# Development of Method to Predict Protein Function of *Escherichia coli* Using Protein Complex Data

Seira Nakamura
Rintaro Saito

## abstract

We have developed a method to predict protein functions using protein complex data of *Escherichia coli* which we obtained from our His-tag pull-down experiments. Our method first converts complex data (ex. protein A, B and C form complex) into binary protein-protein interaction data assuming that all the proteins in the complex interact with each other (ex. protein A-B, B-C and A-C interact). One of the ways to predict protein function from binary protein-protein interaction data is to assume that function of uncharacterized protein is identical to that of interacting partner. The accuracy of prediction in this way is estimated to be similar to rate of interacting pairs sharing common function, which is 16% in the constructed binary interaction network (Schwikowski et al. 2000). To raise this accuracy, the method selects interactions which seem to be biologically meaningful. In particular, it selects pairs of interacting proteins having high expressional correlations and/or those which are conserved in similar bacteria (i.e. those which show similar “phylogenetic profile”). We succeeded to predict 6 uncharacterized proteins, such as yadF and ybeY, at the accuracy of 60-85%.

## 1. Introduction

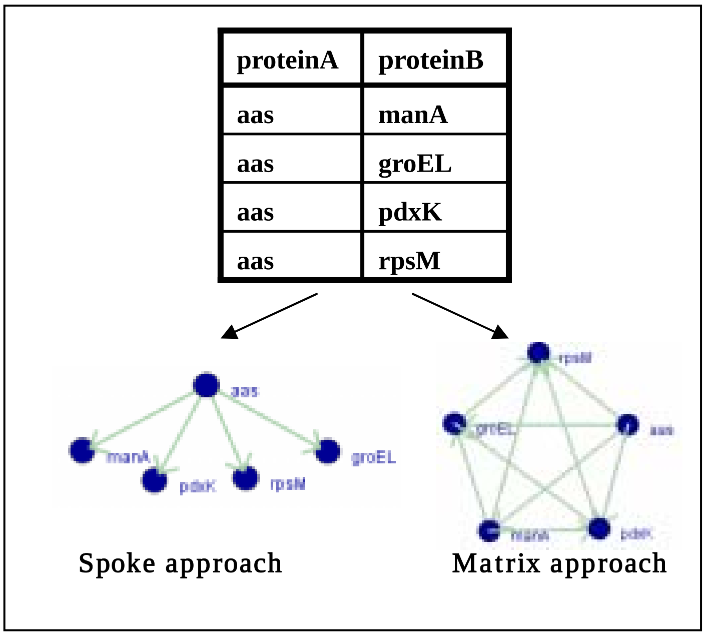
Thousands of proteins that an organism has *in vivo* interact frequently with each other thickly in biological pathways. Protein-protein interaction (PPI) plays an important role in these pathways. PPI information can be used to predict the localizations of proteins, and function of hypothetical proteins. Until now, PPI data of *S.cerevisiae* is available from DIP [1] etc. and analysis of genome-wide PPI of *S. cerevisiae* has been done by many researchers. However, analysis of those of *Escherichia coli* has not been conducted because of the limited amount of publicly available PPI data. In this study, we used genome-wide PPI data of *Escherichia coli* produced by Mori lab.

Our goal of this study is to predict the functions of hypothetical proteins by using PPI data obtained from experiments.

## 2. Description

The two methods we used to describe PPI data are “spoke approach” and “matrix approach” (fig.2-1). PPI experiment data produced by Mori lab is the binary data by “spoke approach”. By the method, the other proteins (not bait protein) within the complex are described to be having interaction with only the bait protein. Within the protein complex, all the proteins do not always have interaction with the bait protein but with the other proteins in the comlex. So we used “matrix approach” to count all of the high-probability protein pairs.

