Title	Prediction of novel DNA/RNA-binding proteins and analysis of tRNA evolution using genomic data
Sub Title	ゲノム情報を用いたDNA/RNA結合蛋白質の予測とtRNAの進化に関する解析
Author	藤島, 皓介(Fijishima, Kosuke) 冨田, 勝(Tomita, Masaru)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2007-03
Jtitle	優秀修士論文
JaLC DOI	
Abstract	ポストゲノム時代を迎えた今、蛋白質と核酸がおりなす複雑な相互作用が一連の遺伝情報の発現系（セントラルドグマ）にどのように寄与しているかを知ることは、生命システムの起源を解き明かす意味でも非常に重要な問題になっている。本研究では古細菌という太古の地球環境に類似していると想像される環境に生育する菌類をモデル生物とし、生命システムの原型を蛋白質及びtRNAという分子から解き明かすことを目的とした。
Notes	先端生命科学プロジェクト2006年
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=0302-0000-0590

ISBN 978-4-87762-175-9
SFC-MT 2006-005

優秀修士論文

Prediction of novel DNA/RNA-binding proteins and analysis of tRNA evolution using genomic data

2006年

藤島 皓介　政策・メディア研究科 修士課程

先端生命科学プロジェクト

慶應義塾大学湘南藤沢学会

Keio SFC Academic Society

優秀修士論文推薦のことば

原始生命に近い古細菌を研究することは、生命システムの起源を知る上で非常に重要である。本論文ではアミノ酸の周期性という独自の要素を定義し、古細菌において、数多くの新規核酸結合蛋白質を予測しさらに同定した。また網羅的な tRNA 配列の比較解析を通して得られた仮説は、遺伝暗号表の起源を明らかにする可能性がある。このように生命情報学と分子生物学を組み合わせたハイスループットな研究手法は SFC が目指す分野融合と最先端の研究を体現していることから、本論文を修士優秀論文として強く推薦する。

慶應義塾大学
環境情報学部教授
冨田　勝

修士論文　　2006 年度 (平成 18 年度)

Prediction of novel DNA/RNA-binding proteins and analysis of tRNA evolution using genomic data

ゲノム情報を用いた DNA/RNA 結合蛋白質の予測と tRNA の進化に関する解析

慶應義塾大学大学院　政策・メディア研究科

藤島　皓介

Prediction of novel DNA/RNA-binding proteins and analysis of tRNA evolution using genomic data

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 (CO) score and Periodicity (PD) score. The *P. furiosus* proteins were classified into three classes (I-III) based on the 2D 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 2D 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. Five out of six candidate proteins in class I possessed DNA/RNA-binding activities, supporting the efficacy of our method.

Recent discovery of the completely separated 5' and 3' halves of tRNA molecules; so-called split-tRNA in Nanoarchaeota *Nanoarchaeum equitans* has brought us question whether ancient form of tRNA was codified on single or separated genes. We propose a new theory that tRNAs are originated from the combination of 5' half and 3' half tRNA fragments in the ancestral archaeons. To verify this hypothesis, we prepared total 1302 tRNA sequences from 30 archaeal genomes based on computational prediction and performed sequence alignment of the exon sequences to observe the relativity of split- intronic- and nonintronic tRNAs at the sequence level. As a result, exon tRNA sequences were clearly separated into 39 different clusters resulting each of the 6 split-tRNAs classified among the intronic and non-intronic tRNAs with same anticodon. Further, we divided 304 tRNAs in 7 representative archaeal species from different genus at the canonical intron position (37/38) to mimic split-tRNA and performed comparative phylogenetics upon 5' and 3' tRNA halves respectively. The topology of the phylogenetic trees of 5' and 3' tRNA halves differed significantly and the consensus sequence of each cluster has shown potential identity determinants suggesting the different evolutionary background. The combination patterns of 5' and 3' tRNA halves strongly correlated with the divergence of amino acids in the codon table. These results suggest that ancient tRNAs could have been emerged through combination of various 5' and 3' tRNA sequence to

establish the genetic code.

Keywords: functional prediction, DNA/RNA-binding protein, amino acid periodicity, Support vector machine, tRNA evolution, phylogenetic analysis

Keio University, Graduate School of Media and Governance
Kosuke Fujishima

ゲノム情報を用いた DNA/RNA 結合蛋白質の予測と tRNA の進化に関する解析

論文要旨

ポストゲノム時代を迎えた今,蛋白質と核酸がおりなす複雑な相互作用が遺伝情報の発現を中心とする生命システムにどのように寄与しているかを知る事は非常に重要な問題となっている．本研究では古細菌における蛋白質及び RNA に着目し、その機能や進化的な側面から先の問題の一端を解明する目的で解析を行った．現在、プロテオームの分野では実験的に確認されたものも含めても機能既知蛋白質の割合は全体の約 5 割程度にとどまっている．一方これまでの先行研究から核酸結合蛋白質の多くは特定のアミノ酸を周期的に有することが確認されていることから，私はアミノ酸の周期性という指標を用いて新規の核酸結合蛋白質の予測方法の確立を目指した．モデル生物には超好熱性古細菌 *P. furiosus* を用いた．その理由は遺伝子数が 2000 程度で進化的に生命の起源に近い生物であること，さらに蛋白質が熱耐性であり精製が容易で,実験的に検証しやすいことが挙げられる．まずアミノ酸を電荷や疎水性などの性質に基づいて 23 のグループに分類し，それぞれのアミノ酸グループが周期性に現れる頻度を蛋白質ごとに計算した.各々の蛋白質の周期性を SVM 法で学習させ，周期スコア(PD score)を独自に定義した．さらに二次元プロット法を用いて PD score をアミノ酸使用頻度と併せて用いることで，非常に高い精度(87%)で核酸結合蛋白質群を分類することに成功した．さらにこの周期性を機能未知蛋白質に適応することにより 10 個の新規核酸結合蛋白質を実験的に同定した．

一方で進化の側面から遺伝情報の形質発現において重要な tRNA 分子に着目し、古細菌における tRNA 進化に関する新たな仮説を提唱する．通常 tRNA はゲノム中のある領域から転写され，プロセシングを受けて成熟 tRNA になるが，ナノ古細菌が有する tRNA のうち 6 つはゲノム中の 2 箇所の領域から転写され,その後トランススプライシングを経て成熟 tRNA なる．このような tRNA を split tRNA と呼ぶ．私は split 型, イントロン型, 非イントロン型の 3 種類の tRNA の間に進化的な関連性があると仮定し，多種間比較を用いた系統解析を行った．まず全ゲノム配列が既知である 30 種の古細菌に対して tRNA 予測ソフト SPLITS を用いて，3 種類の tRNA すべてを網羅的に予測した．その結果，信頼性の高い 1302 本の tRNA が予測された．予測された tRNA に対してイントロン配列を取り除き、エキソン配列のみでアライメントを行ったところ，3 種類の tRNA はその型に関係なく同義コドンを持つ tRNA 間で配列の identity が 80%以上のクラスタを形成した．さらにすべての tRNA をポジション 37/38 で 5'断片と 3'断片に分けて split 型 tRNA を模倣し分類した結果，コドンの遺伝暗号表は特定の 5'及び 3'tRNA 断片の組み合わせと良く対応し，進化的に共通して使われる断片があることが明らかとなった．これらの結果から古細菌における tRNA の進化において 5'と 3'の配列のセレクションが過去に起きたことが示唆された．

キーワード：蛋白質機能予測，DNA/RNA 結合蛋白質，アミノ酸周期性，サポートベクターマシン，tRNA 進化，多種間比較，系統解析

慶應義塾大学大学院　政策・メディア研究科

藤島　皓介

Contents

Chapter 1. Introduction

The RNA molecule can store genetic information just like DNA and acts as a ribozyme to perform various biological processes like protein enzymes. Recently, various systematic screens have identified a new class of RNA known as non-coding RNA (ncRNA) genes., More that 50% of the genome sequence in mouse are known to be transcribed as a ncRNA and some of these transcript seem to play an important role in the gene rexpression, post-transcriptional regulation and even act as a sensor of metabolites, pH and heat to quickly respond to the change in the intracellular environments. Thus, RNA network seem to be supporting and regulating the main stream of genetic information and overall biological processes (Figure 1).

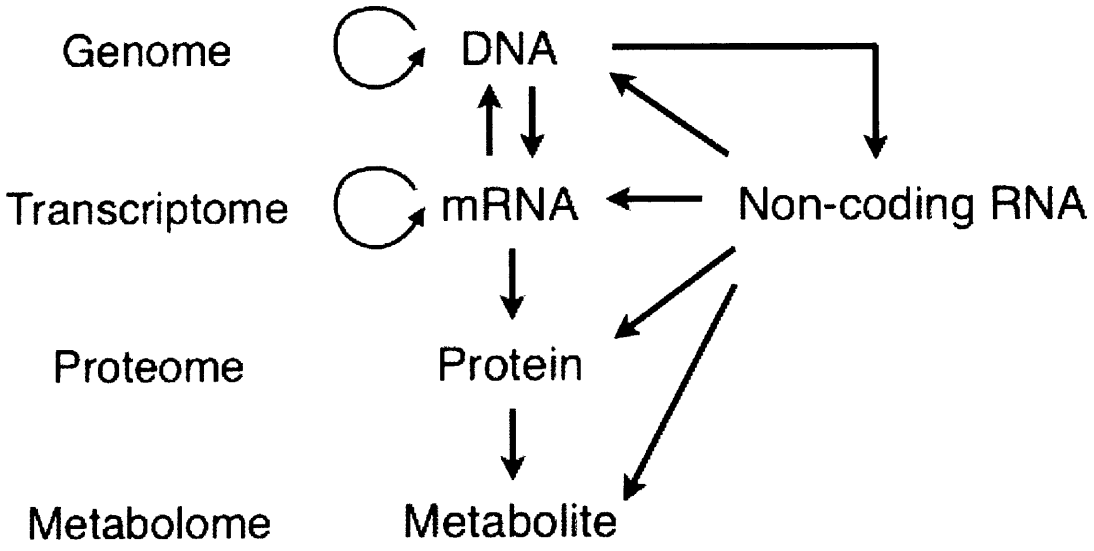


Figure 1. Schematic representation of ncRNA network affecting the central dogma

The ncRNA molecule acts dynamically throughout the main stream of genetic information known as central dogma.

Although some of the biological functions of ncRNAs have been revealed, dynamic

interaction between other molecules especially proteins are yet to be analysed. Considering the fact that about half of the proteins are un-annotated and along with the emergence of mass ncRNAs, discovery of novel RNA-protein regulatory networks are to be expected. Further, revealing the function, origin and evolution of these RNA-protein networks will be an interesting theme for both biologists and molecular evolutionists. In this manuscript, we introduce two different analyses to reveal the dynamics in the ancient RNA-protein world focusing on archaeal proteins and tRNA molecule. In chapter 2, we introduce a new method to predict novel DNA/RNA-binding proteins from functionally unknown proteins using amino acid composition and periodicity in hyperthermophilic archaeon *Pyrococcus furiosus*. In chapter 3, represent a new theory to explain the evolution of tRNAs in archaea by analyzing the sequences of 5' and 3' tRNA halves separately.

Chapter 2. Proteome-wide prediction of novel DNA/RNA-binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon *Pyrococcus furiosus*

2.1 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 (PDB) 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 (so-called 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 those sequences currently in databases. More than 23,000 ORFans have been found in 60 microbial genomes, and, based on 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 based on various amino acid profiles (i.e., overall composition, pseudo-amino acid composition, surface composition, electrostatic potential and hydrophobicity) [13-15]. SVM is one of the most powerful supervised learning algorithm that recently has been widely used in the field of bioinformatics.

We describe here a SVM-based method for classifying known DNA/RNA-binding proteins from *P. furiosus* using amino acid composition and periodicity as a feature vectors. The discriminant value (SVM output) derived from these profiles were defined as a new indices: composition (CO) score and periodicity (PD) score. Amino acid composition are known to be strongly correlated with protein secondary structure class [16] and subcellular localization [17, 18] and are assumed to support the protein function classification. Therefore based on the 2D correlation analysis, we combined amino acid composition (CO score) with PD score to further improve the performance of DNA/RNA-binding protein prediction. The 2D 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 as novel DNA/RNA-binding proteins.

2.2 Materials and methods

2.2.1 Protein dataset and functional annotations

Automated annotations and amino acid sequences of proteins from the two archaeal species, *P. furiosus* (2057 proteins) and *S. 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 the UniProtKB/Swiss-Prot (http://www.ebi.ac.uk/swissprot/ : 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 from the domain databases InterPro [24] and Pfam [25]. Swiss-Prot data were used for the four prokaryotic and eukaryotic species—*B. subtilis* (2799 proteins), *E. coli* K12 MG1655 (4465 proteins), *A. thaliana* (3454 proteins) and *C. 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 dataset, 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).

2.2.2 Amino acid periodicity

To analyze amino acid periodicities, we used eight physico-chemical profiles (Chemical, Sneath, Dayhoff, Stanfel, Functional, Charge, Structural, and Hydrophobicity) [37] to subdivide the 20 common amino acids into groups. For example, the "Charge" profile divided the 20 amino acids into the 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, RNDQEHK, ILMFV, AFILMPVW, ACGILMPSTV, and CDEGHKNQRSTY.

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 (XCXXXCXXCXXXCXXCX) contains five "Cs" with multiple periodicities (two of length 3, two of length 4 and two of length 7). Based on the error range ±1, length 4 is included in length 3; therefore Seq3 is defined to have "C" periods of only 3 and 7.

