DNA Research **14**, 91–102, (2007) doi:10.1093/dnares/dsm011

Proteome-Wide Prediction of Novel DNA/RNA-Binding Proteins Using Amino Acid Composition and Periodicity in the Hyperthermophilic Archaeon *Pyrococcus furiosus*

Kosuke Fujishima[1,2], Mizuki Komasa[1,2], Sayaka Kitamura[1,2], Haruo Suzuki[1,2], Masaru Tomita[1,3], and Akio Kanai[1,3,*]

Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0017, Japan[1]; Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-8520, Japan[2] and Faculty of Environment and Information Studies, Keio University, Fujisawa 252-8520, Japan[3]

(Received 13 March 2007; accepted on April 30, 2007; published online 15 June 2007)

Abstract

Proteins play a critical role in complex biological systems, yet about half of the proteins in publicly available databases are annotated as functionally unknown. Proteome-wide functional classification using bioinformatics approaches thus is becoming an important method for revealing unknown protein functions. Using the hyperthermophilic archaeon *Pyrococcus furiosus* as a model species, we used the support vector machine (SVM) method to discriminate DNA/RNA-binding proteins from proteins with other functions, using amino acid composition and periodicities as feature vectors. We defined this value as the composition score (CO) and periodicity score (PD). The *P. furiosus* proteins were classified into three classes (I–III) on the basis of the two-dimensional correlation analysis of CO score and PD score. As a result, approximately 87% of the functionally known proteins categorized as class I proteins (CO score + PD score > 0.6) were found to be DNA/RNA-binding proteins. Applying the two-dimensional correlation analysis to the 994 hypothetical proteins in *P. furiosus*, a total of 151 proteins were predicted to be novel DNA/RNA-binding protein candidates. DNA/RNA-binding activities of randomly chosen hypothetical proteins were experimentally verified. Six out of seven candidate proteins in class I possessed DNA/RNA-binding activities, supporting the efficacy of our method.

Key words: DNA/RNA-binding protein; amino acid periodicity; support vector machine; archaea

1. Introduction

The last decade has been a remarkable time in the field of genome science. DNA sequences from over 2400 species have been determined,[1] and more are on the way. Correspondingly, the need for reliable functional annotation has become prominent. Most functional annotation is based on a sequence similarity approach,[2] but about half of the proteins registered in protein databases are classified as hypothetical because they lack similarity to functionally known proteins. Proteome-wide functional classification using bioinformatics approaches is becoming an important method for revealing unknown protein functions. For example, the recent exponential growth in Protein Data Bank entries has enabled highly accurate functional predictions to be made on the basis of structural similarities to three-dimensional profiles of proteins.[3,4] Comparative genome analysis using phylogenetic profiling has revealed a diversity of functional linkages among genes, and thus it can be a useful strategy for elucidating the functions of uncharacterized proteins.[5] However, although species-specific genes (ORFans) are known to encode many uncharacterized short peptides,[6] the functions of these peptides are difficult to predict with certainty using comparative genomics because they lack homology to

Edited by Takashi Ito.

* To whom correspondence should be addressed. Tel. +81 235-29-0524. Fax. +81 235-29-0525. E-mail: akio@sfc.keio.ac.jp

those sequences currently in databases. More than 23 000 ORFans have been found in 60 microbial genomes, and, on the basis of structural studies, many are likely to encode expressed, functional, or even essential proteins.[7] Therefore, alternative bioinformatics methods that can predict these uncharacterized protein functions at the proteome level are very useful.

For the past few years, we have been working on RNA metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*[8–10] and reported on our experimental system in which an expression cloning method is used for extracting DNA/RNA-binding proteins at the proteome level. During this work, we observed that charged amino acids—such as aspartic acid, glutamic acid, arginine, and lysine—appeared both in the sequence of the novel RNA-binding protein FAU-1 and Ribonuclease E in a periodic manner.[8] It is possible that certain acidic and basic amino acid periodicities might affect the secondary or tertiary structure of a protein so that it gains DNA/RNA-binding activities. Amino acid periodicities are commonly observed features in the sequences of various proteins such as myosin and amyloids,[11] serine–threonine, and tyrosine protein kinases[12] and are known to be strongly correlated with their secondary structures.

The purpose of the current study was to demonstrate that a bioinformatics approach focusing on the periodicity in a protein's primary structure could be a suitable method for elucidating DNA/RNA-binding proteins. Previously, several support vector machine (SVM)-based methods were developed towards predicting DNA-binding and RNA-binding proteins on the basis of various amino acid profiles (i.e. overall composition, pseudo-amino acid composition, surface composition, electrostatic potential, and hydrophobicity).[13–16] SVM is one of the most powerful supervised learning algorithm that has recently widely been used in the field of bioinformatics. We describe here an SVM-based method for classifying known DNA/RNA-binding proteins from *P. furiosus* using amino acid composition and periodicity as feature vectors. The discriminant values (SVM output) derived from these profiles were defined as two new indices: composition (CO) score and periodicity (PD) score. Amino acid composition are known to be strongly correlated with protein secondary structure class[17] and subcellular localization[18,19] and are assumed to support the protein function classification. Therefore, on the basis of the two-dimensional correlation analysis, we combined amino acid composition (CO score) with the PD score to further improve the performance of DNA/RNA-binding protein prediction. The two-dimensional correlation analysis was then applied to hypothetical proteins of *P. furiosus*, and promising candidates for being novel DNA/RNA-binding proteins were selected. DNA/RNA-binding activities of these candidate proteins were examined experimentally and many of them were confirmed to possess DNA/RNA-binding activities.

2. Materials and methods

2.1. Protein data set and functional annotations

Automated annotations and amino acid sequences of proteins from the two archaeal species, *P. furiosus* (2057 proteins) and *Sulfolobus solfataricus* (2934 proteins), were taken from the EMBL database (http://www.ebi.ac.uk/embl/; Release 83, June 2005). Each protein entry has a UniProt Knowledgebase (UniProtKB) accession code corresponding to its entry in either UniProt KB/Swiss-Prot (http://www.ebi.ac.uk/Swiss-Prot/; Release 47, May 2005) or UniProtKB/TrEMBL (http://www.ebi.ac.uk/trembl/; Release 31, September 2005). Both databases contain information on the gene ontology annotation (GOA: a combination of electronic assignment and manual annotation), and protein data are from the domain databases InterPro[20] and Pfam.[21] Swiss-Prot data were used for the four prokaryotic and eukaryotic species—*Bacillus subtilis* (2799 proteins), *Escherichia coli* K12 MG1655 (4465 proteins), *Arabidopsis thaliana* (3454 proteins), and *Caenorhabditis elegans* (2655 proteins)—as a reliable independent test set.

