

A Novel Model for Prediction of RNA binding Proteins

Shingo Kikugawa¹, Hideki Takehara², Satoru Kuhara²,
and Makoto Kimura^{1*}

¹Laboratory of Biochemistry, Department of Bioscience and
Biotechnology, Faculty of Agriculture, Graduate School,
Kyushu University, Fukuoka 812-8581, Japan

²Laboratory of Molecular Gene Technics, Faculty of Agriculture,
Graduate School, Kyushu University, Fukuoka 812-8581, Japan

*E-mail: mkimura@agr.kyushu-u.ac.jp

(Received December 3, 2004; accepted January 13, 2005; published online February 4, 2005)

Abstract

We have developed an efficient prediction procedure for RNA binding proteins with oligosaccharide/oligonucleotide binding-fold (OB-fold). First, all pairwise superimpositions of 96 OB-fold structures included in Structural Classification of Proteins (SCOP) database classified them into four distinct groups on the basis of structural similarity. The proteins belonging to each group were divided into RNA binding proteins and non-RNA binding proteins. The structure-based sequence alignment of RNA binding proteins in each group were made to build profile hidden Markov models (HMMs). The reliability of HMMs thus obtained was first evaluated by the application to the PDB40 sequence dataset; RNA binding proteins with OB-fold classified into OB-fold in SCOP were selected, giving E-values less than 1.0. The next application of HMMs to sequence database of the hyperthermophilic archaeon *Pyrococcus horikoshii* OT3 detected several RNA binding proteins, including tRNA synthetases, initiation factor, transcriptional regulatory proteins, and ribosomal protein L10E as RNA binding proteins with OB-fold. These results suggested that HMM derived from this study has information about RNA binding proteins with OB-fold. The present analysis strongly suggested 4 hypothetical proteins in *P. horikoshii* to be RNA binding proteins with OB-fold. Furthermore, the application of the present model to the rice full-length cDNA sequence database suggested 14 hypothetical proteins to be RNA binding proteins with OB-fold. It is known that some of the motifs have no specific biological function alone but are part of larger structural and functional assemblies. Thus, the present method would provide clues as to protein functions of unannotated proteins and also be useful for a target selection for structural genomics.

Key Words: hidden Markov models, OB-fold, protein function prediction, *Pyrococcus horikoshii*, RNA binding protein

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

The genome sequencing efforts have now provided biologists with coding information for thousands of new proteins, most of which have no known function that can be predicted using sequence-based methods [1]. Hence, assigning functions to novel proteins is one of the most important problems in the postgenomic era. Recent advances in structural analysis, together with a known fact that three-dimensional structures are conserved across a much greater evolutionary distance than recognizable primary sequences, have led to the concept of “structural genomics”, the determination of three-dimensional protein structures on a genome-wide scale [2]. An important use of three-dimensional protein structural information of proteins is to uncover clues as to protein functions that are not detectable from sequence analysis [3][4]. Although experimental structure determination methods are providing high-resolution structure information about a subset of the proteins [5], computational methods for prediction of protein function are required for large fraction of sequences whose structures will not be determined experimentally.

Recently, several groups have developed algorithms to predict protein function employing sequence-based comparisons. A simple and widely used strategy is the identification of a high sequence similarity between proteins of known and unknown function that is then used to transfer the specific function [6][7][8]. However, lower levels of sequence similarity can only be used to transfer general functions and therefore, this approach is not reliable. On the other hand, non-sequence-based approaches for the protein functional prediction, including the analysis of gene expression patterns, phylogenetic profiles, protein fusions, and protein-protein interactions have been reported. Clustering analysis of gene expression data can be used to predict functions of unannotated proteins based on the idea that coexpressed genes are more likely to have similar functions [9][10]. In addition, functional predictions have been modeled as pattern recognition problems based on sequence homologies and structural information as well as phenotype data [11][12][13]. Although they have been extensively improved, an alternative computational method should be required for high-throughput prediction of the protein functions. As for the protein structural prediction, there have been a number of promising advances in de novo structure prediction [14][15]. A particularly successful method, called Rosetta, uses information from the Protein Data Bank (PDB) to estimate possible conformation for local sequence segments [16][17]. Rosetta was recently used to generate both fold and function predictions for Pfam protein families [18] and *Halobacterium* sp. *NRC-1* [19] and provide general function information for many proteins of unknown function. In spite of these successes, the accuracy of de novo methods is still known to be problematic.

There is growing evidence that RNA molecules play essential roles in biological processes of living cells, such as pre-mRNA splicing in the spliceosome [20] and peptide-bond formation in the ribosome [21]. RNA molecules usually perform these functions in close association with RNA binding proteins, and thus RNA binding proteins play crucial roles in a wide range of biological processes. To date, a large number of structural information about RNA binding proteins has become available, and their knowledge has contributed to full understanding of the biological role of these proteins. In addition, these studies revealed that they share structural motifs, such as oligonucleotide/oligosaccharide binding fold (OB-fold), ribonucleoprotein (RNP), double-stranded RNA binding domain (dsRBD), K homology (KH) and helix-turn-helix motif [22]. During the course of studies on three-dimensional structures of ribosomal proteins, it has been known that many ribosomal proteins share structural motifs, even though the structural resemblance is not reflected in an obvious homology at the sequence level [23]. In our own studies, it was found that ribosomal proteins S1, S12, S17, and L2 fold into OB-fold [24], while S6, L5, and L30 share RNP fold [25], though they share few residues in the corresponding domains. These findings have led us

to the expectation that common features occurred in the structural motifs would be invaluable information for non sequence-based prediction of RNA binding proteins.

In this paper, we describe the building of novel hidden Markov models (HMMs) which were derived from structure-based sequence alignment of RNA binding proteins with OB-fold. Their validity was evaluated and discussed by application to PDB40, *Pyrococcus horikoshii* OT3, and rice full length cDNA sequence databases.