2.2.3 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 based on the individual dataset 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 × 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 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 a feature vector and classified 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 by the default parameters in Gist package version 2.3, which contains software tools for SVM classification [22]. After SVM

training based on the protein training set of amino acid periodicities, maximum-margin hyperplane was applied to the protein test set based on 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 based on amino acid composition and discriminant values were defined as the CO score.

2.2.4 Validation of PD score performance

The performance of the PD score at predicting novel proteins was validated based on 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 dataset of *P. furiosus* was randomly split into 10 mutually exclusive subsets D_1, D_2,..., D_{10} of approximately equal size. Each subset was tested based on the training using the rest of the 9 subsets. Estimated accuracies were derived as average values.

First, the classification accuracy of the PD score was compared with that of single amino acid periodicity using receiver operating characteristic (ROC) studies. 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 PD score cut-off of 0. Equations are represented below:

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

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 dataset representing the standard deviation values derived from the 10-fold cross-validation test.

Second, PD score was compared against CO score (amino acid composition based SVM) and other SVM-based protein function predictor, SVM-Prot. To assess the PD score performance, we calculated the overall accuracy (ACC) for PD score using 10-fold cross-validation test. Training dataset of SVM-Prot is fixed as a combination of 54 functional protein families and predicts several functional classes due to the probability of correct prediction [23]. 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. To equally validate the prediction performance of SVM-Prot with our method, 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 dataset and ACC is calculated as the given equation:

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

To extract proteins that are more likely to be novel DNA/RNA-binding proteins, an index: positive predictive value (PPV) was adapted to measure the percentage of DNA/RNA-binding proteins among proteins above certain thresholds (blue line in Supplementary Figure 1). PPV is calculated as the given equation:

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

The final prediction decision is given by using the calculated value of the Matthews correlation coefficient (MCC) [38] to determine the threshold value of the CO+PD score. The MCC is a popular index for measuring the performance of prediction, maximum MCC provides efficient sensitivity and specificity.

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

where *TP* refers to true positive (number of DNA/RNA-binding proteins with CO+PD score > cut-off), *FP* refers to false positive (number of other proteins with CO+PD score > cut-off),

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

2.2.5 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) as described previously [39] and partially digested with the restriction enzyme Sau3AI. 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 with NdeI and XhoI 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 expression of PF1498. Transformants were grown at 37 °C in Luria–Bertani (LB) medium containing 50 μg/ml ampicillin and supplemented with 0.4 mM isopropylthio-b-galactoside. After 14 to 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 a 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).

2.2.6 Gel-shift assay

5′-end FAM-labeled oligonucleotides were chemically synthesized by Hokkaido System Science Co. (Hokkaido, Japan). Binding reactions containing the oligonucleotide (125 or 500 nM) and 0.1 to 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). For sequencing of the oligonucleotides we used the following two probes (Xiaojing *et al.*, to be published separately):

(1) MPOR-27, 5′-r(GAAACAAGGAGAAAUGGUUCGUGUCCU)-3′,

(2) MPOD-27, 5′-d(GAAACAAGGAGAAATGGTTCGTGTCCT)-3′.

2.3 Results and discussion

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

P. furiosus, *Sulfolobus solfataricus*, *Bacillus subtilis*, *Escherichia. coli*, *Caenorhabditis elegans*, and *Arabidopsis 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 [19] 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 [20, 21]. We defined three categories of proteins based on the number and quality of annotations (see Methods). 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 dataset to 477 non-redundant proteins for the periodicity analysis. To facilitate their use as a training dataset 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).

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

Species	Database	Functionally known					Putative	Hypothetical	Total protein
		*Representative set			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**	1590	2455	56	144	3454
C. elegans	Swiss-Prot	165	401	**566**	1215	2580	0	874	2655

* Representative set consists of proteins with amino acid length > 20 and homology redacted using BLASTP (E-value < 1 x 10^{-4})

2.3.2 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 physico-chemical 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 dataset 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 [22]. 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 dataset of each of the six model species.

Performance of the PD score as a DNA/RNA-binding protein classifier was evaluated by applying the receiver operating characteristic (ROC) curve to the representative set of *P. furiosus* proteins (Figure 2).

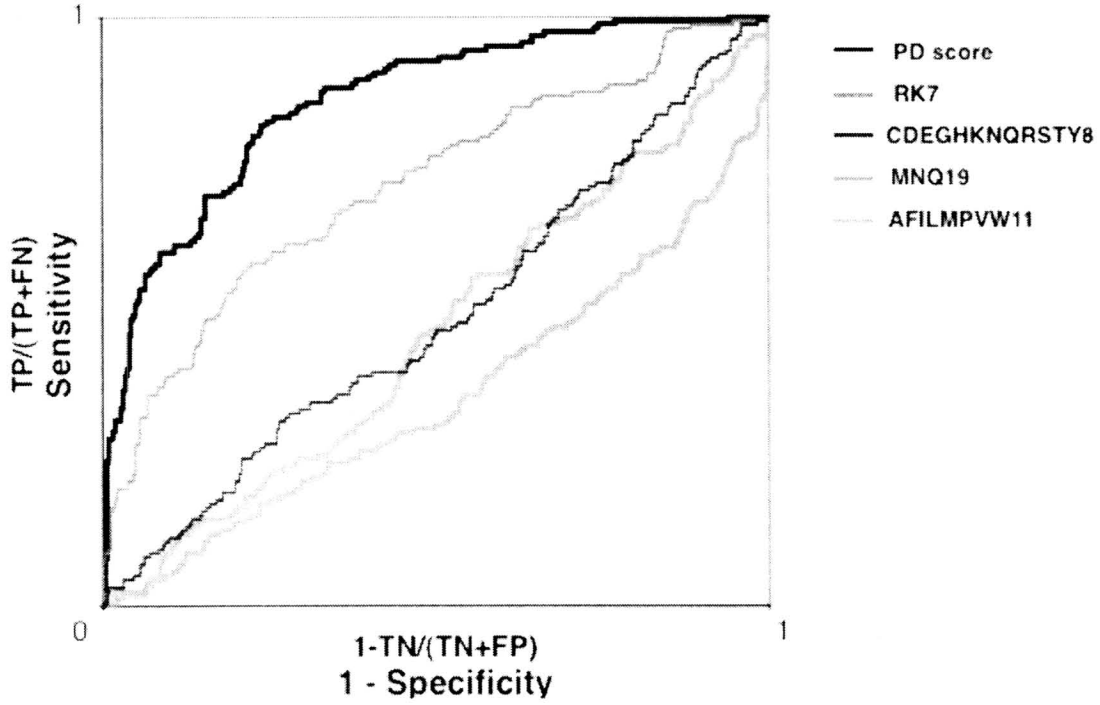


Figure 2. Performance of PD score

Receiver-operator-characteristic (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 periodicity of 7.

Sensitivity and specificity of the PD score overwhelmed that of various individual amino acids periodicities (RK7, CDEGHKNQRSTY8, MNQ19 and AFILMPVW11). This demonstrated that a combination of amino acid periodicities as a feature vector optimizes the system for classification of DNA/RNA-binding proteins.

To further validate the performance of PD score, we conducted a comparative analysis upon amino acid composition and SVM-Prot [23]. 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).

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

		PD score			FQ score			SVM-Prot		
Species	Sample size	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)	79.9 (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

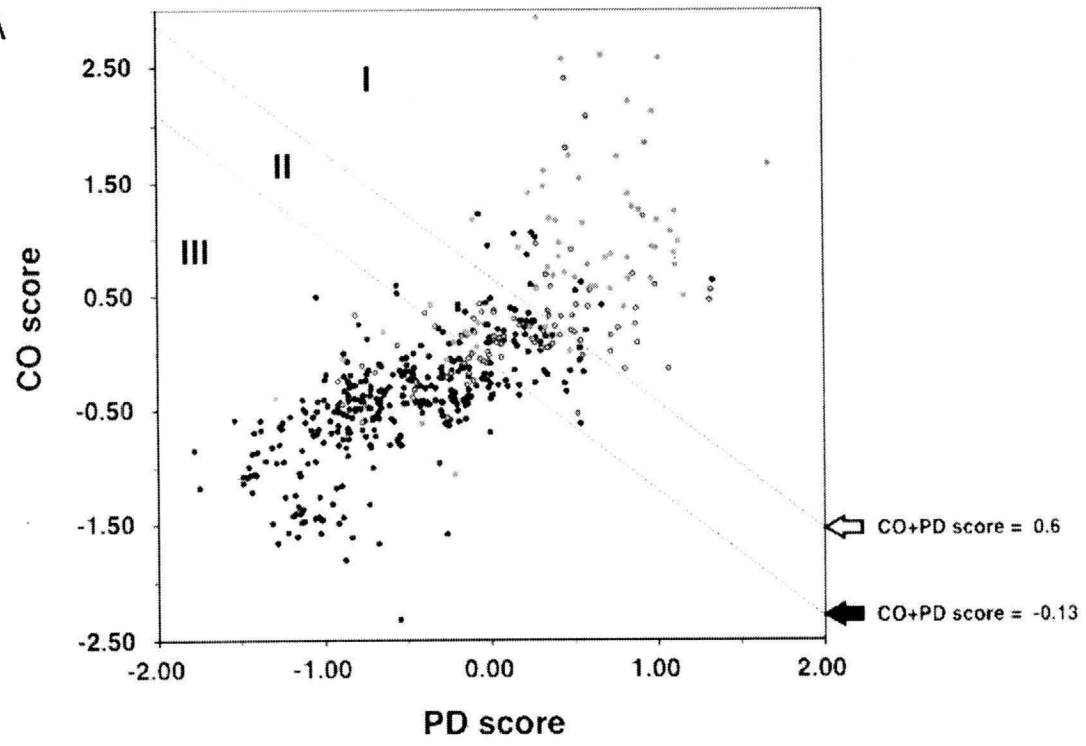
SD stands for standard deviation. Predicted results are shown as SE (sensitivity) = TP/(TP+FN) , SP (specificity) = TN/(TN+FP) and ACC (accuracy) = (TP+TN)/(TP+FN+TN+FP). Numbers in red indicates the highest index among the three classifiers.

The three predictors have shown different characteristic in predicting DNA/RNA-binding proteins due 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 PD score was comparable to other methods but statistical significances cannot be observed based on the comparative analysis. Thus, we combined the two indicator CO score and PD score to improve our DNA/RNA-binding prediction method.

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

We performed 2D correlation analysis of CO score and PD score upon 477 functionally known proteins in *P. furiosus*. The correlation coefficient was r = 0.75 (overall) and r = 0.55 (DNA/RNA-binding proteins only) respectively. The 157 DNA/RNA-binding proteins (red and blue circles in Figure 3A) distributed at the right-upper region of the 2D plot, suggesting that both CO and PD score are required for classifying proteins with DNA or RNA-binding activity (Figure 3A).

A



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 3. 2D correlation analysis of DNA/RNA-binding proteins in *P. furiosus* based on amino acid composition and periodicity.

(A) Total 477 functionally known proteins in *P. furiosus* was plotted on 2D 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 = 0.59 for optimizing sensitivity and specificity (black arrow) and maximum ACC value = 81.8% for optimizing 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 class I to class III.

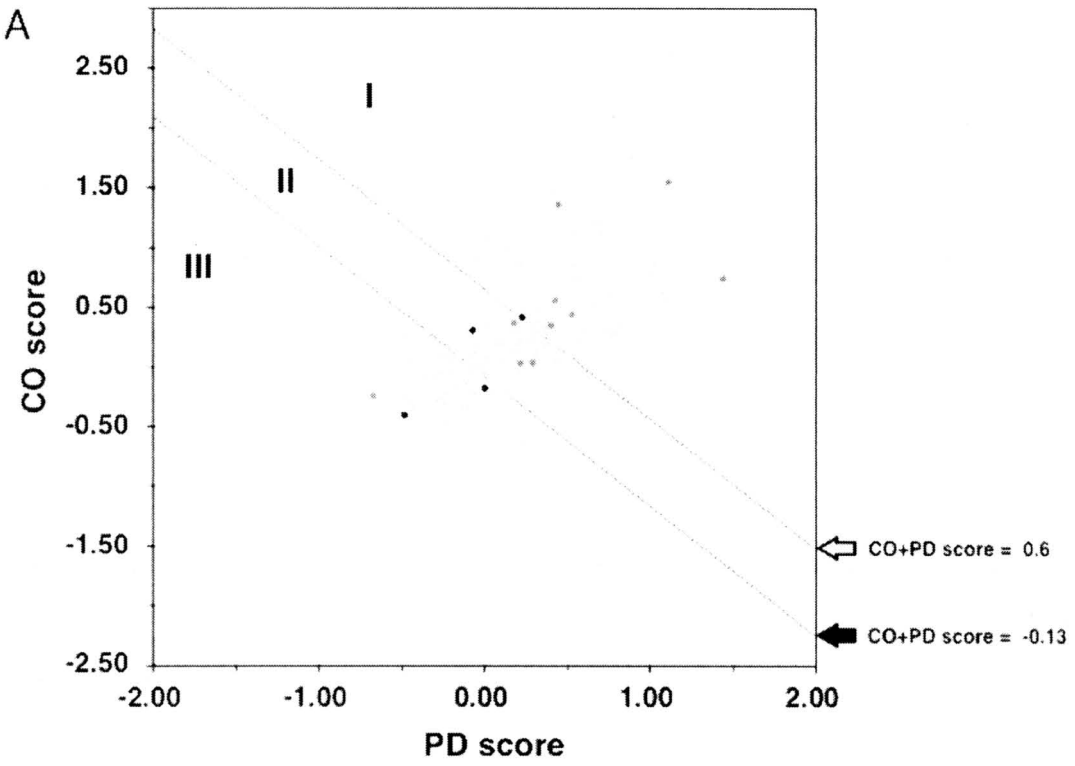
For efficient classification of DNA/RNA-binding proteins, we defined two different thresholds using a value of CO score + PD score (CO+PD score). First threshold is based on the highest overall accuracy (ACC) with CO+PD score = 0.6 and the second threshold is based on the highest Matthews correlation coefficient (MCC) with CO+PD score = -0.13. According to the Supplementary Figure 1, 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%). Based on these thresholds, we classified proteins into three classes (class I – class III) (Figure 3A). As a result, total 94 proteins including 82 DNA/RNA-binding proteins (Figure 3B) were categorized as class I proteins (CO+PD score > 0.6).