We defined 'functionally known proteins' as functionally annotated proteins in the Swiss-Prot or TrEMBL databases with additional GOA. TrEMBL protein entries with no additional annotation were categorized as 'putative functional proteins'. Proteins annotated as 'hypothetical' in the database were defined as 'hypothetical proteins'. DNA/RNA-binding proteins were defined as those proteins whose annotations included the following keywords in Swiss-Prot, TrEMBL, and GOA annotations: DNA, RNA, ribosome(al), RNP, ribonucleo-, helicase, nuclease, or nucleic acid binding. To reduce the bias of functional variety in the protein data set, the functionally known proteins of the six model species were filtered to remove homologous proteins at sequence identity level with E-value $< 1 \times 10^{-4}$ and short peptides <20 amino acids from future analyses. In total, we prepared 477 proteins of *P. furiosus*, 582 proteins of *S. solfataricus*, 914 of *B. subtilis*, 1436 of *E. coli*, 865 of *A. thaliana*, and 566 of *C. elegans* as a 'representative set' for the analysis (Table 1).

2.2. Amino acid periodicity

To analyze amino acid periodicities, we used eight physico-chemical profiles (chemical, Sneath, Dayhoff, Stanfel, functional, charge, structural, and hydrophobicity)[22] to subdivide the 20 common amino acids into groups. For example, the 'charge' profile divided the 20 amino acids into three groups: DE, RKH, and others (ACFGILMNPQSTVW). In total, 23 amino acid groups were identified: DE, RK, NQ, CM, ST, ILV, RKH, FYW, AGP, MNQ, CST, DEQN, FHWY, AGPST, GAVLIP, DERKH, CGNQSTY, ACGPSTWY, RNDQE HK, ILMFV, AFILMPVW, ACGILMPSTV, and CDE GHKNQRSTY.

Table 1. Functional classification table of the proteome data set of six model species

Species	Database	Functionally known					Putative	Hypothetical	Total protein
		Representative set[a]			Redundant Total				
		DNA/RNA	Others	Total					
P. furiosus	TrEMBL + Swiss-Prot	157	320	477	465	942	121	994	2057
S. Solfataricus	TrEMBL + Swiss-Prot	184	398	582	730	1312	302	1320	2934
B subtilis	Swiss-Prot	204	710	914	908	1822	1	976	2799
E. coli	Swiss-Prot	346	1090	1436	1889	3325	1	1139	4465
A. thaliana	Swiss-Prot	223	642	865	2389	3254	56	144	3454
C. elegans	Swiss-Prot	165	401	566	1215	1781	0	874	2655

[a]Representative set consists of proteins with amino acid length > 20 and homology redacted using BLASTP (E-value $< 1 \times 10^{-4}$).

Amino acid periodicity was defined as the regular appearance of a certain amino acid group (X), $Y(Y \geq 3)$ times in a protein sequence with a period (the number of amino acids from one appearance to the next) of Z. Although a previous analysis in *E. coli* defined the range of periodicity as 2 to 50, to eliminate binal periodicities (ex: period 5 includes period 10), we used prime numbers and their multiples [2, 3, 5, 7, 8 (2×4), 9 (3×3), 11, 13, 15 (5×3), 17, 19]. To take into account the fluctuation of periodicities, we set the error range as ± 1. For example, in seq1 (XXXXAXXAXXXX), 'A' appears only twice, so no periodicity can be defined. Seq2 (XXBXXXXBXXXXBX) contains three 'Bs' with a period of five ('B-5' periodicity). Seq3 (XCXXXCXXCX XXCXXCX) contains five 'Cs' with multiple periodicities (two of length 3, two of length 4, and two of length 7). On the basis of the error range $\pm$ 1, length 4 is included in length 3; therefore, Seq3 is defined to have 'C' periods of only 3 and 7.

2.3. SVM classification of DNA/RNA-binding proteins based on amino acid periodicity and composition

SVM is a non-linear classifier creating a maximum-margin hyperplane by applying a kernel trick to the feature vectors. We performed two different SVM analysis on the basis of the individual data set of amino acid periodicity and amino acid composition. For amino acid periodicity, we calculated the relative coverage of the periodic region (R) of each training set (i) with 253 patterns of amino acid periodicities (j): 23 amino acid groups $\times$ 11 kinds (2, 3, 5, 7, 8, 9, 11, 13, 15, 17, 19 periods):

$$R_{ij} = \frac{P}{N}$$

where P is the length of periodic region of periodicity j in a single protein i, and N is the full amino acid length of a single protein i. Thus, a transformed feature space is created from 253-dimensional feature vectors of periodic region R.

For amino acid composition, we calculated the relative composition of amino acid (C) of each training set (i) with 20 types of amino acids (k)

$$C_{ik} = \frac{A}{N}$$

where A is the number of amino acid k in a single protein i, and N is the full amino acid length of a single protein i.

These factors were applied as feature vectors and classified into two distinct members: DNA/RNA-binding proteins and proteins with other functions. For SVM training, the data label for DNA/RNA-binding proteins was denoted as 1 and proteins with other functions was denoted as -1. SVM analysis in this study was performed using the default parameters in Gist package version 2.3, which contains software tools for SVM classification.[23] We have tested two types of kernel function (linear and radial basis) and selected the kernel function with higher prediction performance. As a consequence, maximum-margin hyperplane was applied to the protein test set on the basis of a radial basis function kernel ($r = 1$), and the discriminant value of each protein were defined as the PD score. Likewise, linear kernel-based maximum-margin hyperplane was applied for the protein set on the basis of amino acid composition, and discriminant values were defined as the CO score.

2.4. Validation of PD score performance

The performance of the PD score at predicting novel proteins was validated on the basis of 10-fold cross-validation test. The 10-fold cross-validation test is one of the most reliable methods for estimating the performance of the predictor. For example, the 477 representative data set of *P. furiosus* was randomly split into 10 mutually exclusive subsets D_1, $D_2, \ldots,$ D_{10} of approximately equal size. Each subset was tested on the basis of the training using the rest of the nine subsets. Estimated accuracies were derived as average values.

First, the classification accuracy of the PD score was compared with that of the randomly chosen single

amino acid periodicities using receiver operating characteristic (ROC) studies. The ROC curve is a simple and effective method to compare the overall prediction performance of different methods including SVMs.[24–26] The ROC curve is represented by two indices: sensitivity and specificity. The sensitivity and specificity of the PD score were calculated using a 10-fold cross-validation test with a PD score cut-off of 0. Equations are represented as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TP refers to true positive (number of DNA/RNA-binding proteins with PD score > cut-off), FP refers to false positive (number of other proteins with PD score > cut-off), FN refers to false negative (number of DNA/RNA-binding proteins with PD score < cut-off), and TN refers to true negative (number of other proteins with PD score < cut-off).