2. Materials and methods

2.1 Collection of OB-fold structures

Protein structures with OB-fold were taken from the protein structures belonging to the OB-fold in Structural Classification of Proteins (SCOP) database (release 1.61) [26]. When the structural domains with OB-fold were not notified in the proteins classified into the OB-fold in the SCOP, we manually selected OB-fold domains on the criterion that a five-stranded antiparallel β -sheet forms a closed β -barrel. The atomic coordinates of the OB-folds were obtained from PDB [27]. In this collection, identical structures with a few mutations or in complex with different ligands were omitted from datasets.

2.2 Rectifications of secondary structures

The reliability of our structure alignment method with vector representation of secondary structures is sensitive for their definitions. DSSP [28] included in Sequential Structure Alignment Program (SSAP), which is the program to define secondary structures by the pattern of hydrogen bonding of amino acids in crystal structure coordinates, often gives an inconvenient definition for our structural alignment method because of its strictness. Thus, consecutive long twisting β -strands in OB-fold are defined as a few pieces of short β -strands. This misdefinition of the secondary structure would be caused by a low resolution analysis or disordered structure. When two β -strands in OB-fold form antiparallel β -sheet, and one β -strand forms an angle of more than 100° with the other β -strand, we assumed that the two β -strands would be a single β -strand.

2.3 Pairwise superimposition of OB-fold structures

All combinations of pairwise superimposition of OB-folds were carried out with a vector alignment method of secondary structures, for which the program was written in C++ language on RedHat Linux 8 for IBM PC/AT compatibles with 2.4 GHz pentium4 processor to complete a huge number of superimposition combinations of OB-folds. The vector alignment method is an algorithm partially used at the initial search of the Dali 3D search server [29]. In brief, two pairs of structurally corresponding β -strands of OB-folds were determined by the result of SSAP [30] and transformed to align their structures by overlapping the equivalent β -strands. Structure alignment of a pair of candidates was performed with alignment of corresponding unit vectors along by each secondary structure of the OB-folds. All pairs of OB-folds were superimposed and their original coordinates were transformed to superimposed positions. Then, root-mean-square deviations (RMSD) derived from the superimpositions were calculated with the average distance between two $C\alpha$ points of the SSAP determined equivalent pairs of residues.

2.4 Cluster analysis of OB-fold structures

The cluster analysis with Ward's method [31] using "R 1.7" [32], a language and environment for statistical computing was performed to classify OB-fold structures based on the RMSDs obtained from the structure superimpositions.

2.5 Structure-based multiple alignment

The structure-based multiple sequence alignments were constructed by a heuristic method known as "star alignment method" [33]. Two amino acid sequences, a representative amino acid sequence whose structure has the least average RMSD at the structural superimposition and the other sequence belonging to the same groups, were arranged in two lines:

```
The base sequence      DGIPGRVA
The member sequence    PFRRGVCT
```

where the base and the member sequences indicate amino acid sequences of the representative structure and the other structure in the group, respectively. Equivalent residues of the two sequences were aligned by referring to the SSAP results. When a residue could not be aligned to any residues of the opponent sequence, a gap (-) was inserted into the opponent sequence. All pairwise alignments of the member sequences with the representative sequence were done in the same manner as that described above.

```
The base sequence      D-GIPGRVA-
The member sequence 1  PFR--RGVCT
```

```
The base sequence      DGIPGRV-A--
The member sequence 2  KV-LT--GVVV
```

```
The base sequence      DGIPGR--VA
The member sequence 3  DVI-AGTVV-
```

In this example, there are four members including the representative structure in a group. During pairwise structural alignments, several gaps were inserted at distinct positions in the representative sequences. Then, additional gaps were inserted into the representative sequence so as to give the identical base sequence with several gaps. Next, gaps were inserted into the member sequences so as to give the best alignment with corresponding residues.

```
The base sequence      D-GIPGR--V-A--
The member sequence 1  PFR--RG--V-CT-
```

```
The base sequence      D-GIPGR--V-A--
The member sequence 2  K-V-LT----GVVV
```

```
The base sequence      D-GIPGR--V-A--
The member sequence 3  D-VI-AGTVV----
```

Finally, a set of pairwise alignments was merged to provide a multiple alignment, as shown below.

The base alignment	D-GIPGR--V-A--
The member sequence 1	PFR--RG--V-CT-
The member sequence 2	K-V-LT----GVVV
The member sequence 3	D-VI-AGTVV----

2.6 Building hidden Markov models and database search

Profile hidden Markov models (HMMs) were built from the structure-based multiple alignments of OB-folds to extract common structural properties using the profile HMM software package HMMER [34]. The sequence database of PDB40 [35], a hyperthermophilic archaeon *Pyrococcus horikoshii* OT3 [36] and *japonica* rice full length cDNAs [37] were used for evaluation of HMMs. Databases were searched with an E-value described by HMMER [34].

3. Results and discussion

The five-stranded β -barrel motif designated as OB-fold was first observed in four different proteins which bind oligonucleotides or oligosaccharides: staphylococcal nuclease, anticodon binding domain of asp-tRNA synthetase and B-subunits of heat-labile enterotoxin and verotoxin-1 [38]. The common fold of the four proteins has a five-stranded β -sheet coiled to form a closed

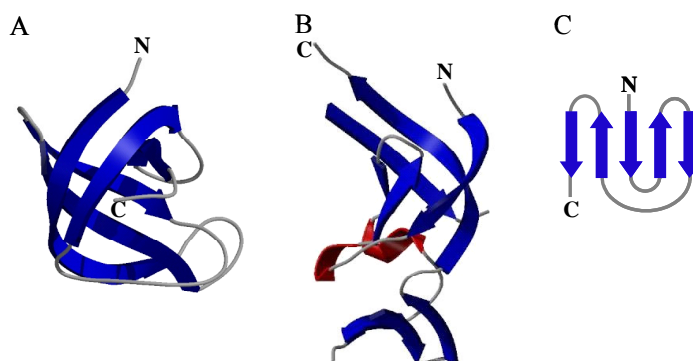


Figure 1. OB-fold structures.