The 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 to 0.5 but some of them were dominant at high PD score region (0.25 – 1.5). This region includes 13 tRNA-processing

enzymes (i.e., tRNA-synthetases, CCA-adding enzyme and Rnase P subunits), 11 DNA-binding proteins (i.e., DNA polymerase, DNA helicase, DNA primase and reverse gyrase), 3 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-antitermination 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.

2.3.4 Selection and experimental verification of novel DNA/RNA-binding protein candidates

The same procedure was applied to 994 hypothetical proteins in *P. furiosus* (Figure 4).



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 4. 2D correlation analysis of hypothetical proteins in *P. furiosus* based on amino acid composition and periodicity.

(A) The vertical and horizontal axis represents (CO score) and periodicity (PD score). Distribution of the 994 hypothetical proteins are 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 were 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 class I to class III.

The 2D plot of hypothetical proteins was similar to that of functionally known proteins as well as the protein ratio in class I to classIII (Figure 2 vs Figure 3). Although, the number of proteins have 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 were classified in to three classes (I-III) due to the CO+PD score thresholds and total 151 proteins were classified as a strong candidate 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 three different class I - III (9 from class I, 5 from class II and 3 from class III, Table 3).

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

			SVM Analysis			Experimenal verification
Class	Gene ID	Mol Mass (kDa)	PD score	CO score	PD + CO	DNA/RNA-binding activity
I	PF1498	16.5	1.11	1.55	2.66	+
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. *The RNA-binding activities of PF1498 was examined by 1.2% agarose gel electrophoresis and ethidium bromide staining due to the co-purification with endogenous RNA in *E. coli* (see Figure 5C).

All 17 recombinant proteins were overexpressed in *E. coli* and purified to near homogeneity (Figure 5).

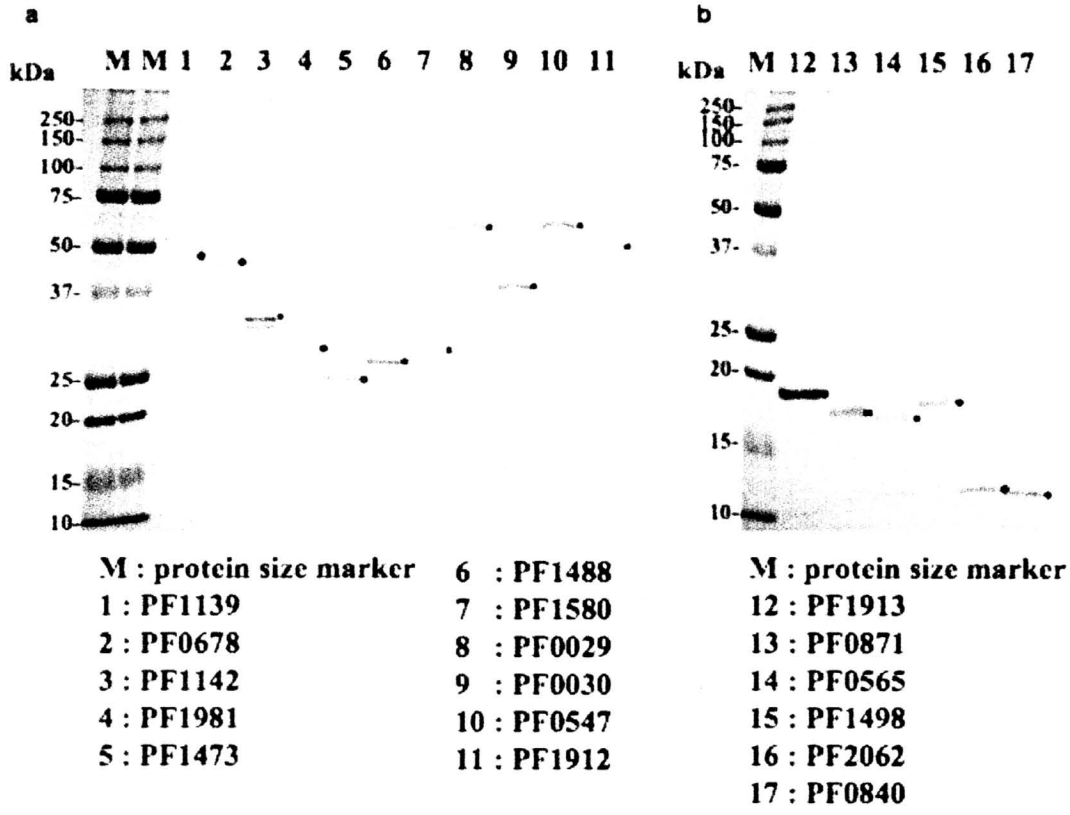


Figure 5. SDS-PAGE analysis of purified 17 hypothetical proteins in *P. furiosus*.

SDS-PAGE analysis of purified candidate proteins. The positions of the 17 purified candidate proteins are marked with black dots. 10–20% (left column :a) and 15–25% (right column :b) gradient gel was used for proteins of large and small molecular size (**M**, protein size marker; Bio-Rad).

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) (Figure 6A). 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. Three proteins, PF0871, PF0678 and PF0840, aggregated in the loading well, so we removed them from the final results.

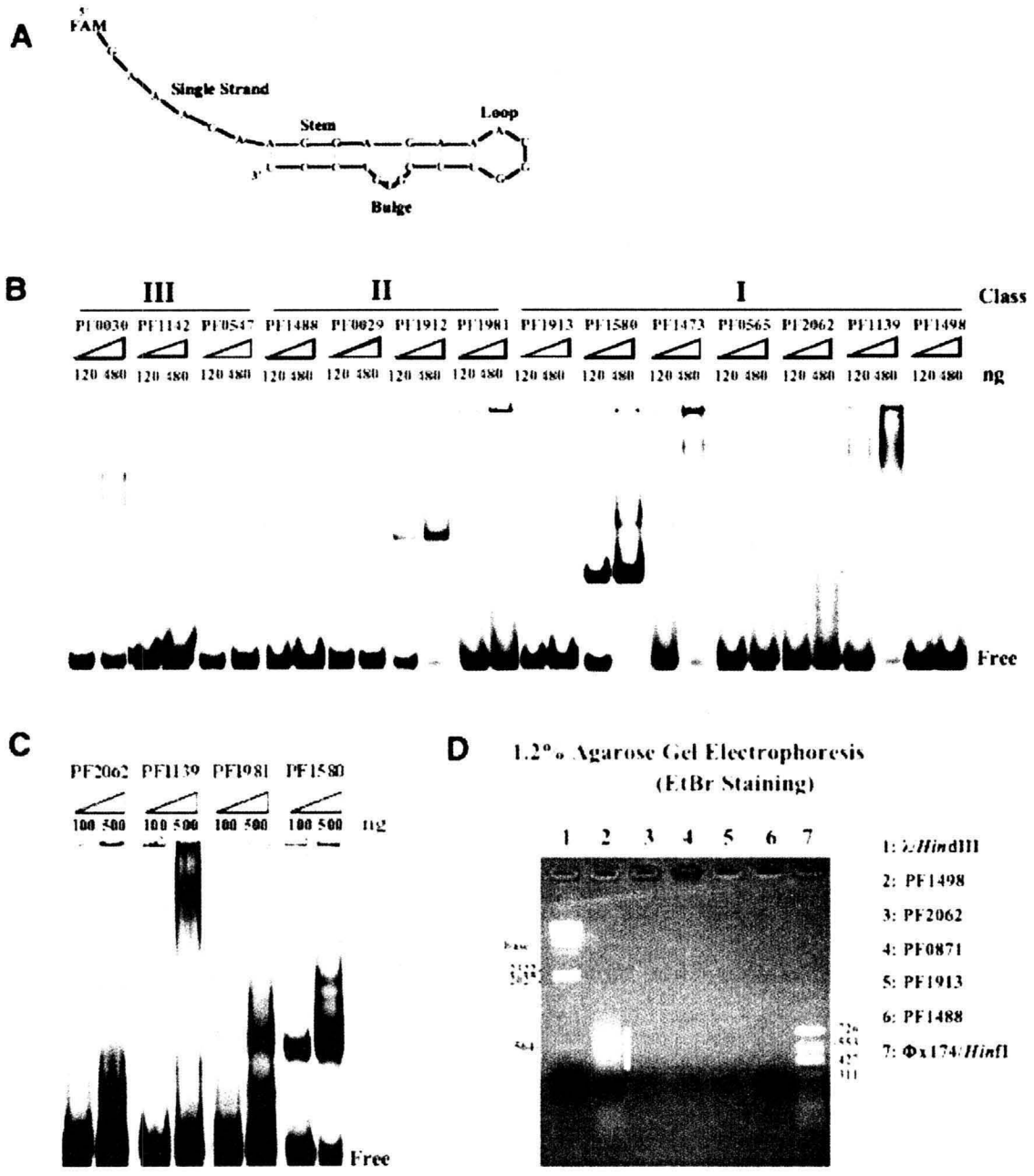


Figure 6. Experimental verification of DNA/RNA-binding activities of 14 candidate proteins.

(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)** 1.2% agarose gel analysis of purified protein peak fractions. White bar indicates the existence of nucleic acids. Lanes 1 and 7 are DNA markers.

A prominent shift of the RNA probe up the gel was observed in candidate proteins PF0029, PF0030, PF0565, PF1139, PF1473, PF1580, PF1912, and PF2062 (Figure 6B). 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 (Figure 6C vs Figure 6B). 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 (Figure 6D).During our investigation, five out of six class I proteins, three out of four class II proteins and one class III proteins were determined as novel DNA/RNA-binding proteins (Table 3).

According to our previous works [8-10], systematic screening of *P. furiosus* genome using the expression cloning method has determined several DNA/RNA-binding proteins (such as Rnase HII, FEN-1, Thy-1 and FAU-1) and its possible biological functions. Our system also demonstrated that approximately 10% to 20% of the *P. furiosus* gene products are to be involved in nucleic acid metabolisms. Thus we suggest that the 10 newly discovered DNA/RNA-binding proteins are determined through experimental procedure with certain specificity. The *in vivo* targets and precise biological functions of the 10 newly identified DNA/RNA-binding proteins are to be further investigated.

According to the domain assignment of the InterPro/Pfam domain database [24, 25] 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 annotation (Supplementary Figure 2). According to Supplementary Table 1, 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 swissprot protein entries. The remaining six proteins have shown sequential similarity to the hypothetical 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. Thus, we believe that combination of CO score and PD score is a powerful indicator for predicting DNA/RNA-binding proteins from sequence specific ORFans and a set of proteins having no obvious functional domains, even allowing that the sample size of validated proteins is still small.

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

As the amino acid composition within proteins varies among taxa [26], our method removes the need to allow for the evolutionary gain and loss of amino acids and increases the generalization capability of SVM training. Figure 7 presents an example of charged amino acid groups that appear periodically in the amino acid sequence of DNA primase.

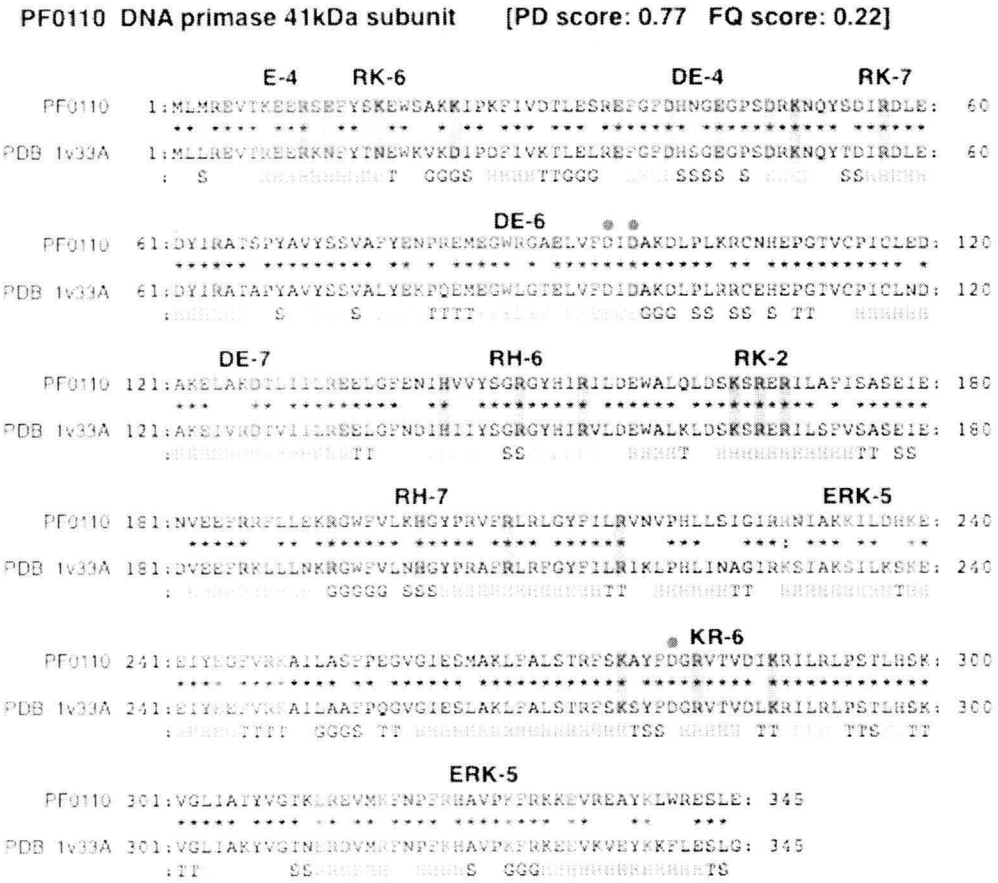


Figure 7. Schematic representation of Amino acid periodicity in the sequence of DNA primase.