By the study of *S.cerevisiae* PPI data [2], the percentage of that PPI pairs have the same function category is 63%. By “spoke approach” and “matrix approach” of *Escherichia coli* PPI data, the values are 6% and 16%. These percentage means that all of the proteins within the complex do not have interactions with each other proteins. To begin with, the experiment data could be including false-positive data. In order to predict the function of hypothetical proteins, the PPI experiment binary PPI data must be refined by any method.



| proteinA | proteinB |
|---|---|
| aas | manA |
| aas | groEL |
| aas | pdxK |
| aas | rpsM |

fig.2-1: Spoke approach and Matrix approach

## 3. Refinement

The experiment method of *Escherichia coli* we used is “His-tag pull-down experiment” and mass spectrometry (MS). The PPI experiment data (15,551 interactions) may include “false positive” and binary data produced by “matrix approach” include much of wrong data. We think that the PPI pairs within *Escherichia coli* tend to have common function as *S.cerevisiae*. According to the percentage of PPI that have same function (”spoke approach”:6%, “matrix approach”:16%), the data includes many false data. The rise of the “the percentage of the same-function PPI pairs” is proportion in the refinement, and we used the percentage as the index of refinement. The goal of refinement is to raise the percentage to about 60% as S.cerevisiase.We used three procedures to refine the PPI experiment data, treatment of disjunction, phylogenetic profiling[3] and expression pattern. By these procedures, we removed uncredible data.

### 3.1 Treatment of Disjunction and homodimer

The PPI experiment data we used includes some of “disjunction”.(fig.3-1) By disjunction, all of PPI experiment data cannot be treated as homogeneous dimension.

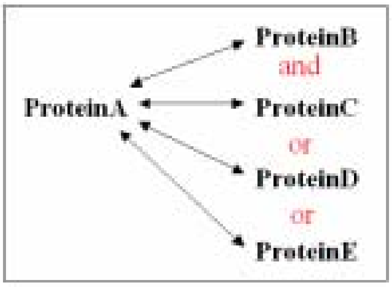


fig.3-1: Disjunction descriptions

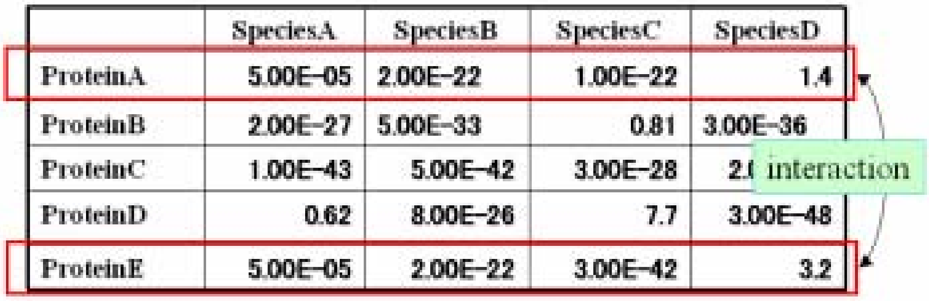
With disjunction, 100568 interactions are produced by “matrix approach”. To pick out the credible PPI pairs from disjunction, we made reference to PPI database (DIP), journals about PPI. By using this reference, we chose strong candidates from disjunction, and removed the other data.

A homodimer is a structure that consists of two identical substructures. Homodimer description in this PPI experiment data, we cannot treat that as PPI data. According to mass spectroscopy, interactions data between the same proteins are uncredible.

As the result of this step, 100,568 interactions are refined to 19,049 interactions.

