Comparison and Characterization of Proteomes in the Three Domains of Life Using 2D Correlation Analysis

Kosuke Fujishima,1,2 Mizuki Komasa,1,2 Sayaka Kitamura,1,2 Masaru Tomita1,2,3 and Akio Kanai1,2,3

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

Proteins are a major regulatory component in complex biological systems. Among them, DNA/RNA-binding proteins, the key components of the central dogma of molecular biology, and membrane proteins, which are necessary for both signal transduction and metabolite transport, are suggested to be the most important protein families that arose in the early stage of life. In this study, we computationally analyzed the whole proteome data of six model species to overview the protein diversity in the three domains of life (Bacteria, Archaea and Eukaryota), especially focusing on the above two protein families. To compare the protein distribution among the six model species, we calculated various protein profiles: hydropathy, molecular weight, amino acid composition and periodicity for each protein. We found a domain-specific distribution of the proteome based on 2D correlation analysis of hydropathy and molecular weight. Further, the merged protein distribution of Archaea and other domains revealed many membrane proteins localized in Bacteria-specific regions with a high ratio of hydropathy and many DNA/RNA-binding proteins localized in Eukaryota-specific regions with a low ratio of hydropathy. Since about half of the proteins encoded in the genome are still functionally unknown, we further conducted Support Vector Machine (SVM)-based functional prediction using amino acid composition (CO score) and periodicity (PD score) as feature vectors to predict the overall number of DNA/RNA-binding proteins and membrane proteins in the proteome. Our estimation indicated that two functional categories occupy approximately 60% to 80% of the proteome, and further, the proportion of the two categories varied among the three domains of life, suggesting that the proteome has gone through different selective pressure during evolution.

§1. Introduction

The main flow of biological information is transferred from DNA to RNA to protein, the so-called central dogma of molecular biology, and all living organisms share this universal system. We are now in the middle of the post-genomic era, with an enormous number of DNA sequences from over 240,000 named organisms,1) and the numbers are rapidly increasing day by day. Accompanied by the accumulation of genomic data, RNAs in over 400 complete genomes (Rfam database)2) and protein sequences from 9,318 families (Pfam database)3) are currently available. Protein molecules are the major regulatory component of the biological system. The model species Escherichia coli is known to possess approximately 4,400 proteins, and the nematode Caenorhabditis elegans possesses approximately 20,000 proteins. These estimated numbers of the proteome are deduced from predicted genes in the genome sequence, and yet about half of the proteins registered in protein databases are still
Comparison and Characterization of Proteomes in the Three Domains

categorized as functionally unknown (hypothetical), because they lack similarity to any of the functionally known proteins. These functionally unknown proteins include products from species-specific genes (so-called ORFans) that are known to encode many uncharacterized short peptides\(^4\) thus, their functions are difficult to predict with certainty using comparative genomics.

The existence of mass functionally unknown proteins in the proteome field cannot be ignored in understanding the complex interaction of cellular molecules. Therefore, a fast and accurate proteome-wide function classification method is expected to be of great benefit in the field of proteome and systems biology. The core system of life, the central dogma of molecular biology, consists of consecutive reaction of various DNA/RNA-binding proteins (polymerases, transcription factors, translation factors, ribosomes, etc.) and key functional RNAs such as transfer RNA and ribosomal RNA. Further, many DNA/RNA-binding proteins exist for cutting, repair, modification and transfer of DNA and RNA; thus, a few recent studies have begun to reveal the complex biological network of proteins and nucleotides.\(^5\) Moreover, evidence of a large proportion of transcribed non-coding RNAs has been discovered by transcriptome analysis in \(E.\ coli\), mouse and human.\(^6\)--\(^8\) The biological function of most of the non-coding RNA is yet to be discovered, and there lies a possibility that novel RNA-binding proteins could be interacting with these RNAs to form an unknown regulatory network in the cell. Therefore, detection of novel DNA/RNA-binding proteins from functionally unknown proteins has topical importance and could lead to finding a novel biological mechanism. For the past few years, we have been working on RNA metabolism in the hyperthermophilic archaeon \(Pyrococcus furiosus\) and reported our experimental system in which an expression cloning method was used for extracting DNA/RNA-binding proteins at the proteome level.\(^9\),\(^10\) Furthermore, we have recently developed a fast, accurate and comprehensive method to predict novel DNA/RNA-binding proteins from functionally unknown proteins using amino acid composition and periodicity.\(^11\) Based on this research, in this paper we performed a proteome-wide comparison and characterization by expanding the target from Archaea to the three domains of life (\(Nanoarchaeum equitans\), \(Methanopyrus kandleri\), \(E.\ coli\) K12 MG1655, \(Synechocystis\) strain PCC 6803, \(Caenorhabditis elegans\) and \(Homo sapiens\)) using various protein profiles: hydropathy, molecular weight, amino acid composition and periodicity. We visualized the proteome diversity among the six model species using a 2D distribution plot based on hydropathy and molecular weight. The two important protein families, DNA/RNA-binding and membrane proteins, were distributed in two distinct regions with domain-specific expansion. The proportion of the two functional families occupied approximately 60% to 80% in the proteome, with wide variation among the three domains of life.