Error bars were added for each data set representing the standard deviation values derived from the 10-fold cross-validation test. Secondly, PD score was compared against CO score (amino acid composition-based SVM) and other SVM-based protein function predictor, SVM-Prot.[27] To assess the PD score performance, we calculated the overall accuracy (ACC) for PD score using 10-fold cross-validation test. Training data set of SVM-Prot is fixed as a combination of 54 functional protein families and predicts several functional classes owing to the probability of correct prediction. SVM-Prot uses 1943 positive set and 1353 negative set for training DNA-binding proteins and 871 positive set and 1120 negative set for training RNA-binding proteins. In total, 104 feature vectors [composition (C), transition (T), and distribution (D)] were calculated for each amino acid group classified by four physicochemical properties, hydrophobicity, Van der Waals volume, polarity, and polarizability, and were introduced to generate the hyperplane for each protein family.[27,28] To equally validate the prediction performance of SVM-Prot with our method, the proteins that were predicted as DNA/RNA-related with the highest probability were regarded as DNA/RNA-binding proteins. SVM-Prot was applied to the representative data set and ACC is calculated as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \times 100(\%)$$

2.5. *Data fusion and threshold determination*

Two heterogeneous data set (amino acid periodicity and amino acid composition) were integrated by three different approaches (early, intermediate, and late integration).[29] Early integration performs single SVM with a feature vector of 253 (periodicity) and 20 (composition) dimensions. For intermediate integration, kernel values (kernel matrix) are separately computed on each data type and the summed kernel values are used for the training of SVM. Late integration performs SVM separately and later, the discriminant values were summed (i.e. CO + PD score). Three evaluation indices, SE, SP, and ACC, were calculated for the evaluation of three different integration methods.

Thresholds for extracting DNA/RNA-binding proteins were determined by considering several indices. An index, positive predictive value (PPV), was adapted to measure the percentage of DNA/RNA-binding proteins among proteins above threshold (blue line in Supplementary Fig. S1). PPV is calculated as follows:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100(\%)$$

The final prediction decision is given by the calculated value of the Matthews correlation coefficient (MCC)[30] to determine the threshold value of the CO + PD score. MCC is a popular index for measuring the performance of prediction; maximum MCC provides efficient sensitivity and specificity.

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP refers to true positive (the number of DNA/RNA-binding proteins with CO + PD score > cut-off), FP refers to false positive (the number of other proteins with CO + PD score > cut-off), FN refers to false negative (the number of DNA/RNA-binding proteins with CO + PD score < cut-off), and TN refers to true negative (the number of other proteins with CO + PD score < cut-off).

2.6. *Construction of expression vectors and purification of His-tagged recombinant proteins*

Genomic DNA of *P. furiosus* DSM3638 was isolated using a GNOME kit (BIO101, La Jolla, CA, USA) and partially digested with the restriction enzyme *Sau*3AI. The resulting DNA fragments were fractionated by electrophoresis on a 0.7% agarose gel. Fragments of 15 kb were extracted from the gel and used as templates for PCR cloning. After PCR amplification using site-specific primers (Supplementary Table S4) with *Nde*I and *Xho*I sites at the 5′ and 3′ termini, respectively, each of the candidate genes was cloned into the pET-23b expression vector (Novagen, Madison, WI, USA). Insert DNA was sequenced and shown to be identical to database sequences.

Recombinant proteins were prepared as described previously.[8] Briefly, *E. coli* strain BL21(DE3) was transformed with each expression plasmid; however, optimal protein production required *E. coli* strain BL21(DE3)pLysS

for the expression of PF0565 and PF1473 proteins and strain HMS174(DE3)pLysS for the expression of PF1498. Transformants were grown at 37°C in Luria–Bertani medium containing 50 μg/mL ampicillin and supplemented with 0.4 mM isopropylthio-b-galactoside. After 14–16 h of further growth at 30°C, cells were harvested by centrifugation (5000*g* for 10 min at 4°C), and the recombinant proteins were released by sonication (2 min) in buffer A (20 mM Tris–HCl, pH 8.0, 5 mM imidazole, 500 mM NaCl, 0.1% NP40). The extracts were heat-treated at 85°C for 15 min to destroy *E. coli* endogenous proteins and then centrifuged at 12000*g* for 10 min at 4°C to remove cellular debris. The recombinant proteins were purified in an Ni^{2+}–Sepharose column according to the manufacturer's instructions (Amersham Pharmacia, Piscataway, NJ, USA). The peaks of the eluted proteins were pooled and dialyzed against buffer B (50 mM Tris–HCl, pH 8.0, 1 mM EDTA, 0.02% Tween 20, 7 mM 2-mercaptoethanol, 10% glycerol).

2.7. Gel-shift assay

Hokkaido System Science Co. (Hokkaido, Japan) chemically synthesized 5′ end FAM-labeled oligonucleotides. Binding reactions containing the oligonucleotide (125 or 500 nM) and 0.1–0.5 μg of purified recombinant protein were incubated for 15 min at either room temperature (24°C) or 75°C in 20 μL of DNA/RNA-binding buffer (10 mM Tris–HCl, pH 7.5, 50 mM NaCl, 0.5 mM EDTA, 2.5 mM $MgCl_2$, 5% glycerol, 1 mM dithiothreitol). The DNA/RNA-protein complexes were analyzed by 6% non-denaturing PAGE. The quantity of DNA/RNA-protein complexes was evaluated by scanning the fluorescent image with a computerized image analyzer, FX Pro (Bio-Rad Laboratories, Hercules, CA, USA). To sequence the oligonucleotides, we used the following two probes (Xiaojing et al., to be published separately):

(i) MPOR-27, 5′-r(GAAACAAGGAGAAAUGGUUCG UGUCCU)-3′,
(ii) MPOD-27, 5′-d(GAAACAAGGAGAAATGGTTCG TGTCCT)-3′.

3. Results and discussion

3.1. Functional annotation of P. furiosus *proteome and those of other model species*

P. furiosus, *S. solfataricus*, *B. subtilis*, *E. coli*, *C. elegans*, and *A. thaliana* were used as model species. The hyperthermophilic archaeon *P. furiosus* was chosen for its topical importance in the evolution of the ancient architecture of DNA/RNA regulation[31] as well as for the thermal stability of its proteins, which enables easy generic purification. In addition, many *P. furiosus* protein functions remain unknown, which further justifies their study.

From the EMBL database (Release 83, June 2005), we extracted reliable protein function data for *P. furiosus* (EMBL accession number AE009950) by unifying information from the three annotated databases Swiss-Prot, TrEMBL, and GOA.[32,33] We defined the three categories of proteins on the basis of the number and quality of annotations (see Methods and materials section). For example, 2057 *P. furiosus* proteins were categorized into 942 functionally known proteins, 121 proteins with putative function, and 994 hypothetical proteins. To eliminate proteins with similar amino acid sequences, we performed a homology search among the 942 functionally known proteins using BLASTP (E-value $< 1 \times 10^{-4}$) and reduced the protein data set to 477 non-redundant proteins for the periodicity analysis. To facilitate their use as a training data set for SVM learning, these functionally known proteins were further divided into 157 DNA/RNA-binding proteins and 320 proteins with other functions. The same procedure was applied to the EMBL data of the archaeon *S. solfataricus* (EMBL accession number AE006641) and the Swiss-Prot entries of the *B. subtilis*, *E. coli*, *A. thaliana*, and *C. elegans* proteomes (Table 1).