The distribution of amino acid periodicities with structural features were observed in many class I proteins with high PD score and low CO score (PD score > 0.25 and CO score < 0.5). Amino acid sequence was aligned with closest structural orthologues registered in the Proteins Data Bank (PDB). The periodic region consists of overall region (colored squares) and amino acids corresponding with periodicities (oblong box). Periodicity includes error range ±1. The conserved amino acids between *P. furiosus* protein and the structurally known proteins with same functions are marked with asterisks. DNA primase 41kDa subunit (PF0110) is aligned with orthologous protein PF0195 in *Pyrococcus horikoshii* (PDB_ID: 1v33A) with 79.1% identity. Putative active site residues are marked with red dot.

An amino acid periodicity of both positively and negatively charged amino acids with various periodicities were widely found throughout the protein primary sequence. The amino acid residues creating the periodicity (oblong boxes in Figure 7) are often conserved in the 3D

structures of orthologous proteins. Similar feature was observed in other proteins with high PD score such as Signal recognition particle 54 kDa subunit (SRP54) and HTH-type transcriptional regulator lrpA (Supplementary Figure 3). Periodic region also covers DNA/RNA recognition motifs of these proteins known as M domain and Helix-Turn-helix as well as DNA primase active sites. We suggest that these periodic features might affect the secondary structures or the net charge of the protein surface to enhance the DNA/RNA-binding capacity.

These 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 [27] 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 site region [28-30]. 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 [31], 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. 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 [32-34], a comparative genomics approach that predicts function based on the comparison of various domains [35], and three-dimensional protein models [36].

2.4 Conclusions

In this paper, 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 2D 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. Similar distribution plot was observed for the hypothetical proteins. Total 10 novel DNA/RNA-binding proteins were determined experimentally including four ORFans and two proteins with no domains. The 2D 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 prediction results derived from CO+PD scores can be further integrated with prediction results from various protein function prediction and annotation methods to validate uncharacterized proteins comprehensively. Further 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.

Chapter 3. A new hypothesis for tRNA evolution in archaea: tRNAs were evolved through the combination of ancestral 5'half and 3'half tRNA fragments

3.1 Introduction

The origin and evolution of tRNAs are widely discussed since the accumulation of genome sequences has provided comparative analysis of numerous tRNAs. A model has been proposed suggesting that the tRNA molecule must have originated by direct duplication of an RNA hairpin structure on the origin of the transfer RNA molecule [41]. Preceding studies have shown that single hairpin RNA structure called minihelix (tRNAacceptor-TΨC arm helix) can be aminoacylated by modern tRNA synthetases [42]. Thus two halves of tRNA, 5' half (containing D arm and anti-codon) and 3' half (containing acceptor stem and T arm) are considered to be originated and evolved independently in the early genomes. Interestingly tRNA-like structure containing 3'-terminal CCA sequence appears in the 3'-end of various RNA virus, retroplasmid and bacteriophage genomes to initiate RNA/DNA replication by replicase or reverse transcriptase [43,44]. Futher, 3' half is recognized by series of tRNA maturation enzymes: RNase P, aminoacyl-tRNA synthetases and CCA-adding enzymes suggesting that 3' half of tRNA is an ancient structural domain acting as a "genomic tag" in early protein-RNA world [45]. Therefore 5' half is considered to be arose after 3' half to provide additional specificity of corresponding amino acids at 3'-end [44]. Recently, 3' and 5' tRNA fragment of total six different tRNA genes were found at separate region in the genome of hyperthermophilic archaeal parasite *Nanoarchaeum equitans* [46]. Both 3' and 5' tRNA half genes possessed conserved polymerase III promoter consensus box A motif at the 5' flanking region suggesting that these sequences are transcribed independently. There is also an argument of an evolutionary link between these split-tRNAs and intron containing tRNAs. Transfer-RNA genes in archaea often have introns intervening between exon sequences. The structural motif at the boundary between exon and intron is the bulge-helix-bulge (BHB) motif [47]. This motif was also found in the processing leader sequence of split-tRNAs [48]. Thus, we consider the evolutional linkage of the three types of tRNAs by analyzing exonic

sequence of the 1302 tRNA as well as 5' and 3' tRNA fragments separately to support the theory that the present tRNAs were emerged from the combination of individual 5' and 3' tRNA fragments. Phylogenetic analysis has shown that exonic sequences of the split-tRNAs branched near other archaeal tRNAs with identical and synonymous anticodon encoding the same amino acids. The combination of 5' and 3' tRNA halves correlated with the variation of amino acid in the genetic code. We have identified a set of tRNAs with conserved 3' tRNA halves but with differed sequence of the 5' region. These results indicates that in the early stage of archaeal tRNA, a variety of 5'-3' tRNA combinations have evolved to construct the genetic code.

3.2 Materials and Methods

3.2.1 Preparation of the genomic data

The genome sequences of 30 archaeal species were downloaded from European Bioinformatics Institute web server (http://www.ebi.ac.uk/genomes/archaea.html) in EMBL file format (Rel. 83) except, the genome sequence for Uncultured methanogenic archaeon RC-I (GI:116077928), was downloaded from National Center for Biotechnology Information. In total five Crenarchaeota (such as *Aeropyrum pernix K1,* ape), twenty four Euryarchaeota (such as *Pyrococcus furiosus*, pfu) and one Nanoarchaeota (such as *Nanoarchaeum equitans*, neq) were prepared. Three types of tRNAs: nonintronic-, intronic- and split- tRNAs were obtained using tRNA predicting software SPLITS [49]. The intron and split introns (sequence constructing the BHLmotif and GC-rich duplex for the split-tRNA) are removed to prepare total 1302 mature tDNA sequences for the analysis.

3.2.2 tRNA sequence analysis

All tRNA sequences (input as tDNA sequences) were first aligned using ClustalW [50] with

the following settings: the Multiple Alignment parameter Gap Opening 10.00; the Multiple Alignment parameter Gap Extension Penalty 0.1; and the Multiple Alignment parameter Delay Divergent Sequence 25% [51]. Several types of phylogenetic trees were constructed using sequence alignment data. A neighbor joining (NJ) tree was constructed using ClustalW. Maximum likelihood (ML) tree and Maximum parsimony (MP) trees with 1,000 bootstrap replicates were constructed using PAUP 4b10 [52]. Bayesian trees were constructed using MrBayes v3.1.2 [53]. The model of sequence evolution was determined by MrModeltest ver2.2 [54]. Finally, these phylogenetic treefiles was described using Hypertree and Treeview..

3.3 Results and discussion

3.3.1 Comprehensive phylogenetic analysis of tRNAs in 30 archaeal species

Total 1302 archeal tRNAs including six known split-tRNAs were aligned based on clustal X algorithm. Pairwise alignment was performed based on the exon sequence of the tRNA gene (using tDNA sequence as an input).

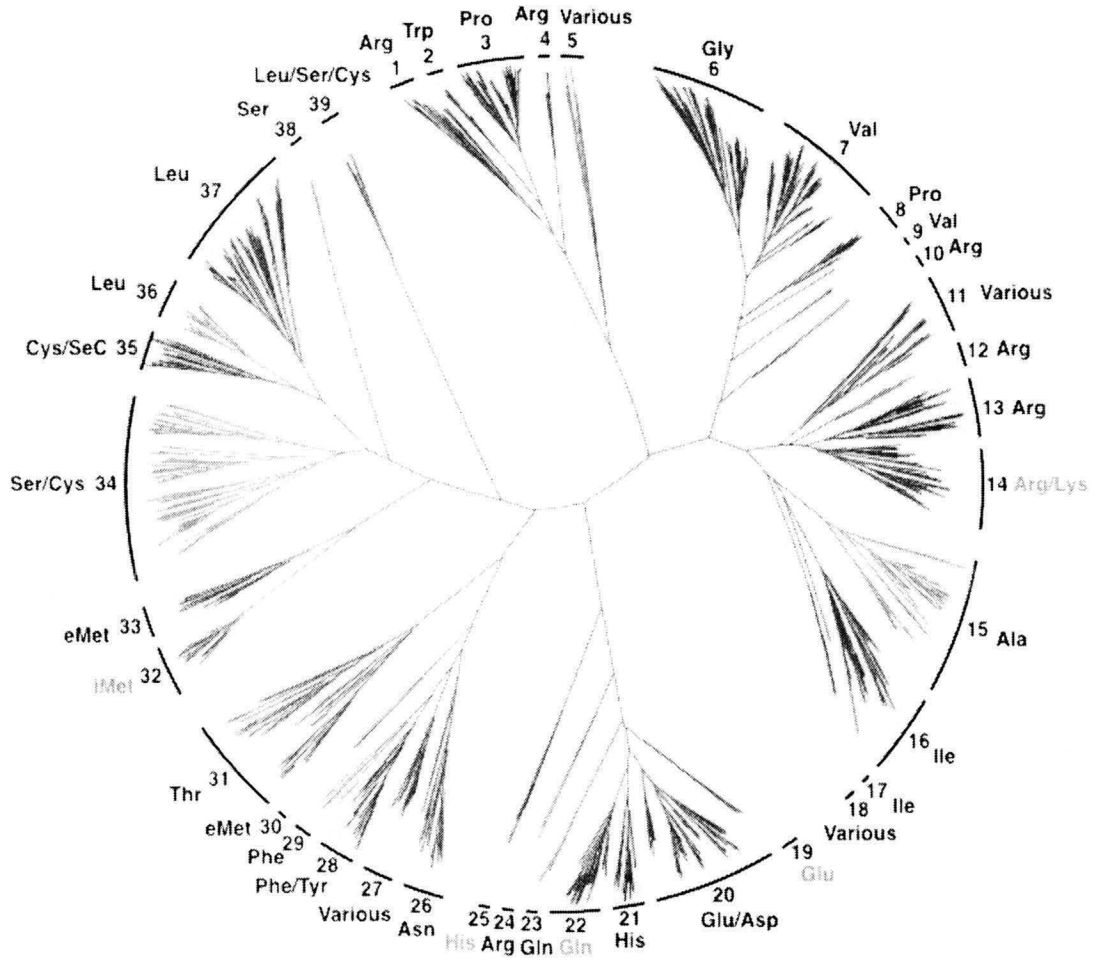


Figure. 8. Phylogenetic tree of 1302 archaeal tRNAs.

The phylogenetic NJ tree was constructed based on the alignment of predicted 1302 tRNA sequences from the complete genome of 30 archaeal species. The tRNA clusters are denoted by number and corresponding amino acids. Red letters indecate the phylogenetic position of the six split-tDNAs. Cluster denoted as "various" include tRNAs corresponding to several amino acids. The tRNA clusters are classified into four categories; clusters including *N. equitans* derived tRNAs (red), without *N. equitans* derived tRNAs (orange), cluster consists of euryarchaeon derived tRNAs only (blue) and crenarchaeon derived tRNAs only (green).

As a result, unrooted NJ tree was constructed and relative distance of each tRNA taxon was determined. The 1302 tRNAs were clustered into 39 groups based on their sequence similarity and most of the tRNA clusters were dominated by synonymous tRNAs with 1 or 2 corresponding amino acids (Figure 8). The exonic sequence of split-tRNAs and other intronic

tRNAs were clustered within the same cluster suggesting an evolutional linkage between the three types of tRNAs. (The pricise phylogenetic locations of the split tRNAs and intronic tRNAs were further discussed in the section 3.3.2). Total 31/39 (79%) clusters were dominated by the tRNAs corresponding to the single amino acids suggesting that these tRNAs with synonymous anti-codons could have common ancestral tRNA. For example, 60 out of 61 archaeal Glycyl-tRNAs (Gly) are clustered in the same branch (cluster no.6 in Figure 8) with consensus of [N]CC (GCC, TCC, CCC) anticodon rules. The feature of first codon in archaeal tRNA severely excludes adenine (A), which were previously shown in Marck and Nikolajewa's works [55,56]. The same rules can be observed in various tRNA clusters for example, Valine (Val) with [N]AC, Proline (Pro) with [N]GG, Alanine (Ala) with [N]GC, Serine (Ser) with [N]GA, etc. These results indicate that in the early stage of archaea, ancestral tRNA anticodon and corresponding amino acids were severely restricted by the strong selectivity of second and third codon as well as lack of adenine in the first codon. Based on the exon tRNA sequence, about 80% of the tRNAs with same anti-codon should have evolved from the single origin, indicating that common ancestral archaea may have already possessed a genetic code similar to that of universal genetic code. In that environment, error in second or third anticodon will alter the charging of amino acid, which can be very critical for the species. As a result, only 17/39 (44%) of the tRNA clades were conserved among all archaea (clades filled in red: Fig.8), suggesting that about half of the tRNAs were sequentially stabilized before the differentiation of three major archaeal clades (nanoarchaeota, crenarchaeota and euryarchaeota).