### 3.2 Phylogenetic Profiling

Phylogenetic profiling allows the prediction of function of uncharacterized proteins. If two proteins are functionally linked, they tend strongly to be found in the same subset of completely sequenced genomes. The patterns of presence or absence of proteins across all known genomes can be used to functionally group a significant fraction of all known protein sequences.(fig.3-2) This leads to a deeper understanding of the role many proteins play in the cell and infers functions for many hundreds of previously uncharacterized proteins. The advantageous effect of PPI prediction by phylogenetic profiling is recognized [3].



| | SpeciesA | SpeciesB | SpeciesC | SpeciesD |
|---|---|---|---|---|
| ProteinA | 5.00E-05 | 2.00E-22 | 1.00E-22 | 1.4 |
| ProteinB | 2.00E-27 | 5.00E-33 | 0.81 | 3.00E-36 |
| ProteinC | 1.00E-43 | 5.00E-42 | 3.00E-28 | 2.( |
| ProteinD | 0.62 | 8.00E-26 | 7.7 | 3.00E-48 |
| ProteinE | 5.00E-05 | 2.00E-22 | 3.00E-42 | 3.2 |

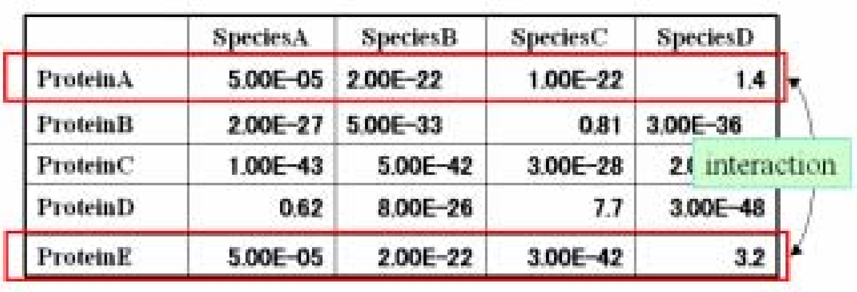
interaction

fig.3-2: Phylogenetic profiling

To assess the pattern of existence information among the plural species, we searched the “homologous gene” among plural species by BLAST and listed the minimum E-value, and calculate the correlation coefficient.

### 3.3 Expression Pattern

Gene expression and protein interaction data shows that protein pairs encoded by co-expressed genes interact with each other more frequently than with random proteins.[4] Furthermore, the mean similarity of expression profiles is significantly higher for respective interacting protein pairs than for random ones. Expressional correlation of interacting pairs are statistically significant but not remarkable.[7] (the PPI protein pairs have similar expression pattern.)



| | SpeciesA | SpeciesB | SpeciesC | SpeciesD |
|---|---|---|---|---|
| ProteinA | 5.00E-05 | 2.00E-22 | 1.00E-22 | 1.4 |
| ProteinB | 2.00E-27 | 5.00E-33 | 0.81 | 3.00E-36 |
| ProteinC | 1.00E-43 | 5.00E-42 | 3.00E-28 | 2.( |
| ProteinD | 0.62 | 8.00E-26 | 7.7 | 3.00E-48 |
| ProteinE | 5.00E-05 | 2.00E-22 | 3.00E-42 | 3.2 |

interaction

fig.3-2: Expression Pattern

To assess the pattern of existence information among the plural species, we used correlation coefficient in common with phylogenetic profiling.

# 4. Result

## 4.1 By Phylogenetic Profiling

| Correlation Coefficient | Percentage*1 | Same Category*2 | Function-Known PPI*3 | Refined PPI*4 |
|---|---|---|---|---|
| More than 0.7 | 38.6% | 51 | 132 | 145 |
| More than 0.8 | 45.2% | 33 | 73 | 75 |
| More than 0.9 | 57.1% | 20 | 35 | 37 |

*1 : the percentage ( *2 / *3 *100)
*2 : the number of binary PPI sharing a common function
*3 : the number of PPI binary data (hypothetical protein excluded)
*4 : the number of PPI binary data (refined by phylogenetic profiling)

Table.4-1: The refinement by phylogenetic profiling

(Table.4-1) The percentage of PPI sharing same function (accuracy of refinement) is up to 57.1% (correlation coefficient is more than 0.9). Although the number of refined PPI data is small, the refinement by phylogenetic profiling is validated.