3.2. Amino acid periodicity score (PD score) and prediction of the DNA/RNA-binding proteins

To ascertain common features of amino acid periodicity throughout the DNA/RNA-binding protein sequences, we defined 23 amino acid groups using eight physicochemical profiles (chemical, Sneath, Dayhoff, Stanfel, functional, charge, structural, and hydrophobicity). We prepared a total of 253 patterns of amino acid periodicities (23 groups × 11 non-redundant periodicities). For each training data set in the six model species, the relative coverage of periodic region R was calculated for 253 individual amino acid periodicities as feature vectors for SVM input. Radial basis function SVM classification was performed with default parameters using the software Gist, which allows users to apply a sophisticated machine-learning algorithm to the data.[23] To quantitatively evaluate a DNA/RNA-binding protein at the proteome level, the discriminant value derived by SVM was defined as a novel index, the periodicity score (PD score), and was assigned to the representative protein data set of each of the six model species.

The performance of the PD score as a DNA/RNA-binding protein classifier was evaluated by applying the ROC curve to the representative set of *P. furiosus* proteins (Fig. 1). Sensitivity and specificity of the PD score overwhelmed that of various individual amino acids periodicities such as RK7, CDEGHKNQRSTY8, MNQ19, and AFILMPVW11. This demonstrated that a combination of amino acid periodicities as a feature vector optimizes the system for the classification of DNA/RNA-binding proteins.

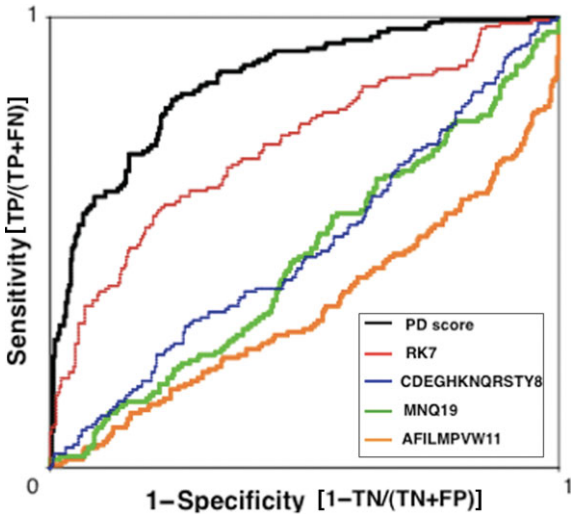


Figure 1. Performance of PD score. ROC curves of PD score (black) and single amino acid periodicities [RK7 (red), CDEGHKNQRSTY8 (blue), MNQ19 (green), and AFILMPVW11 (orange)]. For example, RK7 denotes the frequency of the periodic region of Arginine (R) and Lysine (K) appearing in the protein sequence with a periodicity of 7.

To further validate the performance of PD score, we conducted a comparative analysis upon amino acid composition and SVM-Prot.[27] Amino acid composition is a widely used profile for predicting protein function, subcellular localization, and protein folding. We calculated 20 individual amino acid compositions as feature vectors for SVM input and defined a new indicator named as composition (CO) score. SVM-Prot is a general proteome-wide function-prediction software based on various features of primary sequences. Three indices, sensitivity (SE), specificity (SP), and overall accuracy (ACC), were calculated for the six model species, respectively, using 10-fold cross-validation to calculate the precise prediction efficiency (Table 2). The three predictors have shown different characteristics in predicting DNA/RNA-binding proteins owing to the three indices. The PD score possessed highest overall accuracy; CO score had the highest overall sensitivity and SVM-Prot had the highest overall specificity. As a result, the prediction performance of the PD score was comparable with other methods but statistical significances cannot be observed on the basis of the comparative analysis. Thus, we combined the two indicators CO score and PD score to improve our DNA/RNA-binding prediction method.

3.3. Both CO score and PD score are required for efficient classification of DNA/RNA-binding protein predictions

For the efficient classification of DNA/RNA-binding proteins, we performed three different methods (early integration, intermediate integration, and late integration) for integrating heterogeneous data sets on the basis of the context of SVM learning.[29] Late integration has shown the best performance for sensitivity, whereas intermediate integration has shown the best performance for specificity and accuracy (data not shown). Our observation was not consistent with the previous work by Pavlidis,[29] in which early and intermediate integrations have shown high performance compared with the late integration. It is possible that the prediction performance can be altered by many factors such as evaluation indices, data types, kernel function parameters, and feature-selection algorithms.

To provide further insights into the relative performance of CO and PD scores and to extract efficient DNA/RNA-binding protein candidates, we have chosen the late integration method and carried out two-dimensional correlation analysis on the basis of CO score and PD score upon 477 functionally known proteins in *P. furiosus*. The correlation coefficient was $r = 0.75$ (all functionally known proteins) and $r = 0.55$ (DNA/RNA-binding proteins only), respectively. The 157 DNA/RNA-binding proteins located at the right-upper region of the two-dimensional plot, suggesting that both CO and PD scores are required for classifying proteins with DNA or RNA-binding activity (red and blue circles in Fig. 2A). Then two different thresholds were

Table 2. Prediction performance of PD score compared with CO score and SVM-Prot in the six model species

Species	Sample size	PD score			CO score			SVM-Prot		
		SE (%)	SP (%)	ACC (%)	SE (%)	SP (%)	ACC (%)	SE (%)	SP (%)	ACC (%)
P. furiosus	477	72.7 (10.5)	81.1 (6.6)	**78.1 (5.8)**	72.8 (9.8)	80.8 (9.9)	77.9 (6.9)	**73.1**	**84.9**	72.3
S. solfataricus	582	**68.0 (11.9)**	79.8 (6.8)	76.0 (5.4)	67.7 (18.4)	**84.1 (8.3)**	**78.8 (6.7)**	66.3	**84.1**	77.6
B. subtilis	914	58.5 (11.1)	81.2 (5.5)	**75.9 (4.6)**	**75.5 (13.8)**	72.3 (7.4)	73.1 (6.1)	63.7	**87.9**	75.3
E. coli	1436	58.2 (6.0)	83.4 (2.9)	77.0 (2.3)	**77.7 (8.9)**	69.6 (6.7)	71.7 (4.0)	57.1	**86.5**	**81.3**
A. thaliana	865	63.3 (8.7)	**87.5 (5.8)**	**81.5 (3.7)**	**69.5 (8.7)**	84.9 (3.7)	80.9 (3.0)	66.8	**87.5**	79.1
C. elegans	566	59.3 (8.1)	**84.0 (3.6)**	**77.1 (2.2)**	63.2 (12.1)	81.5 (5.8)	75.6 (4.9)	**65.1**	83.6	72.0
Overall	—	63.3	82.9	**77.6**	**71.1**	78.9	76.3	65.3	**85.8**	76.3

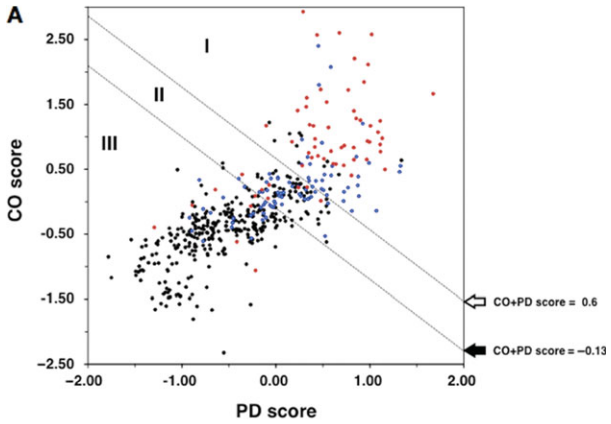
Predicted results are shown as SE (sensitivity) = TP/(TP + FN), SP (specificity) = TN/(TN + FP), and ACC (accuracy) = (TP + TN)/(TP + FN + TN + FP). Numbers in bold font indicate the highest index among the three classifiers.