A, OB-fold structure consists of a five-stranded antiparallel β -sheet with a Greek Key β barrel typified by ribosomal protein S12 (1FJF). B, OB-folds occasionally have an insertion of α - or 3_{10} -helix between β -strands and an additional structure consisting of two or three β -strands, as shown in the structure of RNA polymerase subunit RBP4 (1GO3). C, A topological diagram showing the secondary structure of OB-fold. Figures A and B were produced with the programs MOLSCRIPT [39] and Raster3D [40].

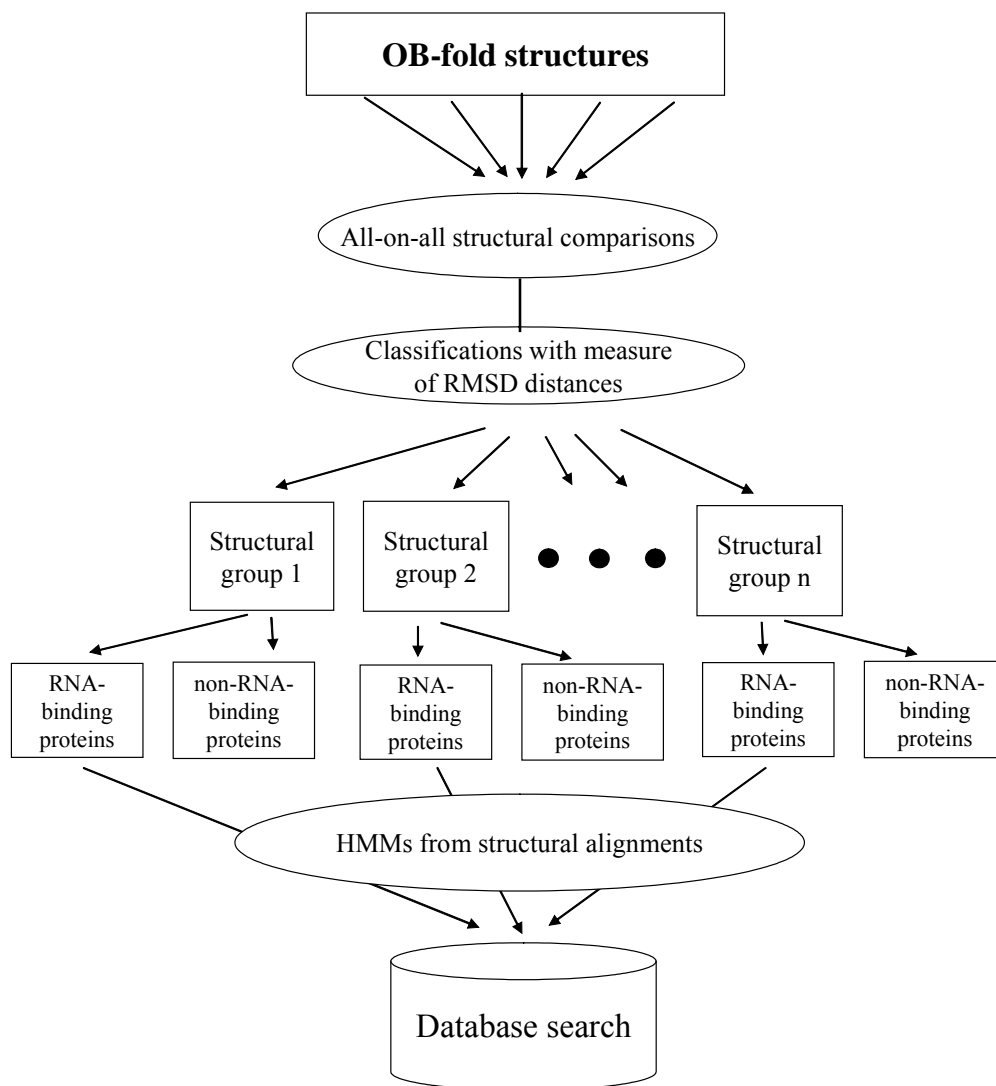


Figure 2. Strategy for developing a novel HMM for prediction of RNA binding proteins with OB-fold.

β -barrel (Figure 1). The barrel frameworks can be superimposed with RMSD of 1.4-2.2 Å, though no similarities can be observed in the corresponding alignment of the four protein sequences. On the assumption that structural information occurred in OB-fold could be extracted from their structural alignments, we attempted to develop a novel computational method for prediction of RNA binding proteins by HMMs. A general strategy for building novel HMMs is presented in Figure 2.

3.1 Superimposition of OB-fold structures

First, protein structures with OB-fold were collected from SCOP database and their atomic coordinates were obtained from PDB, as described under **Materials and methods**. As a result, 84 proteins have OB-fold domains in which 72 proteins have a single OB-fold, whereas 12 proteins are comprised of two OB-folds. In total, 96 OB-fold structures were chosen for the present study (Table 1). Among 96 structures, 36 OB-fold structures are known to have potential to bind RNA.

To gain an insight into a structural relationship, all possible pairwise superimpositions of OB-fold structures were done by using the vector alignment method, as described under **Materials and methods**. It took less than half second to superimpose a pair of OB-folds on the 2.4GHz pentium4 processor excepting the calculation time of SSAP. The reliability of this method for vector alignments of secondary structures depends on definition of secondary structures and their lengths. Indeed, there were 124 small gaps between two consecutive β -strands in 96 OB-fold structures. As described under **Materials and methods**, when the angles between the two β -strands separated by small gaps were more than 100° , they were assumed to be a single β -strand. The results showed that the superimposition of OB-folds from inorganic pyrophosphatase (PDB: 1E9G) and initiation factor eIF2 α (1KL9) gave the least RMSD of 0.939Å, whereas that of tissue inhibitor metalloproteinase (TIMP) (1BR9) and DNA ligase (1FVI) yielded the largest RMSD of 19.677Å. Furthermore, OB-folds from TIMP - eIF2 α , superantigen toxins (1BXT) – ribosomal protein L2 (1RL2), cold shock protein (1G6P) – RNA polymerase subunit RBP8 (1I5O), and molybdate binding protein (1GUT) – bacterial AB5 toxins (1TII) could not be superimposed because of their

Table 1. A list of OB-fold structures used in this study.

