ベイズ統計を用いた
ゲノムワイドな大腸菌 PPI データの精製と予測

政策・メディア研究科　中村征良

要旨

代謝やシグナル伝達など、タンパク質の関わっている多くの生命活動を把握するためには、ゲノムワイドなタンパク質間相互作用(PPI: Protein-Protein Interaction)ネットワークを構築することが必要不可欠となる。モデル原核生物である大腸菌においてはなおさらであるが、その情報量は十分ではない。そこで我々は、His-tag Pull-Down 法によるスクリーニングデータをもとに、computational な手法を用いて信頼性の高い大腸菌 PPI ネットワークを構築することを目指している。

各相互作用の信頼性を評価するために用いた事象(指標)は、系統プロファイル、発現相関、IG (Interaction Generality)、モチーフ間相互作用、遺伝子表現型、大腸菌⇔ピロリ菌間の相互作用の保存情報、そして転写単位(オペロン)の７つである。これらの事象をベイズ統計によって統合し、各タンパク質ペアが相互作用する確率を算出し、実験データにおける個々のデータを対象にし、相互作用データとしての信憑性があるかどうかをその確率を用いて判断した。その結果、我々は信頼性の高い 427 個の相互作用データを取得することが出来た。構築した PPI ネットワークは 3,667 相互作用からなり、1,277 個のタンパク質で表現される。

1. はじめに

生体内には DNA,RNA 等の核酸やタンパク質、代謝物質などが含まれており、それらは互いに反応や結合等の作用を繰り返してその生体の生命活動を維持している。特に、代謝物質間の反応を触媒する酵素や、DNA から mRNA を転写する RNA ポリメラーゼ、タンパク質を翻訳するリボソームなど、生命活動において非常に重要な機能を持った分子のほとんどがタンパク質複合体であるため、現在では、それらの研究はプロテオミクスと称され、盛んに行われている。タンパク質複合体は、個々の遺伝子領域から翻訳されるタンパク質が互いに結合(相互作用)することで構築される。そのため、そのようなタンパク質間の相互作用(Protein-Protein Interaction)の情報を網羅することは、タンパク質複合体の存在を知ることができるだけでなく、タンパク質の機能[1][2]や、代謝[3][4]やシグナル伝達経路[5][6][7][8]の予測を行うことができることを意味する。なお現在、酵母菌においてはゲノムワイドな PPI データが公開されており[9]、それらを用いた解析が頻繁に行われている。

現在、ゲノムワイド PPI データを実験的に取得する有名な手法としては、Y2H(Yeast Two Hybrid Assay)[10][11][12][13]や TAP(Tandem Affinity Purification)[14][15][16]などが知られている。しかしながら、Y2H には「過剰発現させるため、実際の環境とは大きく異なる」「他のタンパク質を仲介して間接的に相互作用している場合がある」、TAP に関しては「タグを付けた部分が実際の相互作用の障害となっている」等の問題が指摘されており、それぞれの結果が多くの false positive, false

negative を含んでいる[17][18]ことが知られている。このようなデータを用いると、誤った解釈につながってしまうため[19]、このような false positive を除く様々な手法が開発されている[20][21][22][23][24][25]。本研究の目的は、ゲノムワイドな実験 PPI データが含んでいる false positive を取り除いて信頼性の高い PPI データを抽出し、ゲノムワイドで信頼性の高い PPI ネットワークの構築を行うことである。

PPI ネットワークからタンパク質の機能予測等の様々な生物学的な考察を行う場合、代謝や個々の遺伝子のアノテーションが揃っていることが望ましい。そのため、我々は原核のモデル生物である大腸菌のゲノムワイド実験 PPI データを奈良先端科学技術大学院大学の森研究室から提供していただき、そのデータセットから信頼性の高いものを抽出する手法の構築を行った。なお、提供していただいたデータセットは、His-tag 付タンパク質を用いた Pull-Down 法(M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, R. Saito, T. Ioka, T. Kawamura, C. Takita, A-U. Amin, A. Hirai et al. in preparation) [26]によるもので、bait タンパク質 2669 個を用いて検出されたものである。今回、これらの実験データを情報処理によるそれぞれの予測手法に対応させるために、まずそれぞれの bait タンパク質とそれに結合して検出された 1 つもしくは複数の prey タンパク質全てのグループを 1 つの複合体とみなし、その複合体に属している全てのタンパク質が総当りで相互作用していると考え、複合体データである実験データを PPI データ(相互作用するタンパク質ペアを最小単位とするバイナリー形式)に変換した(matrix approach[27])。また、この実験手法においてはホモダイマー(同一タンパク質間の相互作用)を認識することが難しいことと、2 つの同じタンパク質を入力とすることができる予測手法が少なかったため、これらのデータセットからホモダイマーを除去した。その結果、精製対象である 9233 個の PPI 実験データを取得した。

