *Structure*, **3**, 279-288 (1995).
[17] K. M. Polyakov, A. A. Lebedev, A. L. Okorokov, K. I. Panov, A. A. Schulga, A. G. Pavlovsky, M. Y. Karpeiskya and G. G. Dodson, *Acta Cryst.*, **58**, 744-750 (2002).
[18] A. Kadziola, J. Abe, B. Svensson and R. Haser, *J. Mol. Biol.*, **239**, 104-121 (1994).
[19] P. M. Alzari, H. Souchon and R. Dominguez, *Structure*, **4**, 265-275 (1996).
[20] C. A. Waddling, T. H. Plummer Jr., A. L. Tarentino and P. V. Roey, *Biochemistry*, **38**, 7878-7885 (2000).
[21] C. Abad-Zapatero, R. Goldman, S. W. Muchmore, C. Hutchins, K. Stewart, J. Navaza, C. D. Payne and T. L. Ray, *Protein Sci.*, **5**, 640-652 (1996).
[22] W. Bode, F. X. Gomis-Rüth, R. Huber, R. Zwilling and W. Stöcker, *Nature*, **358**, 164-167 (1992).
[23] J. Newman, T. S. Peat, R. Richard, L. Kan, P. E. Swanson, J. A. Affholter, I. H. Holmes, J. F. Schindler, C. J. Unkefer and T. C. Terwilliger, *Biochemistry*, **38**, 16105-16114 (1999).
[24] M. Gribskov and N. Robinson, *Comput. Chem.*, **20**, 25-33 (1996).
[25] J. Hou, S. R. Jun, C. Zhang and S. H. Kim, *Proc. Natl. Acad. Sci.* USA, **102**,

3651-3656(2005).
[26] K. R. Acharya, R. Shapiro, J. F. Riordan and B. L. Vallee, *Proc. Natl. Acad. Sci.* USA, **92**, 2949-2953 (1995).
[27] V. S. J. Demel, M. S. Doscher, M. A. Glinn, P. D. Martin, M. L. Ram and B. F. P. Edwards, *Protein Sci.*, **3**, 39-50 (1994).
[28] G. J. Swaminathan, D. E. Holloway, K. Veluraja and K. R. Acharya, *Biochemistry*, **41**, 3341 -3352 (2002).
[29] J. E. Ladner, G. Obmolova, A. Teplyakov, A. J. Howard, P. P. Khil, R. D. Camerini-Otero and G. Gilliland, *BMC Struct. Biol.*, **3**, 7-14 (2003).
[30] V. Oganesyan, D. Busso, J. Brandsen, S. Chen, J. Jancarik, R. Kim and S. H. Kim, *Acta Cryst.*, **59**, 1219-1233 (2003).
[31] T. A. Binkowski, S. Naghibzadeh and J. Liang, *Nucleic Acids Res.*, **33**, 3352-3355 (2003).