\(\S\) 2. Materials and methods

2.1. Protein dataset and functional annotations

Automated annotations and amino acid sequences of proteins in the six model species; \(N.\ equitans\), \(M.\ kandleri\), \(E.\ coli\) K12 MG1655, \(Synechocystis\) (strain PCC
C. elegans and H. sapiens, were taken from the UniProt KnowledgeBase (UniProtKB) accession code corresponding to its entry in either UniProtKB/Swiss-Prot (http://www.ebi.ac.uk/swissprot/: Release 54, December 2007) or UniProtKB/TrEMBL (http://www.ebi.ac.uk/trembl/: Release 37, July 2007). 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\(^\text{12}\) and Pfam\(^\text{13}\). Both the Swiss-Prot and TrEMBL databases were used for two archaeal species; N. equitans (536 proteins) and M. kandleri (1,700 proteins), and Swiss-Prot data alone were used for the four prokaryotic and eukaryotic species; E. coli K12 MG1655 (4,930 proteins), Synechocystis PCC 6803 (3,527 proteins), C. elegans (3,040 proteins) and H. sapiens (7,012 proteins) as a reliable independent test set.

### 2.2. Calculation of protein profiles and definition of PD score and CO score

Amino acid hydropathy score was extracted from Kyte and Doolittle's work\(^\text{14}\) and all scores were summed to calculate the hydropathy score of an individual protein. For amino acid composition, we calculated the relative composition of amino acids \((C)\) 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 \(i\). 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 x 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 the periodic region of periodicity \(j\) in a single protein \(i\), and \(N\) is the full amino acid length of a single protein \(i\). Two heterogeneous profiles; amino acid periodicity and amino acid composition, were integrated by the late integration method. The SVM output (discriminant values) was calculated for each profile as amino acid periodicity (PD score) and amino acid composition (CO score) and later summed as CO+PD score. CO score and PD score were calculated by the same procedure as in a previous study\(^\text{11}\). Thresholds for extracting DNA/RNA-binding proteins were determined by considering several indices. The index, positive predictive value (PPV) was adapted to measure the percentage of DNA/RNA-binding proteins among proteins above the threshold. PPV is calculated by the equation:

\[
PPV = \frac{TP}{TP + FP} \times 100\ (%).
\]

Another index, Matthews correlation coefficient (MCC), was used to determine different threshold values of the CO+PD score. MCC is a popular index for measuring the performance of prediction, and maximum MCC provides high sensitivity and
Comparison and Characterization of Proteomes in the Three Domains

specificity. MCC is calculated by the equation:

\[
MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)\cdot(TP + FN)\cdot(TN + FP)\cdot(TN + FN)}}
\]

taking the DNA/RNA-binding protein classification as an example, \( TP \) refers to true positives (number of DNA/RNA-binding proteins with CO+PD score > cut-off), \( FP \) refers to false positives (number of other proteins with CO+PD score > cut-off), \( FN \) refers to false negatives (number of DNA/RNA-binding proteins with CO+PD score < cut-off), and \( TN \) refers to true negatives (number of other proteins with CO+PD score < cut-off).

§3. Results and discussion

3.1. Comparison of protein distribution in six model species based on 2D plot of hydropathy and molecular weight

A total of six model species; \( N.\ equitans \), \( M.\ kandleri \), \( E.\ coli \), \( Synechocystis \), \( C.\ elegans \) and \( H.\ sapiens \), were used in this study. The two archaeal species \( N.\ equitans \) and \( M.\ kandleri \) are assumed to be two of the most primitive species and were chosen for their topical importance in the evolution of the ancient architecture of the biological system.\(^{15}\) The cyanobacterium \( Synechocystis \) was chosen for its old lineage in the history of the earth, in which putative fossilized cyanobacteria have been found from 3.8 billion years ago. \( E.\ coli \) was chosen as the most well-known standard model bacterium. \( C.\ elegans \) and \( H.\ sapiens \) were chosen as representatives of well-known multi-cellular species. The general concept of the three domains of life was first proposed by Carl Woese,\(^{16}\) separating all living organisms into three large domains (Bacteria, Archaea and Eukaryota). The relative phylogenetic positions of the six model species are shown in Fig. 1. To characterize the proteins in each species at the proteome level, we introduced four protein profiles: hydropathy, molecular weight,