2. PPI 予測手法の検証

信頼性の不明瞭なデータセットから信頼性の高いものだけを抽出(精製)するためには、各データにおける信頼性を裏付ける情報が別個に必要となり、またその情報が多ければ多いほど信用するに足る情報となる。そしてベイズ等の確率統計を用いてそれらの情報を総合的に判断することで、その信頼性を定量的に測ることが可能となる。母集団に含まれる個々のデータの信頼性を測った場合、それは予測することと同意であるため、実験 PPI データの中から信頼性の高いものを抽出する手法と、相互作用を行うタンパク質を予測する手法は同意であるといってもよい。我々は、PPI と高い相関を有している事象を列挙し、予測手法としての成績を評価した。

2.1 各予測手法の成績検証とデータセットの取得

それぞれの予測の成績を評価するためには、その予測による信頼性の評価を行う必要がある。今回、その信頼性の指標としてはベイズ統計における確率変数(Probability)を用いて行った。Probability の算出には、既に相互作用していることが判明しているタンパク質ペアの情報(既知の PPI 情報)である Positive data と、相互作用していないタンパク質ペアの情報である Negative data が必要となる。 Probability 算出に用いた Positive Data 及び Negative Data の取得法は以下の通りである。

2.1.1. Positive Dataset(既知 PPI 情報)

PPI の公的データベース DIP (http://dip.doe-mbi.ucla.edu/)[28][29]には大腸菌 PPI データは 611 個しか揃っておらず、学習に用いるには不足となる。そのため、さらに PubMed から大腸菌の相互作用に関する論文を網羅的に取得し、Curation CGI を用いて手作業で一つずつ PPI データを取得した。そうして得られた PPI データに、DIP のデータ、EcoCyc (http://ecocyc.org/)[30][31]のデータを加えた。その結果、1123 個のタンパク質で表現される 3285 個の PPI データを取得することができた。

2.1.2. Negative Dataset(相互作用しないタンパク質ペア)

相互作用を行う PPI データは、同機能カテゴリに属している傾向が高いとされている(酵母においては約 6 割)。また、発現場所が違うタンパク質も、物理的に相互作用を起こすとは考えられない。同機能カテゴリに属さず、しかも発現場所が違うタンパク質ペアを Negative Data とした。なお、機能カテゴリ情報は GenoBase(http://ecoli.aist-nara.ac.jp/)[32]から取得し、発現場所の情報は、PSORT[33]を用いて調べた。

以上のデータセットを用いて、各タンパク質ペアが相互作用するという Probability は以下の様に算出する。

$$probability(ES) = \frac{Ratio\,Value(ES)}{1 + Ratio\,Value(ES)}$$

ES とは、それぞれの予測手法(ベイズ統計における「証拠・根拠」ES: Evidence Sources)を用いた予測結果のことを指す。オッズ比(Ratio Value)は、各 positive data と negative data から導かれる

$$P'(positive \mid ES) = \frac{P(positive)P(ES \mid positive)}{P(ES)} \qquad P'(negative \mid ES) = \frac{P(negative)P(ES \mid negative)}{P(ES)}$$