determined on the basis of this newly defined CO score + PD score. First threshold is based on the highest overall accuracy (ACC) with a CO + PD score of 0.6 and the second threshold is based on the highest MCC with CO + PD score = −0.13. According to the Supplementary Fig. S1, the first threshold optimizes the extraction of reliable candidates for novel DNA/RNA-binding proteins (SE = 52.2%, SP = 96.3%, ACC = 81.8%, and PPV = 87.2%) and the second threshold optimizes the classification performance of CO + PD score (SE = 82.2%, SP = 80%, ACC= 80.7%, and PPV = 66.8%). On the basis of these thresholds, we classified proteins into three classes (class I–III) (Fig. 2A). As a result, a total of 94 proteins including 82 DNA/RNA-binding proteins (Fig. 2B) were categorized as class I proteins (CO + PD score > 0.6).

Further observation of DNA/RNA-binding proteins has revealed a region-specific distribution of ribosomal proteins and other DNA/RNA-binding proteins. Ribosomal proteins are strongly affected by CO score and are dominant at the high range of CO score (CO > 0.5). The CO score of other DNA/RNA-binding proteins ranged between 0 and 0.5 but some of them were dominant at high PD score region (0.25–1.5). As shown in Supplementary Table S1, this region includes 13 tRNA-processing enzymes (i.e. tRNA-synthetases, CCA-adding enzymes, and RNase P subunits), 11 DNA-binding proteins (i.e. DNA polymerase, DNA helicase, DNA primase, and reverse gyrase), three ribosomal proteins (i.e. ribosomal protein S3P and ribosomal protein L14e), and various transcription/translation-related proteins (i.e. SRP54, HTH-type transcriptional regulator, and transcription termination–anti-termination factor). We assume that PD score is an effective means of classifying DNA/RNA-binding proteins from a set of proteins, which cannot be distinguished by using amino acid compositions.

3.4. *Selection and experimental verification of novel DNA/RNA-binding protein candidates*

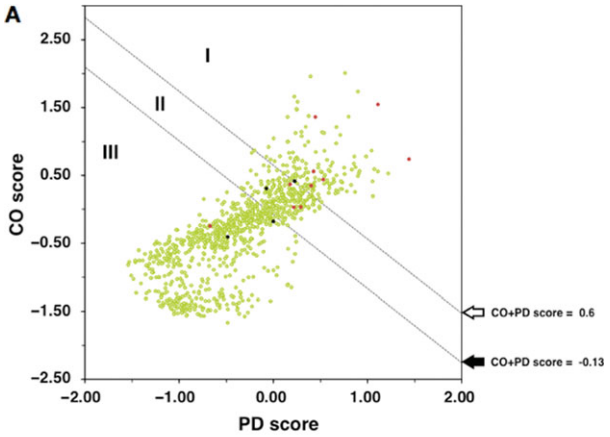
The same procedure was applied to the 994 hypothetical proteins in *P. furiosus*. The two-dimensional plot of hypothetical proteins was similar to that of functionally known proteins as well as the protein ratio in classes I–III (Fig. 2 versus Fig. 3). However, the number of proteins has decreased from the high CO score (CO score > 0.5) region, which is known to be dominated by ribosomal proteins. As a result, 994 hypothetical proteins



B

	Number of proteins			
	●	○	●	Total (%)
class I	48	34	12	94 (19.7)
class II	8	39	52	99 (20.8)
class III	6	23	256	284 (59.5)

Figure 2. Two-dimensional correlation analysis of DNA/RNA-binding proteins in *P. furiosus* on the basis of amino acid composition and periodicity. (**A**) A total of 477 functionally known proteins in *P. furiosus* were plotted on the two-dimensional correlation plot of CO score and PD score. The ribosomal proteins (red), rest of the DNA/RNA-binding proteins (blue), and other functionally known proteins (black) are shown. The two dotted lines represent a threshold of maximum MCC value of 0.59 for optimizing sensitivity and specificity (black arrow) and a maximum ACC value of 81.8% for optimizing the prediction of DNA/RNA-binding protein candidates (white arrow). The ranges of three classes are Class I, CO + PD score > 0.6; class II, 0.6 > CO + PD score > −0.13; and class III −0.13 > CO + PD score. **(B)** The numbers of ribosomal proteins (red), rest of the DNA/RNA-binding proteins (blue), and other functionally known proteins (black) are counted in classes I–III.



B

	Number of proteins			
	●	●	○	Total (%)
class I	6	1	144	151 (15.2)
class II	3	1	205	209 (21.0)
class III	1	2	631	634 (63.8)

Figure 3. Two-dimensional correlation analysis of hypothetical proteins in *P. furiosus* on the basis of amino acid composition and periodicity. **(A)** The vertical and horizontal axes represent CO and periodicity PD scores, respectively. Distribution of the 994 hypothetical proteins is shown. Experimentally validated 14 candidate proteins are denoted as red circles (possessed DNA/RNA-binding activity) and black circles (no detectable DNA/RNA-binding activity) and the remaining proteins are shown as green circles. **(B)** The numbers of experimentally verified proteins with DNA/RNA-binding activities (red), protein with no DNA/RNA-binding activities (black), and the remaining hypothetical proteins (green) are counted for classes I–III.

were classified into three classes (I–III) owing to the CO + PD score thresholds (Supplementary Table S2), and a total of 151 proteins were classified as strong candidates for novel DNA/RNA-binding proteins.

In order to verify that hypothetical proteins in class I actually possess DNA/RNA-binding protein activities, we randomly chose 17 hypothetical proteins from classes I–III (nine from class I, five from class II, and three from class III, Table 3). All 17 recombinant proteins were overexpressed in *E. coli* and purified to near homogeneity (Fig. 4). To study the DNA/RNA-binding properties of the candidate proteins, we first carried out gel-shift assays using 5′ FAM-labeled, 27 bp, multipotential oligoprobe RNA (MPOR-27) (Fig. 5A). MPORs potentially possess four different secondary RNA structures (stem, bulge, loop, and single strand), which encompass the currently known structures corresponding to the activities of various RNA-binding proteins. The three proteins, PF0871, PF0678, and PF0840, aggregated in the loading well, so we removed them from the final results. A prominent shift of the RNA probe up the gel was observed in candidate proteins PF0029, PF0030, PF0565, PF1139, PF1473, PF1580, PF1912, and PF2062 (Fig. 5B). Interestingly, the formation of certain nucleic acid–protein complexes appears to be temperature dependent. For example, PF1981 showed a significant shift at 75°C but not at 24°C (Fig. 5C versus 5B). PF0029, PF0030, PF1139, and PF1580 also showed binding affinity with the multipotential oligoprobe DNA, MPOD-27 (data not shown). No significant shifts were observed in PF0547, PF1142, PF1488, PF1498, or PF1913, though agarose gel analysis of purified PF1498 revealed it to be a potential protein–nucleic acid complex (Fig. 5D).