OB-folds were searched from SCOP and structure files were obtained from PDB.

RNA binding Proteins			
Initiation Factor	1C0A:A:10-106	1FJF:Q:2-105	RNA Polymerase II
1AH9::6-70	1EFW:A:10-104	1JJ2:A:35-84	1GO3:E:80-167
1BKB::75-139	1KRS::65-149	1RIP::8-64	1I50:H:2-62
1D7Q:A:41-109	1LYL:A:63-150	1RL2:A:71-117	1I50:H:92-146
1JT8:A:19-86	Cold Shock Protein	S1 RNA-binding domain	mRNA Capping Enzyme
2EIF:A:74-132	1C9O:A:1-66	1HH2:P:133-198	1CKM:A:240-300
1KL9:A:14-87	1CSP::1-67	1K0R:A:110-183	RHO Termination Factor
1LUZ:A:10-88	1G6P:A:1-66	1SRO::1-76	1A62::46-125
1JJG:A:31-102	1H95:A:9-76	Myf domain	Staphylococcal Nuclease
tRNA Synthetase	1MJC::2-70	1EUJ:A:7-96	1SNC::7-96
1ASZ:A:105-200	Ribosomal Protein	1GD7:A:10-90	
1B8A:A:14-100	1FJF:L:31-105	1PYS:B:41-118	
non RNA binding Proteins			
ssDNA-binding Protein	1BXT:A:30-117	3CHB:D:12-103	DNA Helicase
1GPC::38-195	1ENF:A:23-105	Molybdate-binding Protein	1BVS:A:1-64
1GVP::1-65	1ESF:A:31-120	1B9M:A:127-183	1CUK::1-64
1JE5:A:3-181	1ET9:A:18-95	1B9M:A:200-252	Metalloproteinase Inhibitor
1JMC:A:193-292	1EU3:A:15-95	1FR3:A:1-67	1BQQ:T:16-109
1JMC:A:311-403	1JCK:B:30-121	1G29:1:241-301	1BR9::15-109
1KXL:A:23-147	1STE::28-120	1G29:1:302-358	1UEA:B:15-107
1OTC:A:51-150	2QIL:A:15-96	1GUT:A:2-68	Chemotaxis Protein CheW
1OTC:A:223-300	3SEB::30-122	1H9M:A:1-73	1K0S:A:30-95
1OTC:A:365-495	Superantigen-like Protein	1H9M:A:74-141	Histidine Kinase CheA
1OTC:B:37-129	1M4V:A:23-100	Inorganic Pyrophosphatase	1B3Q:A:561-626
1PFS:A:1-73	Bacterial AB5 Toxin	1E9G:A:90-162	Laminin-binding Domain of Agrir
1QUQ:A:68-151	1EFI:D:12-103	1I40:A:10-112	1JB3:A:14-110
1QUQ:B:21-90	1PRT:B:99-193	1QEZ:A:1010-1110	RecG
1QVC:A:1-115	1PRT:D:1-110	2PRD::10-112	1GM5:A:165-250
3ULL:A:10-124	1PRT:F:2-99	DNA Ligase	Tail-associated Lysozyme gp5
Bacterial Superantigens	1QNU:A:101-169	1A0I::240-331	1K28:A:6-129
1AN8::20-95	1TII:D:14-98	1DGS:A:315-392	
1B1Z:A:28-107	2BOS:A:102-169	1FVI:A:190-293	

dissimilarity. During the process of superimpositions, the ribosomal protein S17 from *Bacillus stearothermophilus* [41] (1RIP) was unable to be superimposed on other structures. The

reinterpretation of the secondary structure of S17 with DSSP revealed a slight distinct topology from other OB-folds. Hence, no superimposition was further carried out on the protein S17.

3.2 Cluster analysis of OB-fold structures

The cluster analysis of OB-fold structures was done with Ward's method based on the RMSD distances obtained from their overall superimpositions. The RMSD distant dendrogram derived from the cluster analysis is shown in Figure 3. It became clear that OB-folds are classified into four structural groups at 20 rescaled distance cluster combine. The OB-fold structures belonging to groups 1 - 4 are given in Table 2. Proteins belonging to the groups 1 and 2 are predominantly composed of β -strands; those included in the group 1 typically have five-stranded anti-parallel β -sheet, while those in the group 2 have additional short β -strands or some insertional residues between β 3 and β 4 strands. Proteins belonging to the groups 3 and 4 have additional α -helix between β 3 and β 4 strands, while β 1 in the group 3 is rather shorter than the corresponding β -strands in the other groups. The group 1 includes 23 proteins, such as initiation factors (aIF5a, eIF1a), Myf domain, single-stranded DNA binding proteins, and molybdate-binding proteins. Among them, 9 proteins are known to be RNA binding proteins. In the group 2, there are 30 proteins including 20 RNA binding proteins. Ribosomal proteins (S1, S12, S17 and L2), cold shock proteins, initiation factors (IF1, aIF1a, eIF2a), inorganic pyrophosphatase and other 6 kinds of proteins belong to the group 2. The group 3 predominantly includes non-RNA binding proteins, such as eight superantigens (toxic proteins), while the group 4 has 6 tRNA-synthetases (RNA binding proteins) and 28 non-RNA binding proteins, such as SSBs, bacterial B5 toxins and other 5 proteins.