![Fig. 1. Schematic representation of three domains of life.](image)

The phylogenetic tree of the three domains of life (Archaea, Bacteria and Eukaryota) is represented.\(^{16}\) The names of the six model species are placed at appropriate positions in the tree.
amino acid composition and amino acid periodicity. Comparison of the distribution of the whole proteome among the six model species was conducted by visualizing the proteome data on a 2D plot based on hydropathy and molecular weight (Fig. 2A). The distribution of proteins expanded according to the total number of proteins that each species possessed. For example, most of the proteins in Archaea possess a molecular weight below 75kDa and hydropathy score between 400 and -200, whereas bacterial proteins show slight expansion with molecular weight below 100kDa and hydropathy score between 600 and -200, while proteins in Eukaryota show a wide range of diversity with molecular weight up to 250kDa and hydropathy ranging from 600 to -1000. To define the area of protein expansion in each species, we formulated two distinct linear borderlines using the top 50 proteins with highest and lowest hydropathy: \( M(\frac{H}{M}) \) and presented the fitted lines in Fig. 2A. We found that the slope of the upper borderline (highest \( \frac{H}{M} \)) in Archaea and Eukaryota possessed certain similarity (from 0.0087 to 0.0099), but differed in the two Bacteria: \( E. coli \) (0.0125) and \( Synechocystis \) (0.0119). On the other hand, the slope of the lower borderline (lowest \( \frac{H}{M} \)) in Eukaryota: \( C. elegans \) (-0.097) and \( H. sapiens \) (-0.119) was significantly lower than that in Archaea and Bacteria (from -0.0044 to -0.0076), suggesting that a certain class of proteins may have evolved and diverged domain-specifically. To visualize proteins distributed in a domain-specific manner, we chose the proteome of the archaeon \( M. kandleri \) as a basis of protein distribution, and merged it with the \( E. coli \) proteome and \( H. sapiens \) proteome (Fig. 2B). As a result, the protein distribution in \( E. coli \) differed in the area surrounding the upper borderline, representing a specific group of proteins possessing high \( \frac{H}{M} \) (‘group A’ in Fig. 2B), while the protein distribution of \( H. sapiens \) showed a specific distribution around the lower borderline possessing low \( \frac{H}{M} \) (‘group B’ in Fig. 2B).

3.2. Prediction of DNA/RNA-binding proteins and membrane proteins based on 2D correlation analysis of CO and PD score

To estimate precise numbers of DNA/RNA-binding and membrane proteins in the proteome of the six model species, we further performed proteome-wide function prediction using two protein profiles (amino acid composition and amino acid periodicity). We previously used a Support Vector Machine (SVM)-based method using functionally known proteins as a training set, and quantitatively evaluated DNA/RNA-binding proteins in \( P. furiosus \).\(^{11} \) The same procedure was applied, and two discriminant values derived by SVM, the so-called composition score (CO score) and periodicity score (PD score), were calculated for the six model species. We summed CO score and PD score by the late integration method: a method for integrating heterogeneous datasets based on the context of SVM learning.\(^{17} \) Then, two different thresholds were determined based on the high positive predictive value (PPV) and highest Matthews correlation coefficient (MCC), separating proteins into three categories: Class I - Class III.

First, by introducing the three categories in the 2D plot based on CO score and PD score, we were able to visualize accurate classification of DNA/RNA-binding proteins in the six species (Fig. 3). Significantly, in \( N. equitans \), the positive pre-
Fig. 2. **Comparison of protein distribution in six model species at proteome level.**

2D correlation analysis based on hydropathy and molecular weight was performed on proteome data of the six model species. (A) Fifty proteins with highest hydropathy per molecular weight (blue circles) and lowest hydropathy per molecular weight (green circles) were plotted with the rest of the proteins (red circles) for each model species. Two borderlines and their equations represent the fitted lines of 50 proteins with highest and lowest hydropathy per molecular weight (H/M). (B) Comparison of the protein distribution between two species from different domains. Merged plot of *M. kandleri* (black circles) and two other species *E. coli* and *H. sapiens* (red circles) are shown. Proteins plotted beyond the borderlines of *M. kandleri* are grouped as ‘group A’ and ‘group B’. (See the online edition for the color version of this figure.)
Total 15 membrane proteins from ‘group A’ in *E. coli* and 15 DNA/RNA-binding proteins from ‘group B’ in *H. sapiens* (see Fig. 2B) are represented.