を用いて以下の様に算出される。

$$Ratio\,Value(ES) = \frac{P'(positive \mid ES)}{P'(negative \mid ES)}$$

P(*positive*)は、大腸菌における全 PPI データの、大腸菌内の全てのタンパク質ペアに対する割合であり、それに準じて P(*negative*)は 1- P(*positive*)となる。大腸菌における全 PPI データは、取得した Positive data の密度と大腸菌の遺伝子数をもとに、40,000 個($_{4000}C_2 \times 3285/_{1123}C_2$)と仮定した。P(*ES*|*positive*)及び P(*ES*|*negative*)は Positive data と Negative data がそれぞれ ES になる確率である。

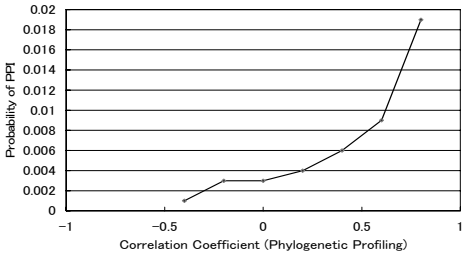
これらをもとに各 PPI 予測手法の結果とその Probability を測ることで、成績を評価した。用いた予測手法は、Phylogenetic Profiling, Expression Pattern, IG(Interaction Generarity), MMI(Motif-Motif Interaction), 遺伝子表現型, Interolog, 転写単位(オペロン)の 7 手法である。

2.2 Phylogenetic Profiling(系統プロファイル)

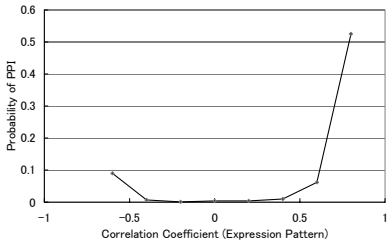
機能的につながりを有するタンパク質ペアは、他生物種における存在情報も似通っている[34]。ようするに、ほぼ同じ進化経路をたどっている傾向が高いということになる。大腸菌の各タンパク質の進化経路パターンを調べるため、大腸菌における各翻訳産物のアミノ酸配列と、全ゲノム配列が解読された 42 生物種(古細菌 6 種、真正細菌 36 種)における各翻訳産物のアミノ酸配列の相同性を BLAST を用いて調べ、大腸菌のタンパク質と系統的につながりを持つものを BLAST の出力である E-value を用いて表現した。例えば、大腸菌におけるタンパク質 A,B,C があったときに、それらを他の生物種 X,Y,Z の有するタンパク質のアミノ酸配列に BLAST を用いてその相同性を調べると、タンパク質 A に関しては X,Y,Z のタンパク質に対してそれぞれ E-value が 1.0×10^{-4}, 2.0, 1.0×10^{-6} となり、タンパク質 B では 1.0×10^{-10}, 1.0×10^{-20}, 2.0、タンパク質 C では 2.0×10^{-4}, 4.0, 2.0×10^{-6} であったとする。この場合、タンパク質 A と C は E-value が互いに相関を持っていることから進化系統的につながりをもっていることとなる。各タンパク質間の相関は、E-value 配列の相関係数を算出することで調べた。この手法を用いて、我々は進化系統的につながりをもっている大腸菌タンパク質ペアと PPI の相関を確認したところ、系統プロファイルのスコア(相関係数)と Probability が正の相関を有しており、予測手法としての有効性を確認した。(図 2-1)

2.2 Expression Pattern(遺伝子発現相関)

酵母において、共発現をする遺伝子の翻訳産物は相互作用を行う傾向が強いことがわかっている[35][36]。これらの傾向が大腸菌に置いても見られるのかを確認した。大腸菌の遺伝子発現データは、KEGG(http://www.genome.jp/kegg/)及び GenoBase から取得し、そこから得られる発現量推移の類似度を、相関係数を算出して比較した。その結果、その相関係数と Probability は、相関係数が 0 以上の領域(共発現の領域)において正の相関を示していた。そのため、酵母と同じく大腸菌においても、発現相関は PPI 予測に有効であるという結果が得られた。(図 2-2)