3.3.2 Phylogenetic location of the six known split-tRNA and other archeal tRNAs

Since the six split-tRNAs have possessed sequential similarity with other tRNAs with same anticodon, I focused on the evolutional relationship of six split-tRNAs derived from

Nanoarchaeum equitans with other intronic and nonintronic tRNAs of archea. Split-tRNAs were aligned with tRNA sequences of the same cluster (cluster number:14, 19, 22, 25 and 32 in Fig. 8) and precise phylogenetic analysis was performed.

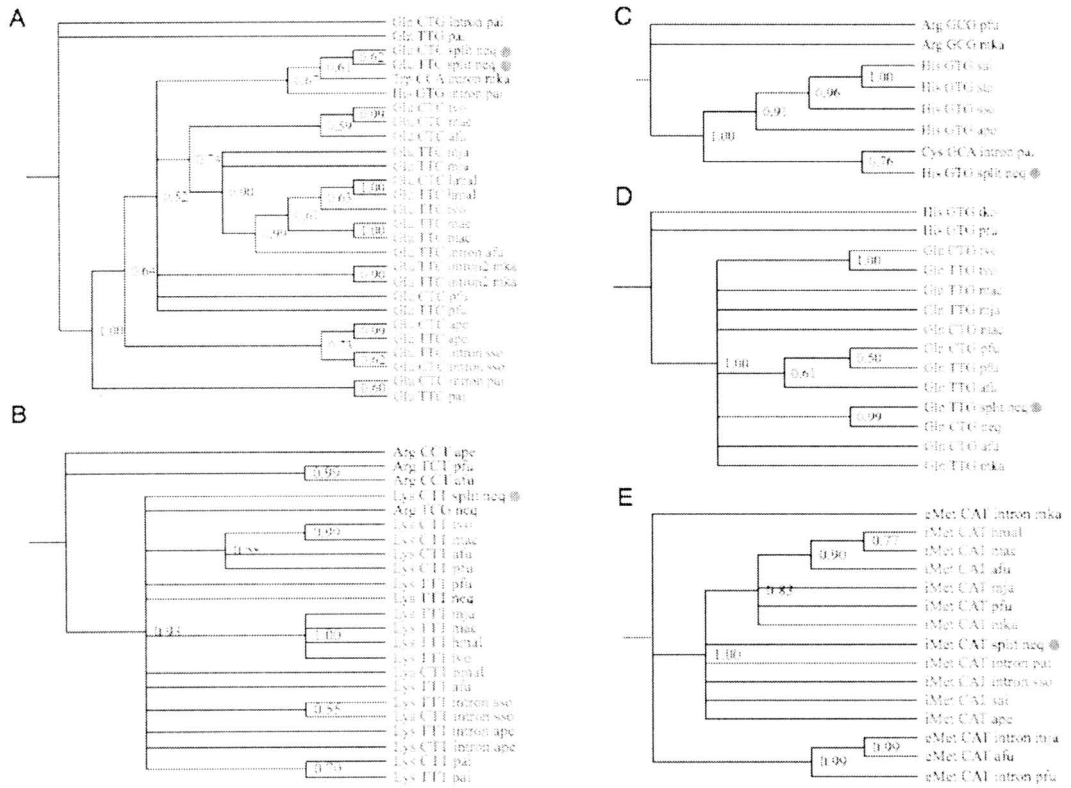


Figure 9. Phylogeny of the six split-tRNAs based on the Bayesian tree

Bayesian approach is performed for the six tRNA clusters including split-tRNA plus adjacent tRNA cluster as an outgroup using MrBayes[53]. Evolutionary model for each tRNA clusters were determined by MrModeltest2[54]. Clade credibility is described with threshold 0.5. The locations of split-tRNA taxons are shown in red circle. The colored square represents other tRNAs with synonymous codons in euryarchaea (orange) and crenarchaeota (green). (A) Bayesian tree of glutaminyl(Glu)-tRNA cluster with Glutamine(Gln)-tRNA used as an outgroup. (B) Bayesian of Lysyl(Lys)-tRNA cluster with Arginyl(Arg)-tRNA used as an outgroup. (C) Bayesian tree of Histidyl(his) -tRNA cluster with Arginyl(Arg)-tRNA used as an outgroup. (D) Bayesian tree of glutamine(Gln)-tRNA cluster with Histidyl(His)-RNA used as an outgroup. (E) Bayesian tree of initiator Methyonyl(iMet)-tRNzzA with elongator Methyonyl(eMet)-tRNA as an outgroup.

As a result, exonic tDNA sequences of the six split-tRNAs (two Glu, Lys, His, Gln and Met) were branched at the root of the subtree cluster with synonymous tRNA anticodons derived from both crenarchaeota and euryarchaeota lineages (Figure 9). The three types of split tRNA (Glu, Lys and iMet) were conserved among all archaea, although Histidyl-tRNA and glutamine(Gln)-tRNA were only conserved with tRNAs derived from crenarchaeota or euryarchaeota, suggesting that these split-tRNAs could be the common ancestor of present tRNAs Although, positive or negative selection seems to be occurred for some tRNAs which can also be confirmed in Fig.8. In addition, several intronic tRNAs were found within the same tRNA cluster, yet further analysis is necessary to discuss the origin and evolution of these introns. These results strongly support the fact that split- intronic and nonintronic tRNAs emerged from evolutionary common ancestral tRNA sequence.

3.3.3 Phylogenetic analysis reveals different origin and evolution of the 5' and 3' tRNA halves

Phylogenetic location of split-tRNAs and intronic-tRNAs have shown evolutional relation between the three types of tRNAs in archaea. Thus, archaeal tRNA sequence may still possess an evidence of the 5' and 3' tRNA selection and combination through evolution. To support this hypothesis, seven archaeal species including one nanoarchaeota: *N. equitans* (neq), three crenarchaeota: *P. aerophilum* (pae), *A. pernix* (ape) and *S. solfataricus* (sso) and three euryarchaeota: *P. furiosus* (pfu), *M. kandleli* (mka) and *M.janaschii* (mja) are chosen, for it is known as the closest taxons to the Last Universal Common Ancestor (LUCA) [56]. In total, 304 tRNA sequences from the seven species were extracted and each tRNA was separated at the cannonical exon/intron junctions (position 37/38). Unrooted NJ tree was constructed for 5' and 3' tRNA halves respectively. According to Figure 10B, the 5'half tRNAs are categorized into seven large taxa (A-F) and five small taxa mainly filled with pseudo sequences (a*-e*).

The 3'half tRNAs are clearly seperated into six large taxa. Each of the large tRNA taxa possesses root as the internal node. These phylogenetic trees tells us only about the phylogenetic relationships of each taxa. although clear difference can be observed between the phylogenetic topology of the two tRNA halves.

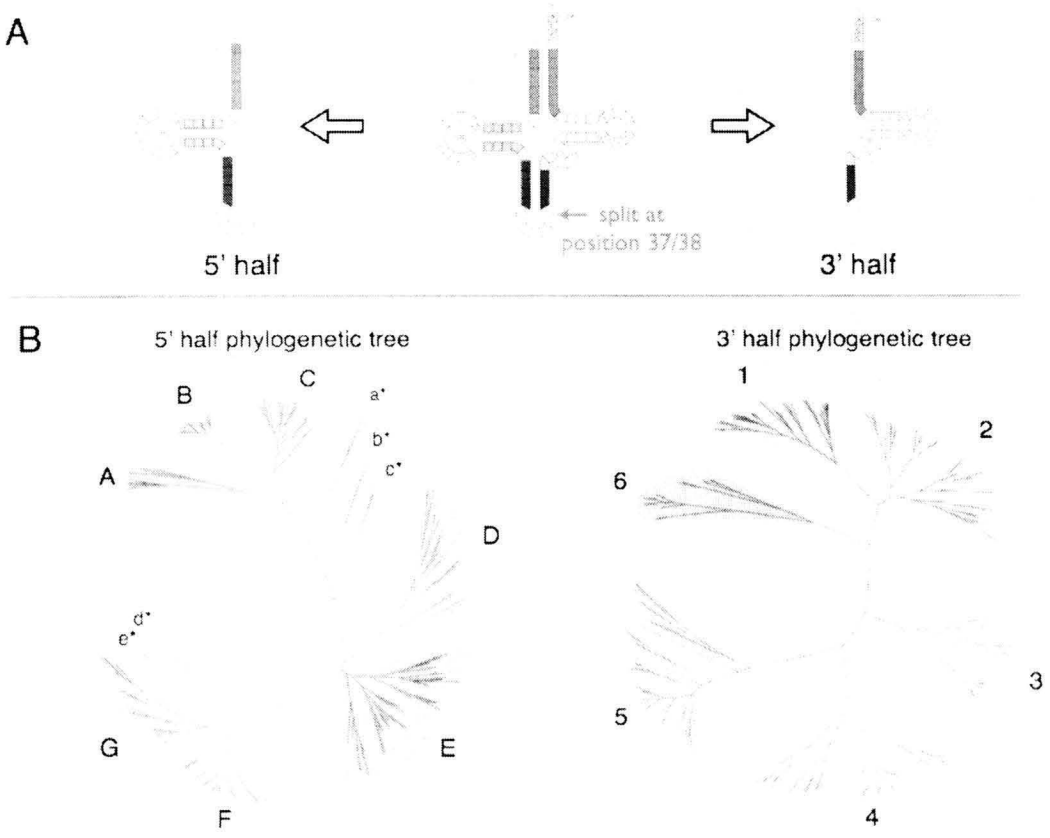


Figure 10. Comparison of the tree topology of 5' half and 3' half tRNA clusters

(A) Schematic representation of 5' and 3' tRNA halves used in the phylogenetic analysis. The tRNA sequence were separated at position 37/38 to produce 5' half (position 1 to 37) and 3' half (position 38 to end) tRNAs respectively. (B) The unrooted NJ tree was constructed based on the 304 tRNA sequences derived from 7 species (neq, pae, ape, sso, pfu, mka and mja). The cluster ID was denoted alphabetically (5' half) and numeric (3'half) for easy comparison. The color of the cluster responds to the genetic code in Figure 11.

To examine the evolutional differences between the two tRNA halves and support our hypothesis that selection of 5' and 3' tRNA combinations have actually happen in the past, we

filled the codon table with the cluster ID of corresponding tRNA halves.

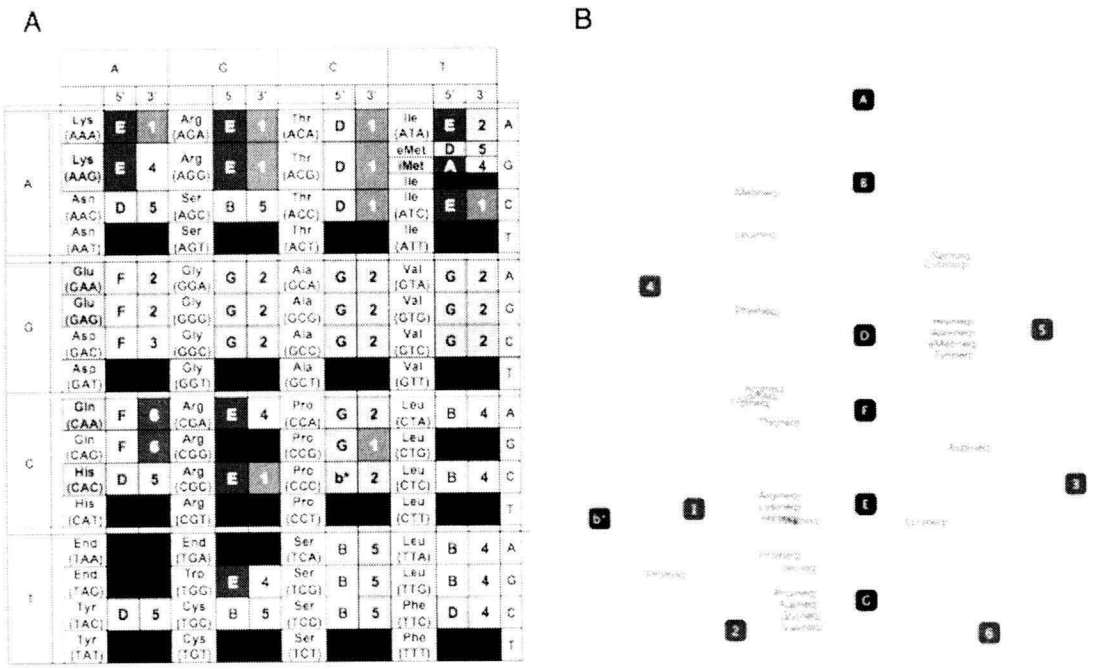


Figure 11. The combination of 5' and 3' tRNA halves in *N. equitans*

(A) Codon table is filled with the cluster ID of 5' and 3' tRNA halves derived from *N. equitans* (neq). The six codon corresponding to the split tRNA is filled with pink. The color of each ID responds to the color of clusters in Figure 10. (B) Combination of 5'- 3' tRNA in *N. equitans* are visualized as a interaction network. Each node represents 5' tRNA fragment (black) and 3' tRNA fragment (blue), thus tRNAs are represented as edge in the network.