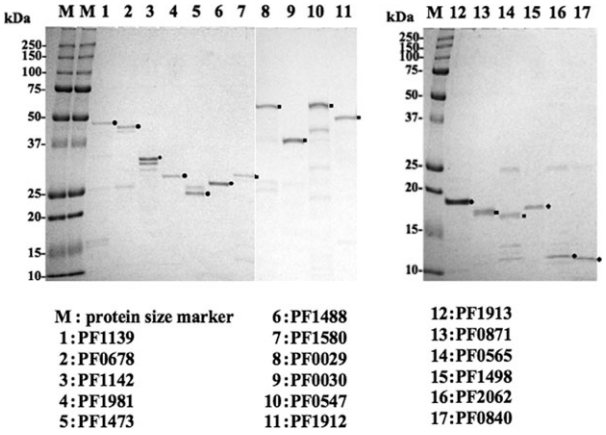


Figure 4. Purification of the hypothetical proteins in *P. furiosus*. SDS–PAGE analysis of 17 purified candidate proteins. The positions of the purified candidate proteins are marked with black dots. For proteins of large and small molecular sizes, 10–20% (left column) and 15–25% (right column) gradient gel was used (M, protein size marker; Bio-Rad).

Table 3 Summary of experimentally validated hypothetical proteins in *P. furiosus*

Class	Gene ID	Molecular mass (kDa)	SVM analysis			Experimenal verification of DNA/RNA-binding activity
			PD score	CO score	CO + PD	
I	PF1498	16.5	1.11	1.55	2.66	+[a]
I	PF1139	44.7	1.44	0.74	2.18	+
I	PF2062	11.0	0.45	1.36	1.81	+
I	PF0565	17.2	0.43	0.56	0.99	+
I	PF1473	26.7	0.53	0.44	0.97	+
I	PF1580	25.5	0.40	0.35	0.75	+
I	PF1913	18.5	0.23	0.42	0.65	−
II	PF1981	27.5	0.18	0.37	0.55	+
II	PF1912	48.2	0.29	0.04	0.33	+
II	PF0029	56.4	0.22	0.03	0.25	+
II	PF1488	26.5	−0.07	0.31	0.24	−
III	PF0547	50.7	0.00	−0.18	−0.18	−
III	PF1142	31.5	−0.48	−0.41	−0.89	−
III	PF0030	40.8	−0.67	−0.25	−0.92	+

The DNA/RNA-binding activities were examined by gel-shift assay at room temperature (22°C) and 75°C.

[a]The RNA-binding activities of PF1498 was examined by 1.2% agarose gel electrophoresis and ethidium bromide staining owing to the co-purification with endogenous nucleic acid in *E. coli* (Fig. 5D).

During our investigation, six out of seven class I proteins, three out of four class II proteins, and one class III protein have shown potential DNA/RNA-binding activities (Table 3). According to our previous works,[8,9] the systematic screening of *P. furiosus* genome using the expression cloning method has determined several DNA/RNA-binding proteins such as Thy-1 and FAU-1. Our system also demonstrated that approximately 10–20% of the *P. furiosus* gene products have shown to possess nucleic acid-binding activity. However, 80–90% of the *P. furiosus* gene products did not show any affinity to the nucleic acid in our gel-shift system. These results suggest that the 10 newly discovered proteins through the experimental procedure are good candidates as novel DNA/RNA-binding proteins, although their binding specificity remains unknown. The in vivo targets and precise biological functions of these proteins are to be further investigated.

According to the domain assignment of the InterPro/Pfam domain database[20,21] against *P. furiosus* proteome, 95–98% of the 942 functionally known proteins possessed domains related to those with known function (functional domains). On the other hand, for 994 hypothetical proteins, only 31–38% of the proteins possessed functional domains, 20% possessed domains of unknown function (DUF/UPF), and the remaining 43–50% lacked domain

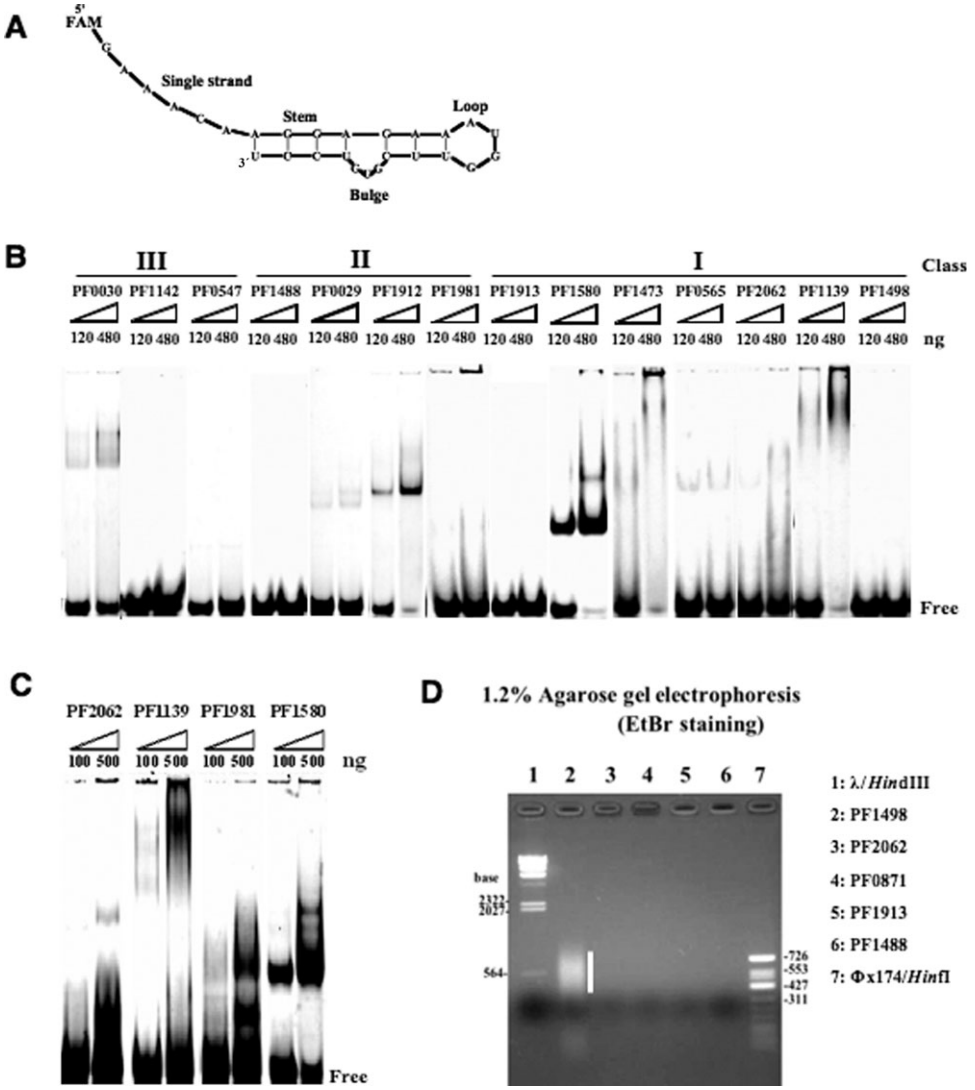


Figure 5. RNA-binding analysis of 14 hypothetical proteins in *P. furiosus*. **(A)** Nucleotide sequence and possible RNA secondary structure of multipotential oligoprobe RNA. (**B**) Detection of RNA-binding activity of 14 candidate proteins by gel-shift assay at room temperature (24°C). (**C**) Gel-shift assay of four candidate proteins with prominent RNA-binding activity at 75°C. (**D**) Analysis of purified protein peak fractions by 1.2% agarose gel. White bar indicates the existence of nucleic acids. Lanes 1 and 7 are DNA markers.