(図 2-1)
Phylogenetic Profilig におけるスコア(縦軸)と、Probability(横軸)の相関グラフ



(図 2-2)
タンパク質ペアの発現の相関係数(縦軸)と Probability(横軸)の相関グラフ

2.3 IG(Interaction Generality)

IG(Interaction Generality)とは、構築した PPI ネットワークを幾何学的に判断してその相互作用の確からしさを算出する手法である[24][37]。出力されるスコアが大きければ大きいほど相互作用している確率が高いことを表現する。その値の大きさにより、0 以上 1 未満を「0」、1 以上2未満を「1」、2 以上を「2」と分類して Probability の算出に使用した。IG 値算出に用いる PPI データは、実験データを Spoke Approach を用いてバイナリー形式に変換したものを使用した。Spoke Approach とは、prey タンパク質全てが bait タンパク質と相互作用していると描写する手法である**[27]**。各分類における Probability を算出した結果、IG 値と Probability には正の相関がみられ、予測手法としての有効性が示された。(図 2-1)

2.4 MMI(Motif-Motif Interaction)

MMI(Motif-Motif Interaction)とは、**Pfam database** (http://pfam.wustl.edu/)から得られるタンパク質の配列モチーフ情報と実験 **PPI** データの情報を合わせることで相互作用を行う可能性が高いアミノ酸配列モチーフペアを取得し、それをもとに PPI を予測する手法である[38]。PPI 予測の値は O/E 値で出力され、この値が大きければ大きいほどそのタンパク質ペアが相互作用している確率が高いことを意味している。この O/E 値が、0 以上 1 未満のものを「0」、1 以上2未満を「1」、2 以上を「2」と分類して Probability の算出に使用した。結果、その値と Probability には正の相関がみられた(図 2-1)。

2.5 Essentiality (遺伝子表現型)

Essential protein は、HUB(多くのタンパク質と相互作用している)である傾向が高いことが分かっている[42]。我々は、**PEC database** (http://www.grs.nig.ac.jp/ecoli/pec/index.jsp)から取得できる大腸菌の Essential protein 情報をもとに、大腸菌のタンパク質を Essential protein と non-essential protein に分類した。実験 PPI データには 341 個の Essential protein、1,922 個の non-essential protein が含まれており、non-essential protein 同士の相互作用を「1」、essential protein と non-essential protein の相互作用を「2」、essential protein 同士の相互作用を「3」と分類し、もしどちらかのタンパク質が unknown であった場合は「0」と分類した。この分類における値が大きければ、そのタンパク質ペアが含む essential protein の数が大きくなることを意味する。結果、この値と Probability においても、正の相関がみられた。(図 2-1)

2.6 Interolog (他生物種 PPI データの参照)

他生物種における PPI データから、相互作用を行う各タンパク質のオーソログを取得して目標生物種(ここでは大腸菌)の PPI を予測する手法である[39]。*H.pylori* の PPI データ[40]が含んでいる各タンパク質を、それらとオーソログな関係にある大腸菌タンパク質に変換した。オーソログは、大腸菌のタンパク質と ssearch を用いて E-value が e^{-10} 以下のものを使用した。その結果、ピロリ菌において相互作用しているタンパク質は、大腸菌においても相互作用している傾向があり、大腸菌 **PPI** 予測に効果を有することが示された。(図 2-1)

2.7 転写単位(オペロン)

原核生物の mRNA は、頻繁に複数のタンパク質を翻訳する。同じ転写因子によって翻訳される複数のタンパク質は機能的にも相関がある傾向が強いことがわかっている**[41]**。それは、実験 **PPI** データに含まれているタンパク質ペアが同じ転写因子による翻訳産物同士であった場合は、相互作用している可能性が向上することを示す。我々は **EcoCyc** から転写単位の情報を取得し、**PPI** との相関を調べた。(表 2-1)