## 4.2 By Expression Pattern

| Correlation Coefficient | Percentage[*1] | Same Category[*2] | Function-Known PPI[*3] | Refined PPI[*4] |
|---|---|---|---|---|
| more than 0.7 | 38.6% | 351 | 909 | 997 |
| more than 0.8 | 51.2% | 297 | 580 | 610 |
| more than 0.9 | 79.8% | 154 | 193 | 196 |

*1 : the percentage ( *2 / *3 *100)

*2 : the number of binary PPIs sharing a common function

*3 : the number of PPI binary data (hypothetical protein excluded)

*4 : the number of PPI binary data (refined by expression pattern)

**Table.4-2: The refinement by expression pattern**

(Table.4-2) The percentage of PPI sharing same function (accuracy of refinement) is up to 79.8% (correlation coefficient is more than 0.9). Remarkably, the expression pattern is useful method to refine the PPI experiment data.

## 4.3 Refined PPI Experiment Data

| Correlation Coefficient | Percentage[*1] | Same Category[*2] | Function-Known PPI[*3] | Refined PPI[*4] |
|---|---|---|---|---|
| more than 0.7 | 36.7% | 364 | 991 | 1090 |
| more than 0.8 | 49.0% | 305 | 623 | 653 |
| more than 0.9 | 75.6% | 167 | 221 | 225 |

*1 : the percentage ( *2 / *3 *100)

*2 : the number of binary PPIs sharing a common function

*3 : the number of PPI binary data (hypothetical protein excluded)

*4 : the number of PPI binary data (refined by phylogenetic profiling & expression pattern)

**Table.4-3: The refinement by phylogenetic profiling & expression pattern**

(Table.4-3) We cumulated the refined data by two methods. The number of refined PPI is 225 interactions, and the percentage of PPI sharing same function (accuracy of refinement) is up to 75.6% (correlation coefficient is more than 0.9). If the rate of PPI, that have the same function, is more than about 60% as *S.cerevisiae*, these refined PPIs have considerable credibility. Although the refined PPI data by phylogenetic profiling reduced the rate of same function (79.8% → 75.6%), the result of each method has very few crossover. We treated this data as refined PPI data to predict functions of hypothetical proteins.

## 5. Prediction of hypothetical protein's function by using PPI data

Two methods to refine PPI experiment data, phylogenetic profiling and expression pattern, have correlation with protein function closely. 75.6% of the refined PPI data by these methods (correlation coefficient is more than 0.9) has the same function. By using these binary data, we can predict hypothetical protein's function with accuracy of 75.6%.

### 5.1 Method

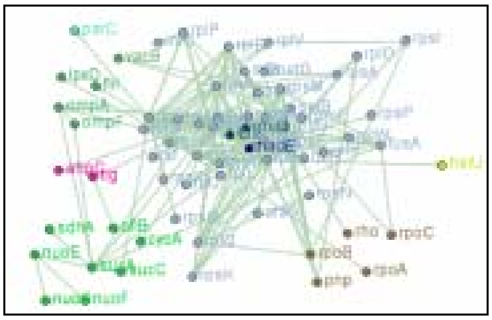


fig.5-1: PPI network (refined PPI data)

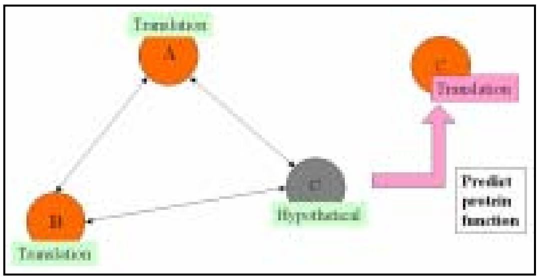


fig5-2: The method to predict protein's function

According to the PPI network described from refined PPI data(fig.5-1) of *Escherichia coli*, the size of protein complex within *Escherichia coli* is huge. Each of proteins is colored by function category. As the color division, the same function proteins tend to interacting closely within the protein complex. Our method to predict functions of hypothetical proteins is based on this feature of complex. The steps of the method is,

**1: to list the proteins having interaction with a hypothetical protein.**
**2: to choose the main function category around a hypothetical protein from the list.**
**3: to predict the function category as the main function category (step2)**

step 1〜3 is performed for every hypothetical protein.