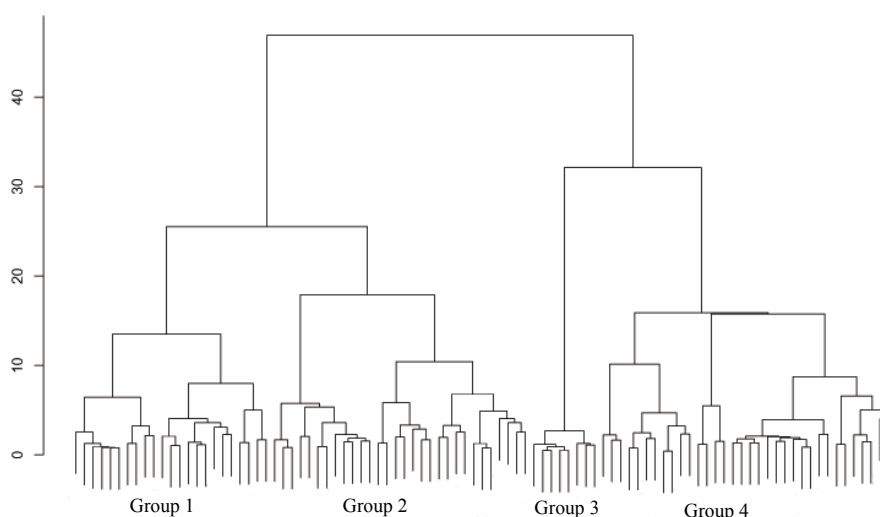


Figure 3. Dendrogram of OB-fold structures.

The dendrogram was drawn from the cluster analysis of OB-fold structures. The result shows that OB-fold structures are classified into four structural groups.

3.3 Structure-based multiple alignment

The amino acid sequences of RNA binding proteins or non-RNA binding proteins belonging to the same OB-fold groups were aligned on the basis of secondary structure, as described under **Materials and methods**.

The structure-based multiple alignments of RNA-binding members in the each group are shown in Figure 4. The average RMSD distances among residues in alignment columns were calculated with the superimposed structure files. In general, the amino acids on β -strands provide short RMSD distances, indicating that the β -strand structures are conserved among OB-folds. Although we manually attempted to find common properties in amino acid sequences of RNA binding proteins belonging to the same groups, general features occurred in the amino acid sequences remained obscure.



Figure 4. Structure-based multiple sequence alignments of RNA binding proteins.

A, B, and C indicate structure-based multiple sequence alignments of RNA binding proteins belonged to groups 1, 2, and 4, respectively. The secondary structures of the base sequence of the each multiple alignment are shown at the top of the alignments.

3.4 Building profile hidden Markov models

Profile HMM is a probabilistic model based on a consensus amino acid sequence; it presents profiles from its multiple sequence alignments as a position-specific scoring system suitable for

searching distant homologous proteins. There are position-specific scores for a distribution of amino acids and the scores for opening and extending a residue insertion or deletion in the position-specific scoring system of profile HMM. This nature of the profile HMM captures valuable information about properties of conserved at positions in the multiple alignments. We built HMMs based on the multiple alignments of RNA binding proteins belonging to each group as shown in Figure 4. HMMs were also built on the basis of the structure-based multiple alignments of non-RNA binding proteins as a control.

To evaluate the HMMs obtained in this study, we tested them by application to the PDB40 sequence dataset [35]. Since a number of domain sequences in the PDB are very similar to others and have redundancy, the sequence database PDB40 included in ASTRAL SCOP 1.63 [35] was used in the present study. The PDB40 has a subset of 5,226 sequences that consists almost entirely of distinct protein domains, derived from the SCOP database to yield the largest set with pairwise sequence identities of 40% or less [35]. The result indicated that RNA binding proteins with OB-fold (35 proteins) and non-RNA binding proteins with OB-fold listed in Table 2 were selected with the E-values lower than 1.0 by HMMs derived from RNA binding proteins and non-RNA binding proteins, respectively. This result suggests that HMMs obtained by structure-based multiple alignment may extract information to predict RNA binding proteins with OB-fold. It is, however, noted that there are a few false positives: that is, non-RNA binding proteins, such as Bacterioferritin [42] and Vinculin [43] are selected, giving E-values of 5.3 and 1.4, respectively, by using HMMs derived from RNA binding proteins. It is reported that proteins which give an E-value lower than 10.00 are assumed to match up to a HMM [34]. Therefore, we have used two different E-value cutoffs (1.0 and 10.0) to predict RNA binding proteins with OB-fold.

To demonstrate the validity of HMMs derived from the structure-based multiple alignment of RNA binding proteins with OB-fold, the results described above were compared with those obtained by HMMs from sequence-based multiple alignment of OB-fold in the SCOP database. For this purpose, the amino acid sequences of RNA binding proteins with OB-fold in the SCOP database were aligned with ClustalW 1.83 [44], HMMs were built by HMMER as with those derived from the structure-based multiple alignment, and then applied to the PDB40. The analysis showed that 22 RNA binding proteins with OB-fold were selected, which was about 63% ($22/35 = 0.63$) compared with those obtained by HMMs derived from the structure-based alignment. The result demonstrates the authenticity of HMMs derived from the structure-based multiple alignment of RNA binding proteins with OB-fold.