Figure 11A shows the codon table filled with the ID of *N. equitans* 5' and 3' tRNA cluster respectively. The combination of 5' and 3' tRNAs elucidate several interesting features of how 5' – 3' combination are selected. Firstly, combination of 5' and 3' tRNA seems to be strongly correlated with the genetic code and specific relativity between the 5' -3' pair and amino acids can be observed (i.e., A-4 for iMet, F-2 for Glu, B-4 for Leu, etc). Some of the combinations are common used among several codons encoding different amino acids. For

example, tRNA encoding Glysine (Gly), Alanine (Ala), Valine (Val) and Proline with anticodon CCA are all consists of G-2 combination. These commonalities were further visualized as an interaction network of 5' and 3' tRNA fragments (Figure 11B). In *N. equitans*, all fragments except cluster C were used for constructing total 43 types of tRNA sequences. The combination of 5'-3' fragments for the six split-tRNAs are various, which includes tRNA fragments from cluster 'A', 'E', 'F' 'D', '2', '4', '5' and '6'. Same features were observed in the codon tables of other species (Supplementary figure 4). To identify the conservation of 5' and 3' pairs among the seven species, the codon tables and the network of each species were integrated and further analyzed.

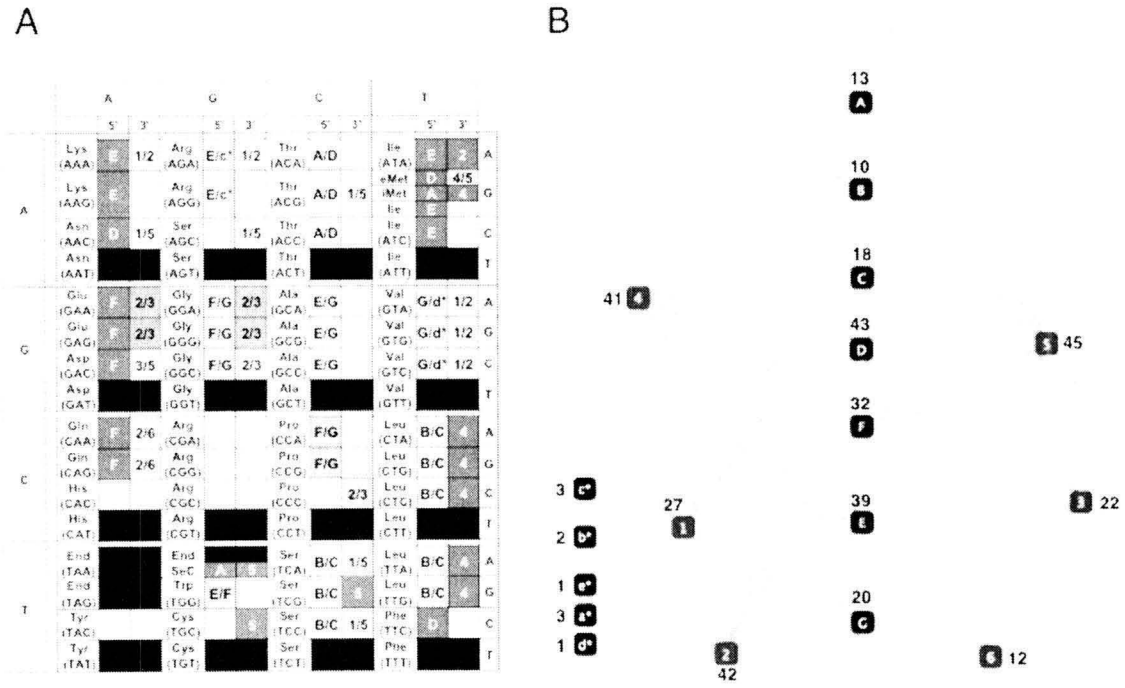


Figure 12. The combination of 5' and 3' tRNA halves in *N. equitans*

(A) A consensus codon table is constructed based on the 7 species (neq, pae, sso, ape, pfu, mka and mja). Conserved 5' and 3' tRNA halves among the 7 species were shown due to the variation of tRNA fragments for each codon. Codon with only single variation is shown in red (conserved in all 7 species). Codon with two variations are shown in yellow (variations conserved in all taxa), green (variations conserved only in neq and crenarchaeota) and blue (variations conserved only in neq). (B) Consensus 5'- 3' tRNA network is visualized. Each node represents 5' tRNA fragment (black) and 3' tRNA fragment (blue), thus tRNAs are represented as

edge in the network. The numbers denoted above the node indicates the number of edge (constructing the tRNA) linked to the node.

The universally conserved tRNA fragment was observed especially for tRNAs with A and T at 2nd codon position (Figure 12A). The 5' half sequences of five out of six split-tRNAs were universally conserved throughout the 7 archeal species. Interestingly initiator Met-tRNA clearly discriminates its sequence by using A-4 combination. Same aspect was observed for the SeC-tRNA using A-5 combination. The tRNA fragments of Prolyl-tRNAs were conserved among neq and crenarchaeota only, differing the two classes of Prolyl-tRNA (cluster number: 3 and 8 in figure 8). The overall network represents four clusters of tRNA fragment '3', '4', 'D', 'E' as a core tRNA fragments, possessing over 40 edges (constructing the tRNA). These fragments was possibly more easier to form the structure of present tRNA (such as clover leaf) as a potential candidates to be integrated in to the ancient translational mechanism.

3.3.4 Possible evidence supporting the 5' – 3' tRNA combination hypothesis

Through the series of comprehensive phylogenetic analysis of archaeal tRNA sequence, several aligned sequences have shown aspects to support 5' – 3' tRNA combination theory.

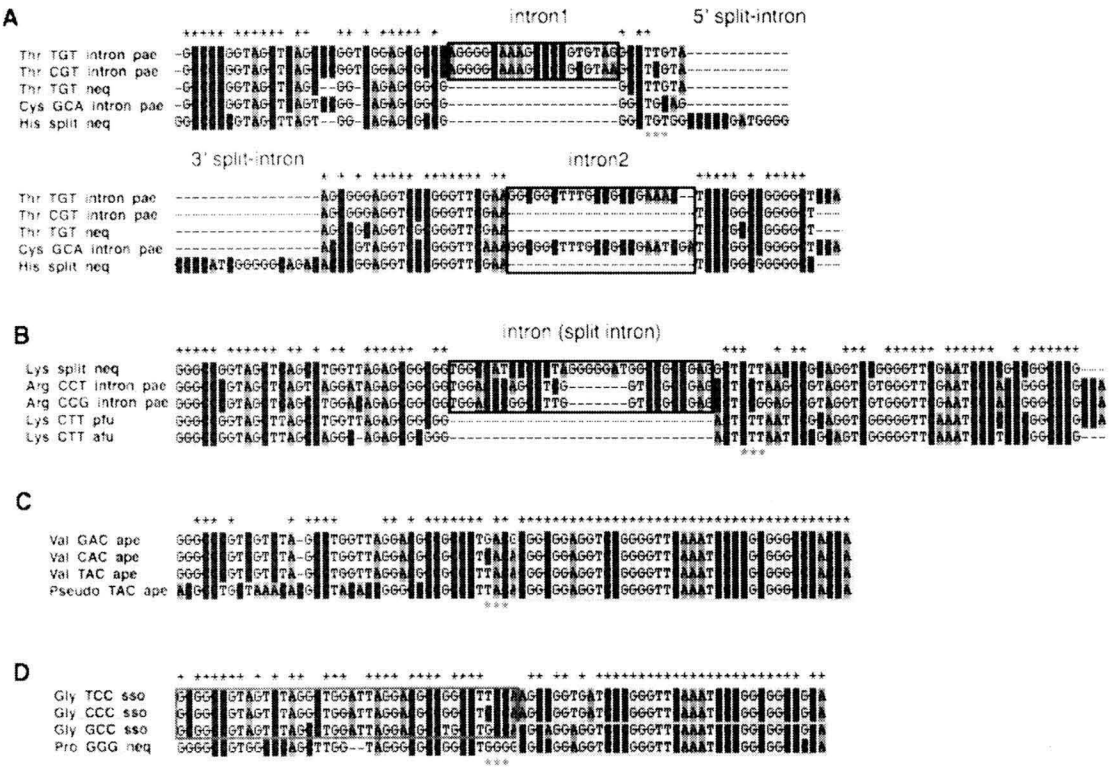


Figure 13. Multiple sequence alignment of tRNAs representing the evidence of 5' and 3' tRNA fragment combination hypothesis

Multiple alignment was performed using clustal X based on the tDNA sequence. Split-tRNAs (red bar) and intronic tRNAs (blue bar) are underlined. **(A)** Five unspliced tDNA sequences including split- intronic- and nonintronic- tRNAs with relatively conserved exon and intron sequences. **(B)** Four unspliced tDNA sequences with conserved sequence among intron and split-introns. **(C)** Four nonintronic tDNA sequences with conserved 3' halves (pink line). **(D)** Four nonintronic tDNA sequences with different conservation among the 5' halves (blue line) and 3'halves (pink line).

In Figure 13A, the double introns in the intronic Thr-tRNA are conserved in separate tRNAs, which Cys-tRNAs is known to be clustered with Histidyl split-tRNA in Figure 9. The Figure 13B represents split-intron and canonical introns share common sequence adjacent at the 3'

halves. Figure 13C and D suggests that different selection of 5' half and 3' half have occurred in the past. Especially 3' half sequences in Fig.13C is 100% conserved from the position 37/38 where 37 half starts, assuming that selection of the 5' – 3' tRNA combination happened very recently in *A. pernix* (ape).

3.3.5 A possible mechanism of the emergence and evolution of tRNAs in archaea

From the sequential point of view, the usage of different tRNA halves in various tRNAs supports the 5' – 3' tRNA combination hypothesis. Although it is important to consider the possible mechanism of how 5' and 3' tRNA are ligased and integrated back in to the genome. I assume that 5' and 3' tRNA fragments might have gone through similar mechanism of present split-tRNA and further, reverse transcribed back into the host genome as a nonintronic or intron containing tDNA sequence (Figure. 14). Several aspects support this hypothesis. As previously explained in the introduction, 3' half of tRNA is highly recognizable by replication factors such as replicase or reverse transcriptase as a genomic tag [44]. For example plant viruses with mono- or multipartite (+)-stranded RNA genomes harbors tRNA-like structure (TLS) at their 3′ -end, which can actually charged by aminoacyl-tRNA synthetases [58]. This structural motif strongly correlates with replication of viral RNA [59,60]. Further, retroplasmid derived reverse transcriptases have shown interesting features that by using this enzyme, *Escherichia coli* derived Tyr-tRNAs were reverse transcribed as mono cDNA and even hybridized cDNAs without any template DNA [43]. In fact these previous works are not based on any archaeal proteins or tRNAs further studies are necessary.

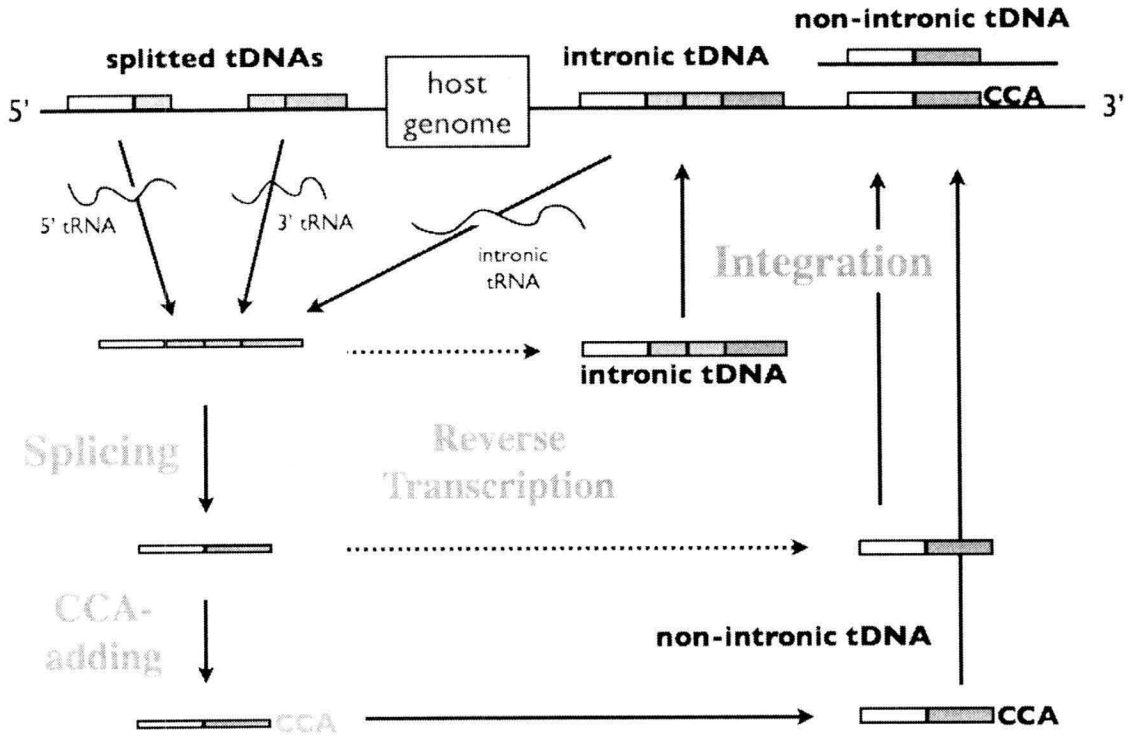


Figure 14. Schematic representation of the possible mechanism explaining the 5' – 3' tRNA combination hypothesis

The red words represents tRNA processing mechanism mediated by individual proteins already determined as a known biological process.