annotation (Supplementary Fig. S2). According to Supplementary Table 3, among the newly discovered 10 DNA/RNA-binding proteins, at least four ORFans are detected (PF0029, PF0030, PF0565, and PF1981), which completely lacked sequence similarity (E-value $>$ 0.1) compared with any of the Swiss-Prot protein entries. The remaining six proteins have shown sequential similarity to the uncharacterized proteins of nearest BLASTP hit ($8.00e{-}07 >$ E-value > 0.0), which were conserved among *Pyrococcus* and *Methanococcus*, including two proteins (PF1473 and PF2062) with no Pfam domain annotation. Hence, we believe that the combination of CO score and PD score is a powerful indicator for predicting proteins with potential DNA/RNA-binding activities from the sequence-specific ORFans and a set of proteins having no obvious functional domains, although allowing that the sample size of validated proteins is still small.

3.5. Possible explanation of charged amino acid periodicity with DNA/RNA-binding activities

As the amino acid composition within proteins varies among taxa,[34] our method removes the need to allow for the evolutionary gain and loss of amino acids and increases the generalization capability of SVM training. Charged amino acids, especially basic amino acids, have previously been suggested as a key component of nucleic acid-binding activity; for example, arginine-rich regions of the *Drosophila melanogaster* suppressor of sable gene[35] are thought to mediate specific RNA-binding activity. Similar features have been observed in the structural

motifs of DNA/RNA-binding proteins that possess positive electrostatic potentials in the binding region.[36] On the basis of electrostatic potential, negatively charged amino acids (DE) conflict with DNA/RNA-binding. However, recent work has revealed that negative peptide charges contribute significantly to the electrostatic free energy of positively charged peptides and affect RNA binding,[37] suggesting the importance of not only basic regions but also, in some cases, acidic regions at the protein surface, for establishing DNA/RNA-binding functions. These results suggest that relative compositions of charged amino acids, especially basic amino acids, are very important for nucleic acid binding. In addition, our study has shown that certain class of DNA/RNA-binding proteins were efficiently classified by integrating amino acid periodicity with amino acid composition. Especially, charged amino acid periodicities have been observed throughout the protein sequence of various DNA/RNA-binding proteins, suggesting that not only amino acid compositions in the DNA/RNA-binding domain region but also the overall sequence feature of amino acid periodicity is useful for classifying DNA/RNA-binding proteins.

To gain insight into the relationship between charged amino acid periodicities and DNA/RNA-binding activity, a schematic representation of charged amino acid groups that appear periodically in the amino acid sequence of various DNA/RNA-binding proteins is given in Supplementary Fig. S3. The three proteins, signal recognition particle of 54kDa subunit (SRP54), DNA primase, and HTH-type transcriptional regulator lrpA, were chosen as an example for their characteristic features of possessing low CO score ($-0.03 <$ CO score < 0.22) but relatively high PD score ($0.56 <$ PD score < 0.88). An amino acid periodicity of both positively and negatively charged amino acids with various periodicities were widely found through protein primary sequence. The amino acid residues creating the periodicity (oblong boxes in Supplementary Fig. S3) are often conserved in the three-dimensional structures of orthologous proteins. Periodic region also covers DNA/RNA-recognition motif known as M domains and helix-turn-helix and active site of DNA primase. Our current study has shown that overall periodic features of charged amino acids throughout the protein primary structure may affect the organization of the secondary structures or the net charge of the protein surface in the tertiary structure in certain class of DNA/RNA-binding proteins. Further detailed analysis of the relationship between DNA/RNA-binding capacity and specific amino acid periodicity will be an important task with the help of other bioinformatics approaches such as the use of DNA/RNA-binding site prediction software,[38] a comparative genomics approach that predicts function on the basis of the comparison of various domains,[39] and three-dimensional protein models.[40]

In conclusion, we have presented a new method for predicting novel DNA/RNA-binding proteins at the proteome level by focusing on compositions and periodicities of amino acids with similar physico-chemical profiles (quantified as a novel index denoted as CO score and PD score). The two-dimensional correlation analysis of CO score and PD score effectively separated DNA/RNA-binding proteins from other functionally known proteins in *P. furiosus* as class I proteins. By applying the same method to the 994 hypothetical proteins, we extracted a list of 151 hypothetical proteins as novel DNA/RNA-binding protein candidates. Ten proteins with potential DNA/RNA-binding activities were determined experimentally, including four ORFans and two proteins with no domains. The two-dimensional correlation analysis of CO score and PD score is applicable to any organisms with complete genomic data. To conclude, our method is highly efficient for evaluating hypothetical proteins on the basis of DNA/RNA-binding function. The CO + PD scores can be further integrated with prediction results from various protein function predictors and annotation methods to validate uncharacterized proteins comprehensively. Further, the investigation of these newly discovered DNA/RNA-binding proteins might elucidate the role of undiscovered protein–DNA/RNA networks and the recognition of many non-conserved proteins throughout entire species.

Acknowledgements: We thank Asako Sato (Keio University, Japan) for technical assistance with the gel-shift assay. We also thank Jun Imoto, Nozomu Yachie, Shinichi Kikuchi, and Rintaro Saito (Keio University, Japan) for their helpful discussions. This research was supported in part by the Project for Development of a Technological Infrastructure for Industrial Bioprocesses in Research and Development of New Industrial Science and Technology Frontiers, the Ministry of Economy, Trade and Industry (METI), the New Energy and Industrial Technology Development Organization (NEDO) of Japan; a Grant-in-Aid for Scientific Research on Priority Areas; a Grant-in-Aid for the 21st Century Center of Excellence (COE) Program entitled 'Understanding and Control of Life's Function via Systems Biology' (Keio University); the Computer Simulation Project, Ministry of Education, Culture, Sport, Science and Technology, Japan; and Keio University.

Supplementary Data: Supplementary data are available online at http://dnaresearch.oxfordjournals.org.

References

1. Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2005, NCBI Reference Sequence RefSeq: a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, **33**, D501–504.

2. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
3. Pazos, F. and Sternberg, M. J. 2004, Automated prediction of protein function and detection of functional sites from structure, *Proc. Natl Acad. Sci. USA.*, **101**, 14754–14759.
4. McLaughlin, W. A., Kulp, D. W., de la Cruz, J., Lu, X. J., Lawson, C. L. and Berman, H. M. 2004, A structure-based method for identifying DNA-binding proteins and their sites of DNA-interaction, *J. Struct. Funct. Genomics.*, **5**, 255–265.
5. Date, S. V. and Marcotte, E. M. 2003, Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages, *Nat. Biotechnol.*, **21**, 1055–1062.
6. Amiri, H., Davids, W. and Andersson, S. G. 2003, Birth and death of orphan genes in Rickettsia, *Mol. Biol. Evol.*, **20**, 1575–1587.
7. Siew, N. and Fischer, D. 2004, Structural biology sheds light on the puzzle of genomic ORFans, *J. Mol. Biol.*, **342**, 369–373.
8. Kanai, A., Oida, H., Matsuura, N. and Doi, H. 2003, Expression cloning and characterization of a novel gene that encodes the RNA-binding protein FAU-1 from *Pyrococcus furiosus*, *Biochem. J.*, **372**, 253–261.
9. Kanai, A., Sato, A., Imoto, J. and Tomita, M. 2006, Archaeal *Pyrococcus furiosus* thymidylate synthase 1 is an RNA-binding protein, *Biochem. J.*, **393**, 373–379.
10. Sato, A., Kanai, A., Itaya, M. and Tomita, M. 2003, Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon, *Pyrococcus furiosus*, *Biochem. Biophys. Res. Commun.*, **309**, 247–252.
11. Cotton, J. L. and Mykles, D. L. 1993, Cloning of a crustacean myosin heavy chain isoform: exclusive expression in fast muscle, *J. Exp. Zool.*, **267**, 578–586.
12. Laskin, A. A., Kudryashov, N. A., Skryabin, K. G. and Korotkov, E. V. 2005, Latent periodicity of serine–threonine and tyrosine protein kinases and other protein families, *Comput. Biol. Chem.*, **29**, 229–243.
13. Bhardwaj, N., Langlois, R. E., Zhao, G. and Lu, H. 2005, Kernel-based machine learning protocol for predicting DNA-binding proteins, *Nucleic Acids Res.*, **33**, 6486–6493.
14. Han, L. Y., Cai, C. Z., Lo, S. L., Chung, M. C. and Chen, Y. Z. 2004, Prediction of RNA-binding proteins from primary sequence by a support vector machine approach, *RNA*, **10**, 355–368.
15. Cai, Y. D. and Lin, S. L. 2003, Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence, *Biochim. Biophys. Acta.*, **1648**, 127–133.
16. Yu, X., Cao, J., Cai, Y., Shi, T. and Li, Y. 2006, Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines, *J. Theor. Biol.*, **240**, 175–184.
17. Ofran, Y. and Margalit, H. 2006, Proteins of the same fold and unrelated sequences have similar amino acid composition, *Proteins*, **64**, 275–279.
18. Xie, D., Li, A., Wang, M., Fan, Z. and Feng, H. 2005, LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST, *Nucleic Acids Res.*, **33**, W105–W110.
19. Sarda, D., Chua, G. H., Li, K. B. and Krishnan, A. 2005, pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties, *BMC Bioinformatics*, **6**, 152.
20. Apweiler, R., Attwood, T. K., Bairoch, A., et al.. 2000, InterPro—an integrated documentation resource for protein families, domains and functional sites, *Bioinformatics*, **16**, 1145–1150.
21. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. 2000, The Pfam protein families database, *Nucleic Acids Res.*, **28**, 263–266.
22. Gatherer, D. and Mc Ewan, N. R. 2003, Analysis of sequence periodicity in *E. coli* proteins: empirical investigation of the 'duplication and divergence' theory of protein evolution, *J. Mol. Evol.*, **57**, 149–158.
23. Pavlidis, P., Wapinski, I. and Noble, W. S. 2004, Support vector machine classification on the web, *Bioinformatics*, **20**, 586–587.
24. Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J. and Zhang, B. T. 2006, miTarget: microRNA target gene prediction using a support vector machine, *BMC Bioinformatics*, **7**, 411.
25. Yu, C., Zavaljevski, N., Stevens, F. J., Yackovich, K. and Reifman, J. 2005, Classifying noisy protein sequence data: a case study of immunoglobulin light chains, *Bioinformatics*, **21**, 495–501.
26. Goldbaum, M. H., Sample, P. A., Chan, K., et al. 2002, Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry, *Invest. Ophthalmol. Vis. Sci.*, **43**, 162–169.
27. Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X. and Chen, Y. Z. 2003, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Res.*, **31**, 3692–3697.
28. Han, L. Y., Cai, C. Z., Ji, Z. L., Cao, Z. W., Cui, J. and Chen, Y. Z. 2004, Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach, *Nucleic Acids Res.*, **32**, 6437–6444.
29. Pavlidis, P., Weston, J., Cai, J. and Noble, W. S. 2002, Learning gene functional classifications from multiple data types, *J. Comput. Biol.*, **2**, 401–411.
30. Matthews, B. W. 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta.*, **405**, 442–451.
31. Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B. and Dunn, D. M. 2001, Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology, *Meth. Enzymol.*, **330**, 134–157.
32. Bairoch, A. and Apweiler, R. 1996, The SWISS-PROT protein sequence data bank and its new supplement TREMBL, *Nucleic Acids Res.*, **24**, 21–25.
33. Camon, E., Magrane, M., Barrell, D., et al. 2003, The Gene Ontology Annotation, GOA project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro, *Genome Res.*, **13**, 662–672.
34. Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., et al. 2005, A universal trend of amino acid gain and loss in protein evolution, *Nature*, **433**, 633–638.

35. Turnage, M. A., Brewer-Jensen, P., Bai, W. L. and Searles, L. L. 2000, Arginine-rich regions mediate the RNA binding and regulatory activities of the protein encoded by the *Drosophila melanogaster* suppressor of sable gene, *Mol. Cell. Biol.*, **20**, 8198–8208.
36. Shanahan, H. P., Garcia, M. A., Jones, S. and Thornton, J. M. 2004, Identifying DNA-binding proteins using structural motifs and the electrostatic potential, *Nucleic Acids Res.*, **32**, 4732–4741.
37. Garcia-Garcia, C. and Draper, D. E. 2003, Electrostatic interactions in a peptide–RNA complex, *J. Mol. Biol.*, **331**, 75–88.
38. Ahmad, S. and Sarai, A. 2005, PSSM-based prediction of DNA binding sites in proteins, *BMC Bioinformatics*, **6**, 33.
39. Anantharaman, V., Koonin, E. V. and Aravind, L. 2002, Comparative genomics and evolution of proteins involved in RNA metabolism, *Nucleic Acids Res.*, **30**, 1427–1464.
40. Manival, X., Ghisolfi-Nieto, L., Joseph, G., Bouvet, P. and Erard, M. 2001, RNA-binding strategies common to cold-shock domain- and RNA recognition motif-containing proteins, *Nucleic Acids Res.*, **29**, 2223–2233.