Next, we tested them by application to the sequence database of *Pyrococcus horikoshii* OT3: a hyperthermophilic archaeon whose genome comprises 2061 open reading frames in which 557 proteins are functionally annotated [36]. The prediction using the HMMs showed that 60 proteins gave E-values less than 10.0, of which 13 proteins were suggested to be RNA-binding proteins with OB-fold, giving E-values lower than 1.0 (Table 3). Among 13 proteins, 9 proteins are annotated to be RNA binding proteins, while 4 proteins are hypothetical proteins in the database. It is thus likely that the four hypothetical proteins may be RNA binding proteins. In addition, two methionyl-tRNA synthetases (PH0285, PH0993), isoleucyl-tRNA synthetase (PH1065), ribosomal protein L10E (PH1999), eIF2 gamma subunit (PH1706) and transcriptional regulatory protein hypF (PH0897) which were not included in the OB-fold dataset in SCOP were predicted to be RNA-binding proteins. Although these *P. horikoshii* protein structures have not been determined, the C-terminal domain of methionyl-tRNA synthetase from *P. abyssi* is categorized in OB-fold in the latest version of SCOP [45]. Moreover, the ribosomal protein L10E from *H. marismortui* is predominantly composed of antiparallel β -sheets with α -helix [18], and the eIF2 gamma subunit from *P. abyssi* is known to be all β protein with a Greek Key β barrel similar to OB-fold [46]. In contrast, the application of the HMMs derived from the structure-based alignments of non-RNA binding proteins

to the *P. horikoshii* database predicted 36 proteins to be OB-fold structures with E-value cutoff of 10.0, of which only one RNA binding protein (RNA methyltransferase) was predicted. These results again suggest that HMMs thus obtained may extract information to predict RNA binding proteins with OB-fold.

Finally, we searched RNA binding proteins with OB-fold in the rice full-length cDNA database release on October 2003 [37]. The result showed that 24 proteins, including 10 annotated and 14 hypothetical proteins, were suggested to be RNA binding proteins with OB-fold, yielding E-values less than 1.0 (Table 4). Actually, all 10 annotated proteins are known to be RNA binding proteins. It is suggested that 14 hypothetical proteins may be RNA binding proteins.

Table 2. Classification of OB-fold structures.

Group 1	Group 2	Group 3	Group 4
<u>RNA binding Proteins</u>	<u>RNA binding Proteins</u>	<u>non RNA binding Proteins</u>	<u>RNA binding Proteins</u>
Initiation factor	Ribosomal Protein	Bacterial Superantigen	tRNA Synthetase
1BKB::75-139	1FJF:L:31-105	1B1Z:A:28-107	1ASZ:A:105-200
1D7Q:A:41-109	1FJF:Q:2-105	1BXT:A:30-117	1B8A:A:14-100
2EIF:A:74-132	1JJ2:A:35-84	1ENF:A:23-105	1C0A:A:10-106
Myf domain	1RL2:A:71-117	1ESF:A:31-120	1EFW:A:10-104
1EUJ:A:7-96	S1 RNA-binding domain	1ET9:A:18-95	1KRS::65-149
1GD7:A:10-90	1HH2:P:133-198	1JCK:B:30-121	1LYL:A:63-150
1PYS:B:41-118	1K0R:A:110-183	1STE::28-120	
Staphylococcal nuclease	1SRO::1-76	3SEB::30-122	<u>non RNA binding Proteins</u>
1SNC::7-96	Cold Shock Protein		ssDNA-binding protein
mRNA Capping Enzyme	1C9O:A:1-66		1GVP::1-65
1CKM:A:240-300	1CSP::1-67		1JE5:A:3-181
RNA polymerase II	1G6P:A:1-66		1JMC:A:311-403
1I50:H:2-62	1H95:A:9-76		1KXL:A:23-147
	1MJC::2-70		1OTC:A:51-150
<u>non RNA binding Proteins</u>	Initiation factor		1OTC:A:365-495
DNA ligase	1AH9::6-70		1OTC:B:37-129
1DGS:A:315-392	1JJG:A:31-102		1PFS:A:1-73
1FVI:A:190-293	1JT8:A:19-86		1QUQ:A:68-151
Tail-associated lysozyme gp5	1KL9:A:14-87		1QUQ:B:21-90
1K28:A:6-129	1LUZ:A:10-88		1QVC:A:1-115
ssDNA-binding protein	RNA polymerase II		3ULL:A:10-124
1GPC::38-195	1GO3:E:80-167		DNA helicase
1JMC:A:193-292	1I50:H:92-146		1BVS:A:1-64
1OTC:A:223-300	RHO termination factor		1CUK::1-64
Molybdate-binding protein	1A62::46-125		RecG
1B9M:A:200-252			1GM5:A:165-250
1B9M:A:127-183	<u>non RNA binding Proteins</u>		Bacterial AB5 toxins
1FR3:A:1-67	DNA ligase		1EFI:D:12-103
1G29:1:241-301	1A0I::240-331		1PRT:B:99-193
1G29:1:302-358	Bacterial Superantigen		1PRT:D:1-110
1GUT:A:2-68	1AN8::20-95		1PRT:F:2-99
1H9M:A:74-141	1EU3:A:15-95		1QNU:A:101-169
1H9M:A:1-73	2QIL:A:15-96		1TII:D:14-98
	Inorganic pyrophosphatase		2BOS:A:102-169
	1E9G:A:90-162		3CHB:D:12-103
	1I40:A:10-112		Superantigen-like Protein
	1QEZ:A:1010-1110		1M4V:A:23-100
	2PRD::10-112		Metalloproteinase Inhibitor
	Histidine kinase CheA		1BQQ:T:16-109
	1B3Q:A:561-626		1BR9::15-109
	Laminin-binding domain of agrin		1UEA:B:15-107
	1JB3:A:14-110		Chemotaxis protein CheW
			1K0S:A:30-95

Table 3. Predicted RNA binding proteins in *P. horikoshii* OT3.

The database was searched by HMMER with restriction of E-value < 10.0. Among 31 annotated proteins, 16 proteins with boldface are annotated as RNA binding proteins, while 15 other proteins are indicated by underlines. RNA binding proteins boxed were not used to build the HMMs.