The main function category around a hypothetical protein is the protein that is the largest number around the hypothetical protein.(ex. the functionA is largest around a hypothetical protein, main function is functionA). The concept of this method is called "guilt-by-association".[7] The protein complex is the assembly of PPI pairs sharing same function.

### 5.2 Result

We predicted 6 hypothetical protein’s function.(Table.5-2)

| gene name | Function of hypothetical protein predicted by refined PPI data |
|---|---|
| pdxK | Biosynthesis of cofacters, prosthetic groups, carriers |
| vacB | Translation |
| ycdX | Biosynthesis of cofacters, prosthetic groups, carriers |
| yihK | Translation |
| ylaB | Other categories |
| yleA | Translation |

table.5-2: predicted function of hypothetical protein predicted by refined PPI data

## Summary and Discussion

We recognized the efficiency of each two methods to refine PPI experiment data. The number of PPI experiment data is refined from 100,568 to 225 at the accuracy of 75.6%. By using this refined PPI data, we predicted 6 functions of hypothetical proteins at the accuracy of 75.6%. Although the value is high enough to predict the functions, we guess the number of refined PPI is fewer than real data. We need the study focused on the number of PPI with a certain level of accuracy.

Some of the protein complexes of *Escherichia coli* are extremely huge and it is difficult to recognize and study the network of interaction. The system to compart the huge complex to small matrix is needless. (fig6-1)

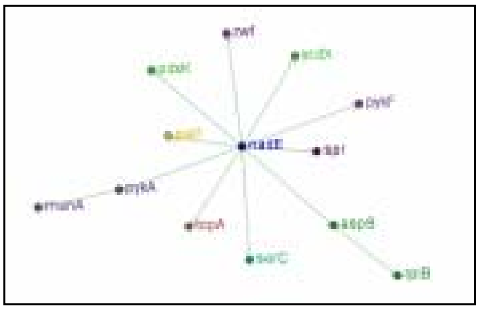


fig.6-1: the example of comparted PPI network

## 7. Outlook

Essential gene tends to have interactions with many proteins. A large amount of connections around the essential genes make the recognition of protein network difficult. To recognize the protein network of *Escherichia coli*, we will find essential gene from refined PPI network and classify the refined PPI data by removing essential genes.

## 8. Acknowledgement

We have benefited from discussions and offering PPI expression data with many people on the topics covered here. In particular, we thank Aya Itoh, Takeshi Ara and Mori lab. I also thank faculties and staffs of Institute for Advanced Bioscience. Finally, I gratefully acknowledgement to Professor Masaru Tomita for offer the opportunity and environment for my study.

# 9. References

[1] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. “**DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.**” Nucleic Acids Res. 2002 Jan 1;30(1):303-5.

[2] Schwikowski B, Uetz P, Fields S. “**A network of protein-protein interactions in yeast..**” Nat Biotechnol. 2000 Dec;18(12):1257-61.

[3] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. “**Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.**” Proc Natl Acad Sci U S A. 1999 Apr 13;96(8):4285-8.

[4] Grigoriev PS, Lobocka MB. “**Determinants of segregational stability of the linear plasmid-prophage N15 of Escherichia coli.**”

[5] Grigoriev A. “**relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.**” Nucleic Acids Res. 2001 Sep 1;29(17):3513-9.

[6] Jansen RC, Nap JP. “**Genetical genomics: the added value from segregation.**” Trends Genet. 2001 Jul;17(7):388-91.

[7] Stephen Oliver. “**Guilt-by-association goes global**” Nature. 2000 Feb 10;403(6770):601-3.