3.4 Conclusions

In this study, we have considered the evolutionary relationship of total 1302 tRNAs in 30 archaea (including split-tRNAs) based on comparison of the pre-processed (before splicing) and exonic (after splicing) tRNA sequences. Phylogenetic analysis has shown that exonic sequences of the split-tRNAs branched near other archaeal tRNAs with identical and synonymous anticodon encoding the same amino acids. Further analysis of 5' and 3' tRNA fragments leads to the hypothesis that the emergence of 5' and 3' tRNA halves happened

individually. The combination of 5' -3' tRNA fragment correlates strongly with the genetic code but with certain level of divergence among nanoarchaeota, crenarchaeota and euryarchaeota. Some tRNAs possess completely conserved 3' tRNA half but the sequence of the 5' region differs significantly. These results indicates that in the early stage of archaeal tRNA, a variety of 5'-3' combination has been evolved to construct the genetic code. Although try and error of tRNA fragments left an variety in the current tRNA sequences which we currently see as a network of "consensus genetic code".

Acknowledgements

Firstly, I would like to thank Prof. Akio Kanai for supervising and carrying out overall design of the study and providing the main concept of biology, which truly influenced my stand point as a scientist. I would like to give special thanks to Komasa (Mitan) Mizuki, Sayaka Kitamura, Hikaru Taniguchi, Kahori Takane, Nobuto Saito and Hiromi Toyoshima as well as other RNA members for being an great advisee as well as great partner to overcome many problems in the scientific field as a family. I also like to appreciate my parents as well as Atsuko Kishi and Yoshiteru Negishi for her/his kind support and sharing many values of life. Yuka Watanabe, Yuki Okada, Chikako Oki and Kazuhide Sekiyama have been giving me wonderful time in Tsuruoka which encouraged my motivation toward the research. Sugahara Junichi has provided tRNA prediction software SPLITS for this manucript. Suzuki Haruo participated in the statistical analysis. Asako Sato made effort in the technical assistance with the gel-shift assay. We also thank Jun Imoto, Motomu matsui, Mikiko Hattori, Nozome Yachie, Dr. Rintaro Saito and Associate. Prof. Yasuhiro Naito for their helpful discussions. Finally I would like to sincere thanks to Prof. Masaru Tomita for supervising the whole project as a lab leader and as a great OB of the Keio University.

References

1. Pruitt KD, Tatusova T & Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes transcripts and proteins. *Nucleic Acids Res* **33**, D501-504.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
3. Pazos F & Sternberg MJ (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A* **101**, 14754-14759.
4. McLaughlin WA, Kulp DW, de la Cruz J, Lu XJ, Lawson CL & Berman HM (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 SV & Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-1062.
6. Amiri H, Davids W & Andersson SG (2003) Birth and death of orphan genes in Rickettsia. *Mol Biol Evol* **20**, 1575-1587.
7. Siew N & 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 & 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 & 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 & Tomita M (2003) Cooperative regulation for Okazaki fragment processing by RNase HII & FEN-1 purified from a hyperthermophilic archaeon Pyrococcus furiosus. *Biochem Biophys Res Commun* **309**, 247-252.
11. Cotton JL & Mykles DL (1993) Cloning of a crustacean myosin heavy chain isoform: exclusive expression in fast muscle. *J Exp Zool* **267**, 578-586.
12. Laskin AA, Kudryashov NA, Skryabin KG & Korotkov EV (2005) Latent periodicity of serine-threonine & tyrosine protein kinases & other protein families. *Comput Biol Chem* **29**, 229-243.
13. Bhardwaj N, Langlois RE, Zhao G & Lu H (2005) Kernel-based machine learning

protocol for predicting DNA-binding proteins. *Nucleic Acids Res* **33**, 6486-6493.

14. Han LY, Cai CZ, Lo SL, Chung MC & Chen YZ (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *Rna* **10**, 355-368.
15. Cai YD & Lin SL (2003) Support vector machines for predicting rRNA- RNA- and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta* **1648**, 127-133.
16. Ofran Y& Margalit H (2006) Proteins of the same fold and unrelated sequences have similar amino acid composition. *Proteins* **64**, 275-279
17. Xie D, Li A, Wang M, Fan Z & Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res* **33**, W105-110.
18. Sarda D, Chua GH, Li KB & Krishnan A (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics* **6**, 152.
19. Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB & Dunn DM (2001) Genomic sequence of hyperthermophile *Pyrococcus furiosus*: implications for physiology & enzymology *Methods Enzymol* **330**, 134-157.
20. Bairoch A & Apweiler R (1996) The SWISS-PROT protein sequence data bank & its new supplement TREMBL. *Nucleic Acids Res* **24**, 21-25
21. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A & Apweiler R (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT TrEMBL and InterPro. *Genome Res* **13**, 662-672.
22. Pavlidis P, Wapinski I, Noble WS (2004) Support vector machine classification on the web. *Bioinformatics* **20,** 586-587.
23. Cai CZ, Han LY, Ji ZL, Chen X & Chen YZ (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.
24. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F & Zdobnov EM (2000) InterPro--an integrated documentation resource for protein families domains and functional sites. *Bioinformatics* **16**, 1145-1150
25. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S & Eddy SR (2004) The Pfam protein families database. *Nucleic*

Acids Res **32**, D138-141.

26. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS & Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**, 633-638.
27. Turnage MA, Brewer-Jensen P, Bai WL & Searles LL (2000) Arginine-rich regions mediate the RNA binding & regulatory activities of the protein encoded by the Drosophila melanogaster suppressor of sable gene. *Mol Cell Biol* **20**, 8198-8208.
28. Shanahan HP, Garcia MA, Jones S & Thornton JM (2004) Identifying DNA-binding proteins using structural motifs & the electrostatic potential. *Nucleic Acids Res* **32**, 4732-4741.
29. Ikegami T, Kuraoka I, Saijo M, Kodo N, Kyogoku Y, Morikawa K, Tanaka K & Shirakawa M (1998) Solution structure of the DNA- & RPA-binding domain of the human repair factor XPA. *Nat Struct Biol* **5**, 701-706.
30. Bayer TS, Booth LN, Knudsen SM & Ellington AD (2005) Arginine-rich motifs present multiple interfaces for specific binding by RNA. *Rna* **11**, 1848-1857.
31. Garcia-Garcia C & Draper DE (2003) Electrostatic interactions in a peptide--RNA complex. *J Mol Biol* **331**, 75-88.
32. Ahmad S & Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics **6**, 33.
33. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *Rna* **8**, 1450-1462.
34. Wang L, Brown SJ. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **34**, W243-8.
35. Anantharaman V, Koonin EV & Aravind L (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* **30**, 1427-1464.
36. Manival X, Ghisolfi-Nieto L, Joseph G, Bouvet P & Erard M (2001) RNA-binding strategies common to cold-shock domain- and RNA recognition motif-containing proteins. *Nucleic Acids Res* **29**, 2223-2233.
37. Gatherer D & McEwan NR (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.
38. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442-451.
40. Ramakrishnan V & Adams MWWW (1995) Preparation of Genomic DNA from

sulfur-dependant hyperthermophilic Archaea. Archaea: A Laboratory Manual 95-99.

41. Nagaswamy, U & G.E. Fox (2003) RNA ligation and the origin of tRNA. *Orig Life Evol Biosph* **33:** 199-209.
42. Hipps, D., K. Shiba, B. Henderson & P. Schimmel. (1995) Operational RNA code for amino acids: species-specific aminoacylation of minihelices switched by a single nucleotide. *Proc Natl Acad Sci U S A* **92:** 5550-5552.
43. Chiang, C.C & A.M. Lambowitz (1997) The Mauriceville retroplasmid reverse transcriptase initiates cDNA synthesis de novo at the 3' end of tRNAs. *Mol Cell Biol* **17:** 4526-4535.
44. Maizels, N & A.M. Weiner (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci U S A* **91:** 6729-6734.
45. Weiner, A.M & N. Maizels (1999) The genomic tag hypothesis: modern viruses as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis. *Biol Bull* **196:** 327-328; discussion 329-330.
46. Randau, L., R. Munch, M.J. Hohn, D. Jahn & D. Soll (2005) Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. *Nature* **433:** 537-541.
47. Marck, C & H. Grosjean. (2003) Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *Rna* **9:** 1516-1531.
48. Randau, L., M. Pearson & D. Soll (2005) The complete set of tRNA species in Nanoarchaeum equitans. *FEBS Lett* **579:** 2945-2947.
49. Sugahara, J., Yachie. N., Sekine. Y., Soma. A., Matsui. M., Tomita. M & Kanai A (2006) SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level. *In Silico Biol* **5**: 411-418.
50. Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin & D.G. Higgins (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.
51. Hall, B (2005) Phylogenetic Trees Made Easy: A How-To Manual 28-30.
52. Swofford, D. L. (1998). PAUP. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.
53. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **8**: 754-755.

54. Nylander, J. A. A. (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

55. Nikolajewa, S., M. Friedel, A. Beyer & T. Wilhelm (2006) THE NEW CLASSIFICATION SCHEME OF THE GENETIC CODE, ITS EARLY EVOLUTION, AND tRNA USAGE. *J Bioinform Comput Biol* **4:** 609-620.

56. Marck, C & H. Grosjean (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *Rna* **8:** 1189-1302.

57. Brochier C, Forterre P & Gribaldo S (2005) An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol Biol.* **5**: 36.

58. Dreher, T.W (1999) Functions of the 3'-Untranslated Regions of Positive Strand Rna Viral Genomes. *Annu Rev Phytopathol* **37:** 151-174.

59. Rudinger-Thirion, J., R.C. Olsthoorn, R. Giege, & S. Barends (2006) Idiosyncratic behaviour of tRNA-like structures in translation of plant viral RNA genomes. *J Mol Biol* **355:** 873-878.

60. Brown, D.M., C.T. Cornell, G.P. Tran, J.H. Nguyen & B.L. Semler (2005) An authentic 3' noncoding region is necessary for efficient poliovirus replication. *J Virol* **79:** 11962-11973.

APPENDIX

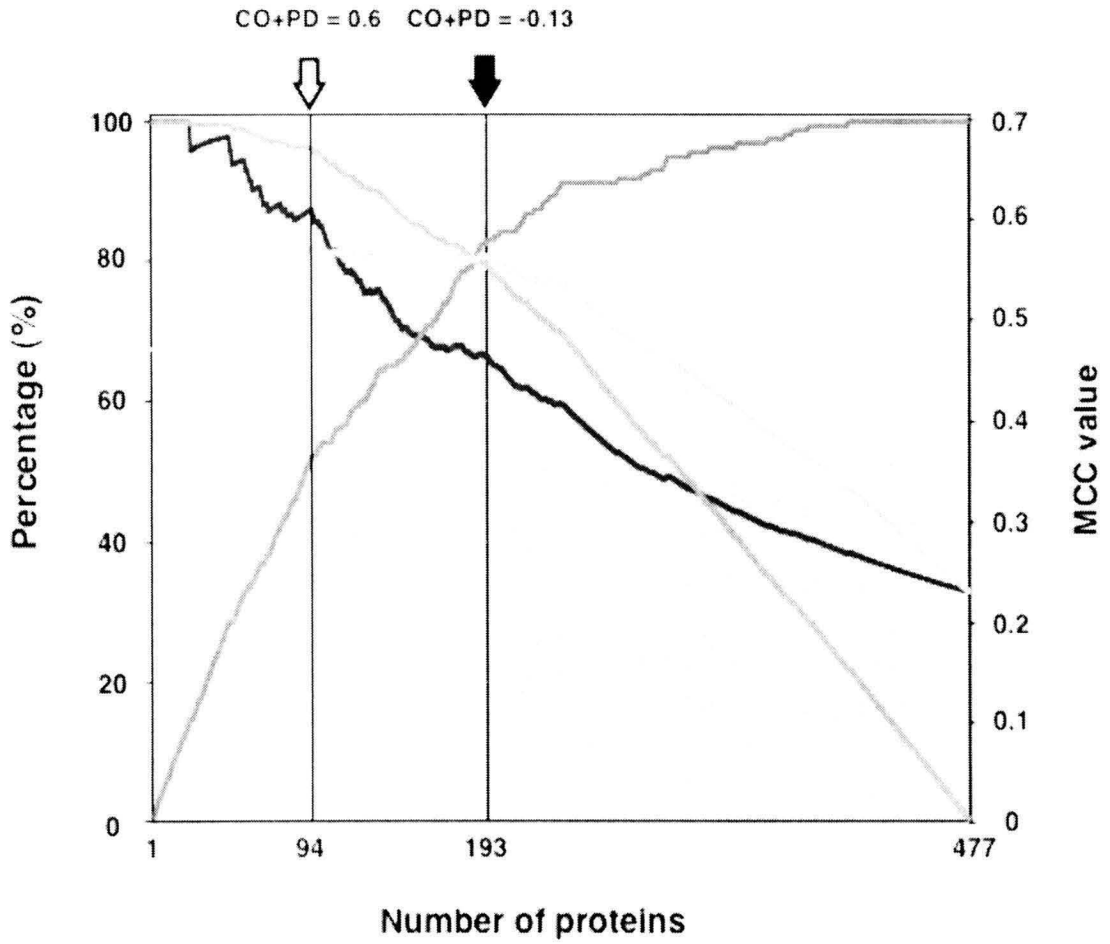
Supplementary Tables

Supplementary Table 1. Summery of 10 novel DNA/RNA-binding proteins in *P. furiosus*