表 2-1: 各 PPI 予測手法(2.3～2.7)の出力と、Probability の相関

Interaction Generality	Probability
0	0.003
1	0.069
2	0.217

Essentiality pattern	Probability
0	0
1	0.009
2	0.005
3	0.103

MMI score	Probability
0	0.003
1	0.021
2	0.024

Potential Interolog	Probability
No	0.003
Yes	0.163

Operon	Probability
No	0.003
Yes	0.925

3. PPI 予測手法の統合と評価

3.1 手法の統合

３章で紹介した７つの PPI 予測手法は、それぞれその出力と Probability が正の相関を有していることから、どれも PPI 予測にとって効果を有していることがわかる。しかし、それらの予測結果は独立とまではいかないが重複が少なく、それぞれを個別に使用した予測結果は予測手法に準じた非常に大きな偏りを持っていた。そこで取得できる PPI データの増加とその偏りの削減のために、ベイズを用いてこれらの７つの PPI 予測手法を統合した。ベイズによる統合の手法には”Fully Connected Bayes”と” Naïve Bayes”の 2 通りがあり、統合する手法が互いに相関を持っているときは前者を選択し、互いに独立であるならば後者を選択する必要がある。今回は、この７つの手法が互いに相関を有しているために前者の”Fully Connected Bayes”を選択した。

統合した予測手法における出力の確率変数 Probability(P($positive|ES1,ES2,ES3\ldots$))は、個々の PPI 予測手法と同様に Positive data、Negative data を学習に用いた以下の式で算出される。ここで用いている ES1, ES2, ES3…は各予測手法のスコアである。

$$RatioValue(ES1,ES2,ES3,\cdots)=\frac{P'(positive\mid ES1,ES2,ES3,\cdots)}{P'(negative\mid ES1,ES2,ES3,\cdots)}$$

$$P(positive\mid ES1,ES2,ES3,\cdots)=\frac{RatioValue(ES1,ES2,ES3,\cdots)}{1+RatioValue(ES1,ES2,ES3,\cdots)}$$

3.2 ベイズにより統合した手法の評価

ただし、このように出力される Probability は、学習したデータに大きく依存しており、学習したデータではないデータセットを用いて評価をすることで初めて使用することが可能となる。今回我々がその評価に用いた指標は *True Positives* / (*True Positives* + *False Positives*)で得られる Accuracy 変数である。True Positive と False Positive の算出手順は以下のとおりである。

・True Positive

1.Positive Data から任意のデータを 1 つ選び、取り除いてベイズによる学習を行う。
2.予測した結果に Positive Data から除いた 1 データが含まれているかを判断する。
3.1~2 の手順を全 Positive Data を選び終わるまで繰り返す。前に選んだデータは選ばない。
4.繰り返した回数のうち、除いたデータを出力した回数の割合を True Positive とする。

・False Positive

1.Negative Data から任意のデータを 1 つ選び、取り除いてベイズによる学習を行う。
2.予測した結果に Negative Data から除いた 1 データが含まれているかを判断する。
3.1~2 の手順を全 Negative Data を選び終わるまで繰り返す。前に選んだデータは選ばない。
4.繰り返した回数のうち、除いたデータを出力した回数の割合を False Positive とする。

こうして得られる Accuracy を用いて Probability を評価したところ、二つの指標は互いに正の相関を持っており、唯一の推定変数である大腸菌における全ての PPI 数を変化させた場合でも変わることはなく(図 3)、十分 PPI の予測に効果的な手法であると評価することができた。

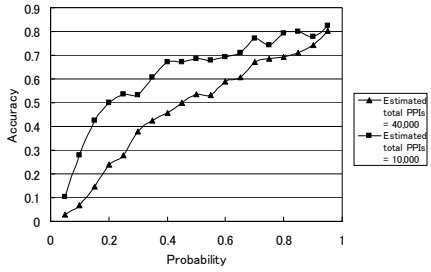


図 3: Probability(縦軸)とその Accuracy(横軸)の相関 (大腸菌における全ての PPI データ数を 10,000 とした場合と 40,000 とした場合)

4. 実験 PPI データの精製及びゲノムワイドな予測結果

統合した手法を用いて PPI の信頼性(Accuracy)を評価した場合、その Accuracy の閾値により取得できる PPI データ数は変化する。この手法を用いて実験 PPI データの精製を行ったところ、図 4-1 から相互作用を行う確率と取得できる PPI 数は Trade-Off の関係にあることがわかる。

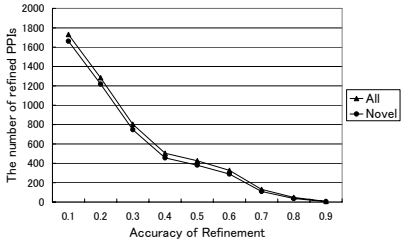


図 4-1: PPI 精製における Accuracy(横軸)と、取得できる PPI 数(縦軸)の推移(ALL:取得できる PPI 数 Novel:取得した PPI データに含まれ、Positive Data に含まれない新規 PPI データ)