Table 1. Examples of functionally known proteins localized in species-specific regions

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Function</th>
<th>Mol Weight (M)</th>
<th>Hydropathy (H)</th>
<th>H/M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples of membrane proteins in ‘group A’ in <em>E. coli</em> specific region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>znuB</td>
<td>High-affinity zinc uptake protein</td>
<td>27713.0</td>
<td>344.3</td>
<td>0.012</td>
</tr>
<tr>
<td>soiB</td>
<td>Sugar efflux transporter</td>
<td>42523.2</td>
<td>520.2</td>
<td>0.012</td>
</tr>
<tr>
<td>btaC</td>
<td>Vitamin B12 import system permease</td>
<td>34934.7</td>
<td>419.4</td>
<td>0.012</td>
</tr>
<tr>
<td>manY</td>
<td>Mannose permease IIC component</td>
<td>27620.4</td>
<td>330.8</td>
<td>0.012</td>
</tr>
<tr>
<td>gntP</td>
<td>High-affinity gluconate transporter</td>
<td>47124.0</td>
<td>555.1</td>
<td>0.012</td>
</tr>
<tr>
<td>fepD</td>
<td>Ferric enterobactin transport system permease</td>
<td>33856.8</td>
<td>393.4</td>
<td>0.012</td>
</tr>
<tr>
<td>idnT</td>
<td>Gnt-II system L-idonate transporter</td>
<td>46026.7</td>
<td>532.1</td>
<td>0.012</td>
</tr>
<tr>
<td>dcuC</td>
<td>Anaerobic C4-dicarboxylate transporter</td>
<td>48398.3</td>
<td>558.9</td>
<td>0.012</td>
</tr>
<tr>
<td>gntU</td>
<td>Low-affinity gluconate transporter</td>
<td>46401.6</td>
<td>527.6</td>
<td>0.011</td>
</tr>
<tr>
<td>dsdX</td>
<td>DsdX permease</td>
<td>47149.0</td>
<td>534.3</td>
<td>0.011</td>
</tr>
<tr>
<td>arsB</td>
<td>Arsenical pump membrane protein</td>
<td>45431.8</td>
<td>512.2</td>
<td>0.011</td>
</tr>
<tr>
<td>glcA</td>
<td>Glycolate permease</td>
<td>58905.9</td>
<td>662.6</td>
<td>0.011</td>
</tr>
<tr>
<td>nepl</td>
<td>Purine ribonucleoside efflux pump</td>
<td>41826.9</td>
<td>469.6</td>
<td>0.011</td>
</tr>
<tr>
<td>fmsB</td>
<td>Ferrichrome transport system permease</td>
<td>70411.3</td>
<td>785.2</td>
<td>0.011</td>
</tr>
<tr>
<td>entS</td>
<td>Enterobactin exporter</td>
<td>43268.0</td>
<td>481.8</td>
<td>0.011</td>
</tr>
<tr>
<td>tyrP</td>
<td>Tyrosine-specific transport protein</td>
<td>42804.2</td>
<td>475.8</td>
<td>0.011</td>
</tr>
</tbody>
</table>

| **Examples of DNA/RNA-binding proteins in ‘group B’ in *H. sapiens* specific region** |
| ZRANB2 | Zinc finger Ran-binding protein 2 | 37390.5 | -558.2 | -0.015 |
| RNPS1 | RNA-binding protein with serine-rich domain 1 | 34194.0 | -464.4 | -0.014 |
| TRA2A | Transformer-2 protein homolog | 32674.0 | -433.8 | -0.013 |
| FUSIP1 | FUS-interacting serine-arginine-rich protein 1 | 31285.3 | -405.6 | -0.013 |
| EIF4B | Eukaryotic translation initiation factor 4B | 69213.4 | -867.6 | -0.013 |
| RSRC1 | Arginine/serine-rich coiled coil protein 1 | 38663.2 | -472.0 | -0.012 |
| TCEAL2 | Transcription elongation factor A protein-like 2 | 25834.3 | -308.9 | -0.012 |
| RBMY1A1 | RNA-binding motif protein | 55771.8 | -643.1 | -0.012 |
| SRRP130 | Splicing factor, arginine/serine-rich 130 | 92568.4 | -927.3 | -0.010 |
| FIP1L1 | Pre-mRNA 3’-end-processing factor FIP1 | 66515.3 | -656.2 | -0.010 |
| UPF3A | Regulator of nonsense transcripts 3A | 54683.8 | -516.4 | -0.009 |
| PRPF38A | Pre-mRNA-splicing factor 38A | 37462.1 | -352.7 | -0.009 |
| GPATCH4 | G patch domain-containing protein 4 | 50369.2 | -462.7 | -0.009 |
| NCL | Nucleolin | 76604.1 | -701.7 | -0.009 |
| GTF2F1 | Transcription initiation factor IIF subunit alpha | 58228.3 | -526.2 | -0.009 |