Group 1			Group 2			Group 3			Group 4		
Name	Score	E-value	Name	Score	E-value	Name	Score	E-value	Name	Score	E-value
PH0993	44.4	9.1E-11	PH0961	39.3	3.1E-09	PH0928	-5.3	3.6	PH1507	-7.6	6.7
PH1381	22.4	0.00037	PH1908	38.3	5.9E-09	PH1100	-5.4	3.6	PH1882	-7.9	7.2
PH0536	4.8	0.16	PH1568	23.1	0.00024	<u>PH1647</u>	-5.9	4.2	PH1706	-8.0	7.3
PH0285	-2.1	1.0	PH1775	18.4	0.0059	<u>PH1399</u>	-6.0	4.3	PH1179	-8.0	7.4
PH1212	-4.0	1.7	PH0340	2.2	0.48	<u>PH0275</u>	-6.0	4.4	PH1740	-8.1	7.5
PH0117	-4.2	1.8	PH0043	0.3	0.79	PH0128	-6.6	5.1	PH0988	-8.1	7.6
PH0822	-5.3	2.4	PH1770	-1.5	1.3	<u>PH1907</u>	-6.6	5.1	PH0326	-8.2	7.7
PH1995	-5.8	2.8	PH0484	-1.7	1.3	PH0235	-6.9	5.5	PH0121	-8.3	8.0
PH1740	-5.9	2.8	PH1065	-2.3	1.6	PH0346	-7.0	5.6	PH0863	-8.9	9.4
PH1568	-6.1	3.0	<u>PH1353</u>	-3.4	2.2	<u>PH1349</u>	-7.2	6.0	PH0387	-9.0	9.8
PH1999	-8.3	5.2	PH0117	-3.5	2.2	<u>PH0800</u>	-7.3	6.2	PH0897	-9.1	9.8
<u>PH1622</u>	-10.3	8.9	<u>PH1262</u>	-4.0	2.5	PH0841	-7.5	6.4	<u>PH1593</u>	-9.1	9.9
			PH0650	-5.1	3.4	<u>PH0764</u>	-7.5	6.5			

Table 4. Predicted RNA binding proteins in the rice full-length cDNA database.

The database was searched by HMMER with restriction of E-value < 10.0. Among 14 annotated proteins, 11 proteins with boldface are annotated as RNA binding proteins, while 3 other proteins are indicated by underlines. RNA binding proteins boxed were not used to build the HMMs.

Group 1			Group 2			Group 3			Group 4		
Name	Score	E-value	Name	Score	E-value	Name	Score	E-value	Name	Score	E-value
J023012P17	43.0	2.4E-09	J013001O08	44.1	1.1E-09	J023095J08	4.2	2.9	002-160-A04	61.6	6.1E-15
002-157-A08	40.7	1.2E-08	J023036A08	42.5	3.4E-09	J023136G11	3.8	3.2	J013071I10	58.6	5E-14
002-137-H02	36.5	2.3E-07	J013106B18	39.2	3.4E-08	J033046J06	2.0	5.2	002-117-F07	41.2	8.5E-09
001-036-F11	21.1	0.0097	002-159-D04	35.2	5.4E-07	001-027-E01	0.6	7.6	002-126-D07	32.5	3.5E-06
			002-171-F11	33.9	1.4E-06	001-044-D07	0.3	8.2	002-165-A01	27.8	9.4E-05
			J023075D13	31.2	8.6E-06	002-149-B07	-0.3	9.7	J023105D18	23.2	0.00035
			J023061H10	30.2	1.8E-05				001-121-B05	13.2	0.0028
			001-020-G05	29.5	2.9E-05				J033070O21	2.9	0.025
			J023088B11	27.8	9.3E-05				<u>J023027K10</u>	15.2	1.1
			001-203-D03	24.3	0.001				J023123B03	22.2	5.0
			J033149P22	22.4	0.0039				<u>J013157O09</u>	24.7	8.4
			J013110C21	14.2	0.19				J023001J17	24.9	8.8
			<u>J013106A17</u>	5.9	1.8				J013146B15	25.4	9.7

3.5 In conclusion

We are currently concerned with structural genomics project on proteins involved in transcription and translation. One major motivation of structural genomics projects is that the determination of the structure of a protein provides insight into its molecular function, which is a step toward understanding its cellular function. It is now common knowledge that RNA molecules play essential roles in transcription and translation. Usually, RNA molecules perform these functions in close association with RNA binding proteins. Hence, there is a pressing requirement for computational methods for prediction of RNA binding proteins in vast amounts of genome sequence information. Recently, several groups have developed algorithms to predict protein function often employing either sequence-based or structure-based comparisons. However, the methods specific for a given function or fold are limited. The result obtained in this study demonstrates feasibility of HMMs derived from structure-based multiple alignments for prediction

of RNA binding proteins with OB-fold. Although OB-fold selected in the present study is a characteristic fold, predominantly comprising of β -strands, the strategy presented in this paper can be in principle applicable to any other structural motifs found in RNA binding proteins. Since some of the motifs have no specific biological function alone but are part of larger structural and functional assemblies, the present method would provide clues as to protein functions of unannotated proteins and also be useful for a target selection for structural genomics.

This work was supported in part by a grant of Rice Genome Project PR-3007, MAFF, Japan and a grant from the National Project on Protein Structural and Functional Analyses, Japan.