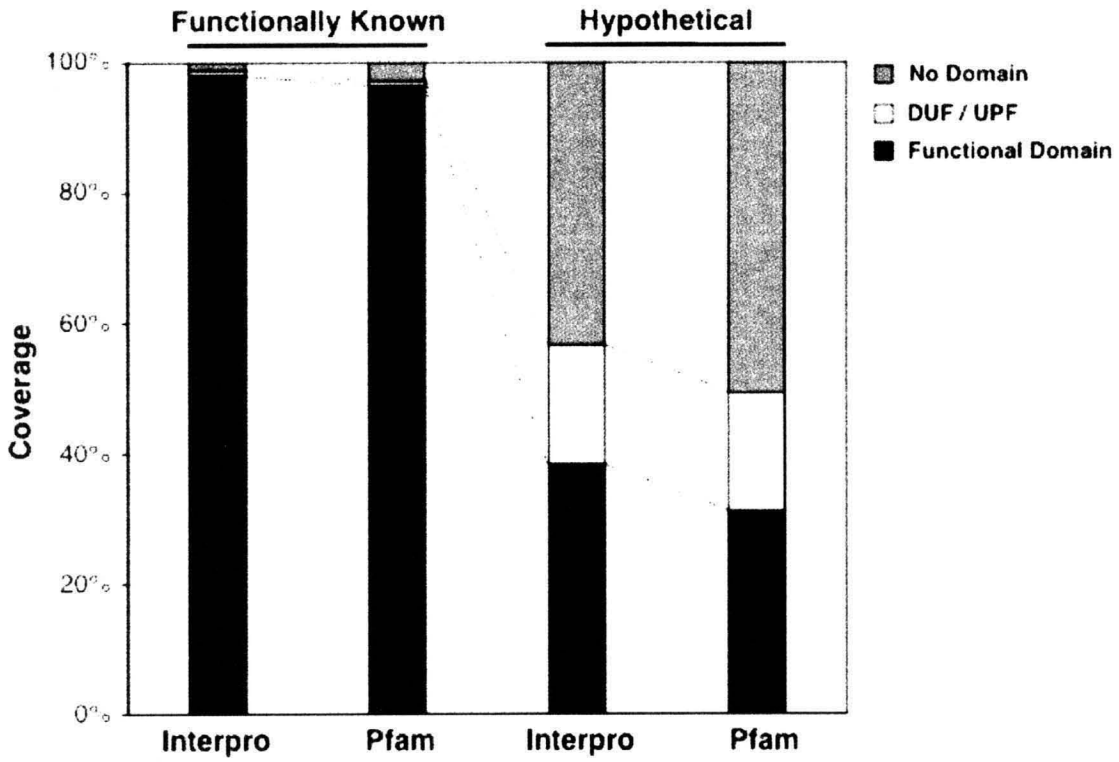
***P. furiosus* Gene ID**	**Pfam Domain**	**NCBI BLASTP search hit (*Species*)**	**E-value**
PF1912	Radical_SAM, TRAM, UPF0004	UPF0004 protein PH1875 (*Pyrococcus horikoshii*)	0.0
PF1139	THUMP	Hypothetical protein MJ0041 (*Methanococcus jannaschii*)	1.00E-81
PF1580	KH_1	Hypothetical protein MJ0443 (*Methanococcus jannaschii*)	2.00E-37
PF1498	CRS1_YhbY	UPF0044 protein MJ0652 (*Methanococcus jannaschii*)	9.00E-11
PF1473		Hypothetical protein MJ1211 (*Methanococcus jannaschii*)	2.00E-09
PF2062		Hypothetical protein PYRAB14350 precursor (*Pyrococcus abyssi*)	8.00E-07
PF0030	DUF1611	Homoserine O-acetyltransferase (*Burkholderia pseudomallei*)	0.3
PF0565		Proteasome activator complex subunit 1 (*Rattus norvegicus*)	0.7
PF0029		1-deoxy-D-xylulose-5-phosphate synthase (*Haemophilus ducreyi*)	0.7
PF1981	NTP_transf_2	Exopolysaccharide production protein exoQ (*Sinorhizobium meliloti*)	1.3

BLASTP search and Pfam domain annotation is performed on 10 experimentally identified novel DNA/RNA-binding proteins. The functional annotation of the closest swissprot hit is shown and sorted with relative E-value. Pfam domain is assigned due to the Uniprot protein data.

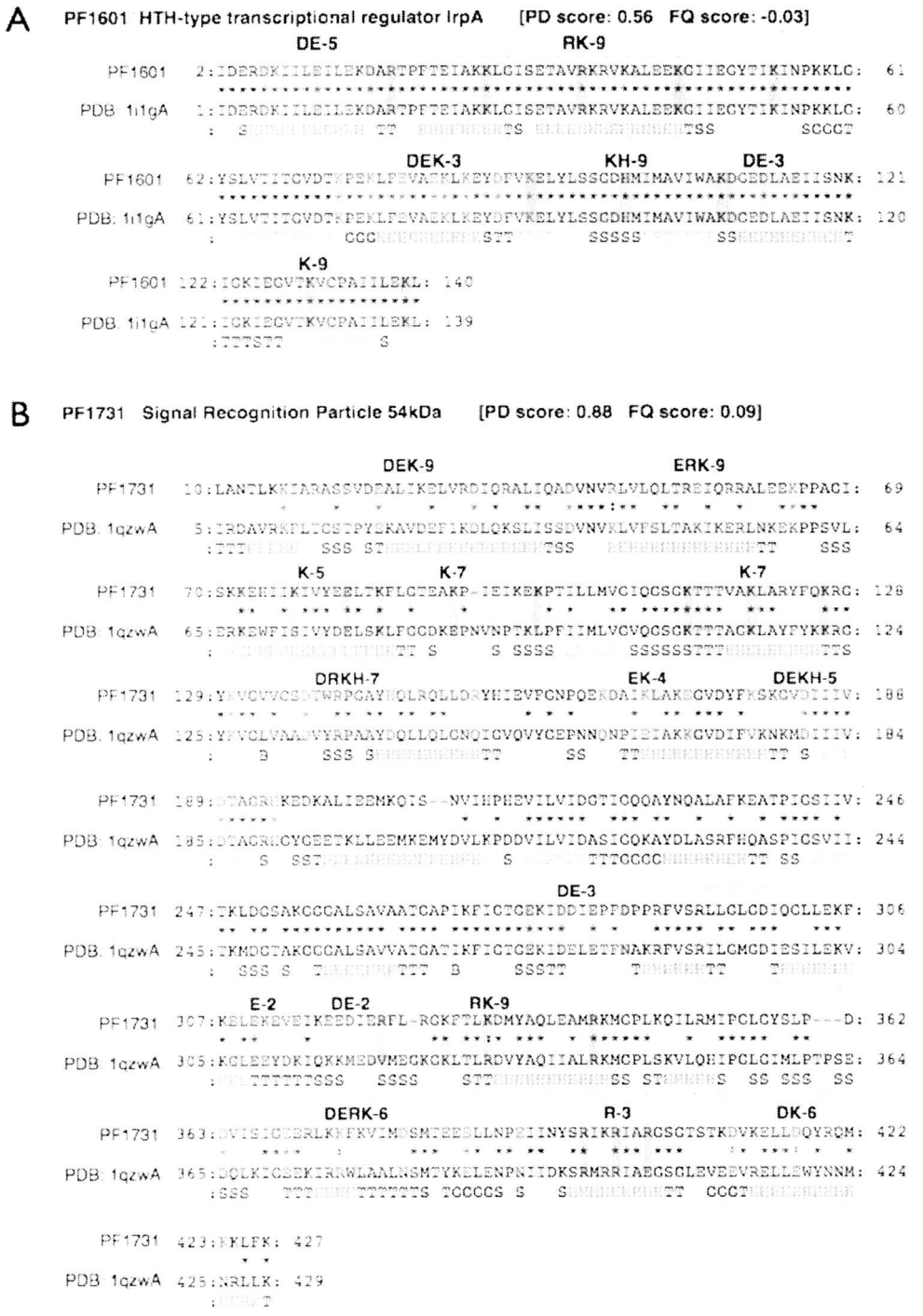
Supplementary Figures



Supplementary Figure 1. Five index values and two thresholds for DNA/RNA-binding protein classification. The left vertical axis represents the percentage (%) of four indices, positive predictive value: PPV (blue), sensitivity: SE (red), specificity: SP (orange) and overall accuracy: ACC (green), shown in lines. The right vertical axis represents the value of Matthews correlation coefficient (MCC) shown in gray bars. These indices were calculated at every CO+PD score of the sorted 477 functionally known proteins. Two thresholds (dotted lines) were determined at maximum ACC and MCC value.



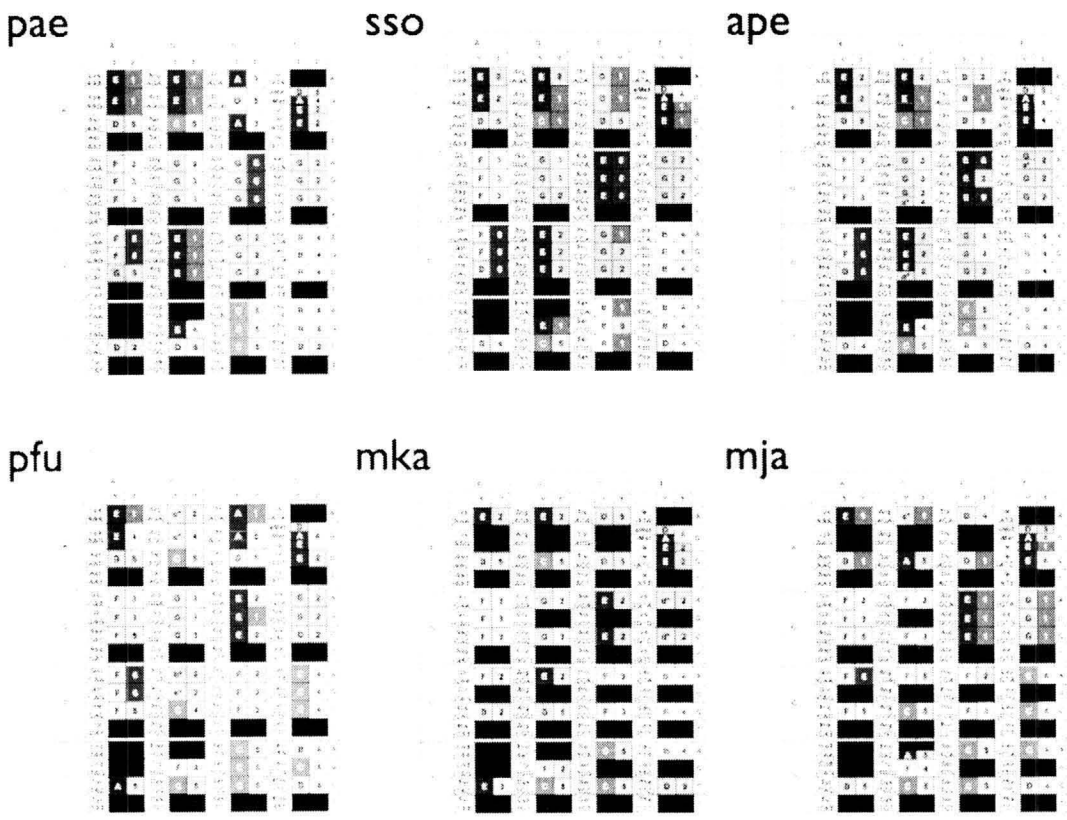
Supplementary Figure 2. Domain analysis of *P. furiosus* proteome. Histogram showing the coverage of three domain categories for 942 functionally known proteins and 994 hypothetical proteins annotated by Interpro and Pfam domain databases respectively. Three categories are proteins with functionally known domains (black), proteins with DUF: domains of unknown functions or UPF: unknown protein functions (white) and proteins with no domain (gray).



Supplementary Figure 3. Amino acid periodicity in the sequences of proteins with high PD score and low CO score.

The distribution of amino acid periodicities with structural features were observed in three class I proteins with high PD score and low CO score (PD score > 0.25 and CO score < 0.5)

through sequence alignment with closest structural orthologues registered in Proteins Data Bank (PDB). The periodic regions consists of overall region (colored squares) and amino acids corresponding with periodicities (oblong box). Periodicity includes error range ±1. The conserved amino acids between *P. furiosus* protein and the structurally known proteins with same functions are marked with asterisks. (A) HTH-type transcriptional regulator lrpA (PF1601) is aligned with its own structure (PDB_ID: 1i1gA) with 100% identity. DNA-binding motif (Helix-turn-Helix) are framed with red dashed line. (B) Signal Recognition Particle 54kDa (PF1731) is aligned with orthologous protein (srp54) in *S. solfataricus* (PDB_ID: 1qzwA) with 47.1% identity. DNA-binding domain (M domain) are framed with red dashed line.



Supplementary Figure 4. Codon tables of the six archaeal species filled with 5' and 3' tRNA fragments.

Each codon in the codon tables are filled with the corresponding cluster ID of 5' and 3' tRNA halves (see Figure 9B). The codon table on the upper region represents Crenarchaeota (pae, sso and ape) and the codon table on the lower region represents Euryarchaeota (pfu, mka and mja).

Prediction of novel DNA/RNA-binding proteins and
analysis of tRNA evolution using genomic data

2007年3月30日　　初版発行

著者　藤島皓介

監修　冨田　勝

発行　慶應義塾大学 湘南藤沢学会
〒252-0816　神奈川県藤沢市遠藤5322
TEL:0466-49-3437

Printed in Japan　　印刷・製本　　ワキプリントピア

ISBN 978－4－87762－175－9
SFC-MT 2006-005

■本論文は修士論文において優秀と認められ、出版されたものです。