我々は Probability の閾値を 0.5 とし(Accuracy の閾値も約 0.5)、実験 PPI データを対象に Probability0.5 以上のタンパク質ペアを抽出した。その結果が以下の表 4-1 である。

表 4-1

	総数	既知データ	新規PPIデータ
実験PPIデータ	9233	94	
精製後データ	427	46	381

実験 PPI データには、全体(9,233 個)のおよそ 1%である 94 個の Positive Data が含まれているが、精製の結果、46 個の Positive Data を含む 427 個の PPI データを取得することが出来た。全体の約 11%が Positive Data を含んでいることになり、使用した Positive Data を反映した学習が行われていることを確認した。次に、大腸菌における全タンパク質ペア(9,290,205 ペア)を対象として Probability0.5 以上のタンパク質ペアを抽出したところ、330 個の Positive Data を含む 1,407 個(新規 PPI データ:1,077 個)のタンパク質ペアを PPI 候補として抽出した。

最終的に、精製した PPI データ(427 個)を既知データである Positive Data と統合させることで、1,277 個のタンパク質からなる 3,667 個の信頼性の高い PPI データセットを構築することができた。このデータセットにおける各タンパク質の有する相互作用数の平均は 6 個である。図 4-2 は、相互作用相手タンパク質の数と、相互作用相手をその数所有しているタンパク質数の分布であり、このような傾向はその他多くの相互作用ネットワークにおいて観察されている[42]。

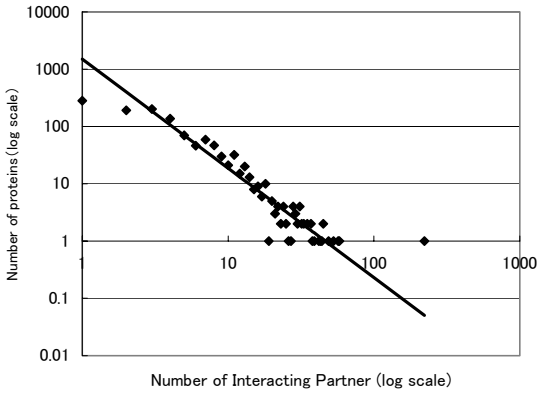


図 4-2: 相互作用相手タンパク質の数(横軸)と、相互作用相手をその数所有しているタンパク質の数(縦軸)の相関

それに加え、構築した PPI データセットでは、相互作用相手の数と、その数を相互作用相手として保有しているタンパク質における essential protein の割合の間には正の相関があることがわかった。例えば、5 個以下の相互作用相手を保有しているタンパク質のうち、それが essential protein である割合は 11.6%(P<0.002)であるのに対し、5 個以上の場合は 20.6%となっている。これらの傾向は、酵母においても観察されている[42]。このように、我々が構築した PPI データの傾向は他に報告されたものと一致しており、実際の細胞内 PPI を反映した結果であると考えられる。

5. 考察

構築した PPI データセットは、生物学的な考察をするのに妥当な信頼性と量を兼ね備えたデータセットであるといえる。それを示唆する事実として、このデータセットの中に、今回使用した Positive PPIs には含まれていない既知データ(例：FlgN-FlgL[43])を新たに文献により見つけることができた。

5.1 機能未知タンパク質の機能予測

現在、大腸菌における多くのタンパク質の機能が未知のままとなっている。しかしながら、PPI ネットワークを用いることで、これらの機能未知タンパク質の機能を予測することが可能となる[18][44][45][46]。相互作用をしているタンパク質は、互いに同じ機能を共有している傾向が強い。信頼性の高い PPI ネットワークにおいてこの手法を用いることで、機能未知タンパク質の機能を予測することができる[24]。我々は、大腸菌におけるタンパク質の機能情報を GenoBase から取得し、構築した大腸菌 PPI ネットワークを用いて機能予測を行った。構築した PPI ネットワークには、98 個の機能未知タンパク質が含まれており、そのうち 45 個の機能未知タンパク質が、1 個もしくはそれ以上の機能が分かっているタンパク質と相互作用していた。特に、そのうち 9 個の機能未知タンパク質に関しては、相互作用相手が全て共通の機能を有しており、機能予測の信頼性が他の機能未知タンパク質と比べて非常に高いといえる[25]。表 5 が、9 個の機能未知タンパク質の機能予測結果である。