References

- [1] H.W. Mewes, K. Albermann, K. Heumann, S. Liebl, F. Pfeiffer, *Nucleic Acids Res.*, **25**, 28-39 (1997).
- [2] E. Pennisi, *Science*, **279**, 978-979 (1998).
- [3] T.I. Zarembinski, L.W. Hung, H.J. Mueller-Dieckmann, K.K. Kim, H. Yokota, R. Kim, S.H. Kim, *Proc. Natl. Acad. Sci. USA.*, **95**, 15189-15193 (1998).
- [4] G.T. Montelione, S. Anderson, *Nature Struct. Biol.*, **6**, 11-12 (1999).
- [5] D. Christendat, A.Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J.R. Cort, V. Booth, C.D. Mackereth, V. Saridakis, I. Ekiel, G. Kozlov, M.A. Kennedy, A.R. Davidson, F.F. Pai, M. Gerstein, A.M. Edwards, C.H. Arrowsmith, *Nature Struct. Biol.*, **7**, 903-909 (2000).
- [6] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, *Proc. Natl. Acad. Sci. USA.*, **95**, 14863-14868 (1998).
- [7] M. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M.Jr. Ares, D. Houssler, *Proc. Natl. Acad. Sci. USA.*, **97**, 262-267 (2000).
- [8] C.A. Wilson, J. Kreychman, M. Gerstein, *J. Mol. Biol.*, **297**, 233-249 (2000).
- [9] D. Devos, A Valencia, *Proteins*, **41**, 98-107 (2000).
- [10] B. Rost, *J. Mol. Biol.*, **318**, 595-608 (2002).
- [11] D.B. Kell, R.D. King, *Trends Biotechnol.*, **18**, 93-98 (2000).
- [12] R.D. King, A. Karwath, A. Clare, L. Dehaspe, *Bioinformatics*, **17**, 445-454 (2001).
- [13] A. Clare, R.D. King, *Bioinformatics*, **18**, 160-166 (2002).
- [14] A.R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, *Proteins Suppl*, **3**, 177-185 (1999).
- [15] J. Pillardy, C. Czaplowski, A. Liwo, J. Lee, D.R. Ripoll, R. Kazmierkiewicz, S. Oldziej, W.J. Wedemeyer, K.D. Gibson, Y.A. Arnautova, J. Saunders, Y.J. Ye, H.A. Scheraga, *Proc. Natl. Acad. Sci. USA.*, **98**, 2329-2333 (2001).
- [16] K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, D. Baker, *Proteins*, **34**, 82-95 (1999).
- [17] R. Bonneau, D. Baker, *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173-189 (2001).
- [18] A. Bateman, L. Coin, R. Darbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, et al., *Nucleic Acids Res.*, Database issue: D138-D141 (2004).
- [19] R. Bonneau, N.S. Baliga, E.W. Deutsch, P. Shannon, L. Hood, *Genome Biology*, **5**, R52 1-15 (2004).
- [20] C.B. Burge, T. Tuschl, P.A. Sharp, Splicing of precursors to mRNAs by the spliceosomes. In *The RNA word*, 2nd ed., Gesteland, R.F., Cech, T.R., and Atkins, J.E. eds. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press pp 525-560 (1999).

- [21] N. Ban, P. Nissen, J. Hansen, P.B. Moore, T.A. Steitz, *Science*, **289**, 905-920 (2000).
- [22] C.G. Burd, G. Dreyfuss, *Science*, **265**, 615-621 (1994).
- [23] V. Ramakrishnan, S.W. White, *Trend Biochem. Sci.*, **23**, 208-212 (1998).
- [24] A. Nakagawa, T. Nakashima, M. Taniguchi, H. Hosaka, M. Kimura, I. Tanaka, *EMBO J.*, **18**, 1459-1467 (1999).
- [25] T. Nakashima, M. Yao, S. Kawamura, K. Iwasaki, M. Kimura, I. Tanaka, *RNA*, **7**, 692-701 (2001).
- [26] A.G. Murzin S.E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.*, **247**, 536-540 (1995).
- [27] E.E. Abola, J.L. Sussman, J. Prilusky, N.O. Manning, *Methods Enzymol.*, **276**, 556-571 (1997).
- [28] W. Kabsch, C. Sander, *Biopolymers*, **22**, 2577-637 (1983).
- [29] L. Holm, C. Sander, *J. Mol. Biol.*, **233**, 123-138 (1993).
- [30] R. William, R. Taylor, A. Christine, A. Orenge, *J. Mol. Biol.*, **208**, 1-22 (1988).
- [31] J.H. Ward, *J Am Stat Assoc.*, **58**, 236-244 (1963).
- [32] A. Richard, A. Becker, J.M. Chambers, A.R. Wilks, *The New S Language*. Chapman & Hall, New York. (1988).
- [33] S.F. Altschul, D.J. Lipman, *SIAM J. Appl. Math.*, **49**, 197-209 (1989).
- [34] S. Eddy, HMMER: profile HMMs for protein sequence analysis. <http://hmmer.wustl.edu/> (2003).
- [35] S.E. Brenner, P. Koehl, M. Levitt, *Nucleic Acids Res.*, **28**, 254-256 (2000).
- [36] Y. Kawarabayashi, et al., *DNA Res.*, **5**, 55-76 (1998).
- [37] S. Kikuchi, et al., *Science*, **301**, 376-379 (2003).
- [38] A.G. Murzin, *EMBO J.*, **12**, 861-867 (1993).
- [39] P.J. Kraulis, *J. Appl. Crystallogr.*, **24**, 946-950 (1991).
- [40] E.A. Merrit, M.E.P. Murphy, *Acta Crystallogr.*, **D50**, 869-873 (1994).
- [41] B.L. Golden, D.W. Hoffman, V. Ramakrishnan, S.W. White, *Biochemistry*, **32**, 12812-12820 (1993).
- [42] F. Frolow, A.J. Kalb, J. Yariv, *Nat.Struct. Biol.*, **1**, 453-460 (1994).
- [43] C. Bakolitsa, J.M. de Pereda, C.R. Bagshaw, D.R. Critchley, R.C. Liddington, *Cell*, **99**, 603-613 (1999).
- [44] J.D. Thompson, D.G. Higgins, T.J. Gibson, *Nucleic Acids Res.*, **22**, 4673-4680 (1994).
- [45] T. Crepin, E. Schmitt, S. Blanquet, Y. Mechulam, *Biochemistry*, **43**, 2635-2644 (2004).
- [46] E. Schmitt, S. Blanquet, Y. Mechulam, *EMBO J.*, **21**, 1821-1832 (2002).