Total 15 membrane proteins from ‘group A’ in *E. coli* and 15 DNA/RNA-binding proteins from ‘group B’ in *H. sapiens* (see Fig. 2B) are represented.
Fig. 3. 2D correlation analysis of DNA/RNA-binding proteins in six model species based on amino acid composition and periodicity. Functionally known proteins in the six model species are plotted on a 2D correlation plot of CO score and PD score. Ribosomal proteins (red), rest of the DNA/RNA-binding proteins (blue) and other functionally known proteins (black) are shown. The two lines represent the threshold of optimum PPV value for accurate prediction of DNA/RNA-binding protein candidates (white arrow) and maximum MCC value for optimizing sensitivity and specificity (black arrow). The ranges of the three classes (Class I - Class III) are defined for each species. (See the online edition for the color version of this figure.)
Table 2. Summary of DNA/RNA-binding protein classification in six model species

<table>
<thead>
<tr>
<th></th>
<th>Functionally known proteins</th>
<th>Hypothetical proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ribosome</td>
<td>DNA/RNA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N. equitans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>Class II</td>
<td>13</td>
<td>65</td>
</tr>
<tr>
<td>Class III</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>156</td>
</tr>
<tr>
<td><strong>M. kandleri</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>36</td>
<td>95</td>
</tr>
<tr>
<td>Class II</td>
<td>23</td>
<td>159</td>
</tr>
<tr>
<td>Class III</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>61</td>
<td>314</td>
</tr>
<tr>
<td><strong>E. coli</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>39</td>
<td>185</td>
</tr>
<tr>
<td>Class II</td>
<td>25</td>
<td>293</td>
</tr>
<tr>
<td>Class III</td>
<td>13</td>
<td>267</td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>745</td>
</tr>
<tr>
<td><strong>Synechocystis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>38</td>
<td>145</td>
</tr>
<tr>
<td>Class II</td>
<td>15</td>
<td>139</td>
</tr>
<tr>
<td>Class III</td>
<td>4</td>
<td>106</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>390</td>
</tr>
<tr>
<td><strong>C. elegans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>45</td>
<td>132</td>
</tr>
<tr>
<td>Class II</td>
<td>13</td>
<td>229</td>
</tr>
<tr>
<td>Class III</td>
<td>12</td>
<td>256</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>617</td>
</tr>
<tr>
<td><strong>H. sapiens</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>24</td>
<td>503</td>
</tr>
<tr>
<td>Class II</td>
<td>99</td>
<td>1643</td>
</tr>
<tr>
<td>Class III</td>
<td>41</td>
<td>825</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>2971</td>
</tr>
</tbody>
</table>

Functionally known proteins and hypothetical proteins were classified into three categories (Class I - Class III) based on 2D correlation analysis for DNA/RNA-binding protein classification (Fig. 3). The numbers of proteins in each category are shown. PPV stands for positive predictive value: composition of DNA/RNA-binding proteins (Ribosome + DNA/RNA) within the category. Predicted numbers of DNA/RNA-binding proteins in the hypothetical proteins were calculated as PPV x total number of hypothetical proteins.

97%, suggesting that amino acid composition and periodicity are effective protein profiles to predict not only DNA/RNA-binding proteins but also membrane proteins (Table 3). As an exception, PPV for predicting membrane proteins in the *N. equitans* proteome was low, only 35.7%. Since only 26 membrane proteins are annotated in *N. equitans*, we assumed that the size of the training set was too small to perform accurate classification. However, we were still able to estimate the total number of membrane proteins and DNA/RNA-binding proteins in the proteome of the three domains of life.
Fig. 4. 2D correlation analysis of membrane proteins in six model species based on amino acid composition and periodicity. Functionally known proteins in the six model species are plotted on a 2D correlation plot of CO score and PD score. The membrane proteins (gray) and other functionally known proteins (black) are shown. The two lines represent the threshold of optimum PPV value for accurate prediction of membrane protein candidates (white arrow) and maximum MCC value for optimizing sensitivity and specificity (black arrow). The ranges of the three classes (Class I - Class III) are defined for each species.
Functionally known proteins and hypothetical proteins were classified into three categories (Class I - Class III) based on 2D correlation analysis for membrane protein classification (Fig. 4). The numbers of proteins in each category are shown. PPV stands for positive predictive value: composition of membrane proteins (Membrane) within the category. Predicted numbers of DNA/RNA-binding proteins in the hypothetical proteins were calculated as PPV x total number of hypothetical proteins.

### Table 3. Summary of membrane protein classification in six model species

<table>
<thead>
<tr>
<th></th>
<th>Functionally known proteins</th>
<th>Hypothetical proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Membrane</td>
<td>Other</td>
</tr>
<tr>
<td><strong>N. equitans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Class II</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>Class III</td>
<td>13</td>
<td>244</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>291</td>
</tr>
<tr>
<td><strong>M. kandleri</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>Class II</td>
<td>74</td>
<td>33</td>
</tr>
<tr>
<td>Class III</td>
<td>76</td>
<td>899</td>
</tr>
<tr>
<td>Total</td>
<td>178</td>
<td>937</td>
</tr>
<tr>
<td><strong>E. coli</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>514</td>
<td>2</td>
</tr>
<tr>
<td>Class II</td>
<td>240</td>
<td>104</td>
</tr>
<tr>
<td>Class III</td>
<td>512</td>
<td>2187</td>
</tr>
<tr>
<td>Total</td>
<td>1266</td>
<td>2293</td>
</tr>
<tr>
<td><strong>Synechocystis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>127</td>
<td>25</td>
</tr>
<tr>
<td>Class II</td>
<td>82</td>
<td>61</td>
</tr>
<tr>
<td>Class III</td>
<td>208</td>
<td>1067</td>
</tr>
<tr>
<td>Total</td>
<td>417</td>
<td>1153</td>
</tr>
<tr>
<td><strong>C. elegans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>240</td>
<td>3</td>
</tr>
<tr>
<td>Class II</td>
<td>167</td>
<td>113</td>
</tr>
<tr>
<td>Class III</td>
<td>272</td>
<td>1500</td>
</tr>
<tr>
<td>Total</td>
<td>679</td>
<td>1616</td>
</tr>
<tr>
<td><strong>H. sapiens</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class I</td>
<td>453</td>
<td>11</td>
</tr>
<tr>
<td>Class II</td>
<td>644</td>
<td>454</td>
</tr>
<tr>
<td>Class III</td>
<td>927</td>
<td>3604</td>
</tr>
<tr>
<td>Total</td>
<td>2024</td>
<td>4069</td>
</tr>
</tbody>
</table>

Functionally known proteins and hypothetical proteins were classified into three categories (Class I - Class III) based on 2D correlation analysis for membrane protein classification (Fig. 4). The numbers of proteins in each category are shown. PPV stands for positive predictive value: composition of membrane proteins (Membrane) within the category. Predicted numbers of DNA/RNA-binding proteins in the hypothetical proteins were calculated as PPV x total number of hypothetical proteins.

### §4. Conclusions

In this manuscript, by using several protein profiles, we were able to visualize and compare the overall distribution of proteome data in the three domains of life and further classified two important protein families: DNA/RNA-binding proteins...
and membrane proteins with high accuracy. We further adapted the threshold to hypothetical proteins and were able to estimate the proportions of the two protein families in the six model species. Our results indicated that approximately 60% to 80% (with overlapping genes sharing 2%) of the proteome consists of DNA/RNA-binding and membrane proteins (Fig. 5), suggesting the importance of these two protein families in the biological system. Further, comparison of the proportions of the two protein families in the three domains of life demonstrated that dynamic fluctuation of protein numbers and their characteristics has occurred throughout the evolution of life. We believe that our study will provide a molecular basis toward understanding the diversity of proteins and their functions in the three domains of life.

Acknowledgements

We thank all RNA project members at the Institute for Advanced Biosciences for helpful discussion. This research was supported in part by 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); research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan; and a grant from Keio University.

References

4) N. Siew and D. Fischer, Structure 11 (2003), 7.