表 5：構築した大腸菌 PPI ネットワークを用いた機能未知タンパク質の機能予測結果

遺伝子	機能予測結果(括弧の中は相互作用相手の数)
frvA	Transport/binding protein (3)
Tas	Cellular process (7)
yadI	Transport/binding protein (5)
yajC	Cellular process (7)
ybeV	Cellular process (3)
yhjK	Energy metabolism (3)
yidC	Cellular process (10)
yjcC	Energy metabolism (3)
yliB	Transport/binding protein (3)

5.2 新規PPIデータの考察

全ての生体細胞内には、複数の機能を有したタンパク質が多く存在する。我々の構築したPPIネットワークにおいては、複合体に含まれる複数のタンパク質が、互いに異なった機能を有しているケースが存在した。我々は、そのような相互作用とお互いの機能情報をもとに、タンパク質が有している更なる新規の機能を考察した。

・NusAタンパク質

転写に関わるタンパク質と複製に関わるタンパク質がNusAを仲介して複合体を形成しているものが観察された(図5-A)。NusAは、別の因子であるNusB,NusG,Rhoと複合体を形成してRNA polymeraseに直接結合し、転写の終結を調節するタンパク質である[47]。そのNusAは、我々が構築したPPIデータに含まれているPositive Dataにおいて転写因子複合体のサブユニットであるRpoB-RpoCと相互作用を有していることがわかっている。一方で、我々が実験データから精製したPPIデータにおいて、DNAポリメラーゼであるDnaEが、DNAヘリカーゼであるDnaBと、さらにNusAとも相互作用していることが判明した。DnaBとDnaEはDNA複製因子複合体に含まれるタンパク質である。そのことからNusAは転写においてもDNA複製においても抗終結因子としても働いている可能性が示唆される。

・IscSタンパク質

染色体分離機構と細胞骨格に関わる2つのタンパク質複合体が、硫黄代謝のタンパク質複合体と相互作用している可能性が示唆された(図5-B)。MukBとIscS間の相互作用は、我々の精製したPPIデータに含まれているが、我々が使用したPositive Dataには含まれていない。いわゆる実験PPIデータから精製されたデータである。しかし、この相互作用を新規に文献において見つけることができた[14]。

MreBとIscS間の相互作用は実験PPIデータから精製されて得られた相互作用である。MukBは染色体の構造維持(SMC: Structural Maintenance of Chromosomes)タンパク質の1つであり、細胞分裂における染色体分離にとって不可欠なタンパク質である[48]。MreBは細胞の中心軸に沿って伸びたらせん状の繊維を形成しているタンパク質であり、原核細胞におけるアクチンと構造的にも機能的にも相同性があり、細胞骨格の1つの要素であると考えられている。また、細胞の形作りや染色体の分離を調整する役割を担っていることが知られている[49][50][51]。IscSは、もともとtRNAチオウリジン合成酵素であるシステインデスルフラーゼとして知られており[52]、IscSとIscUは鉄-硫黄クラスターを形成するタンパク質複合体である[53]。そして、IscAもまた、鉄-硫黄クラスターに属していると考えられている[54]。iscSUAは、バクテリアにおいて広く保存されている遺伝子クラスターである。最近においてIscAは、細胞分裂のための収縮リングを形成するFtsZや、分裂部位の決定を行うMinCDEとは別に、極細胞もしくはその周辺の細胞分裂部位に位置していることが報告されている[55]。しかしながら、IscAとMreBの関係はこの報告には触れられていない。そのことから、MukBとIscS、MreB間の相互作用は染色体分離因子の候補となると考えられ、さらにIscSには細胞分裂の際の染色体分離に関係する未知の機能を有していることが示唆された。

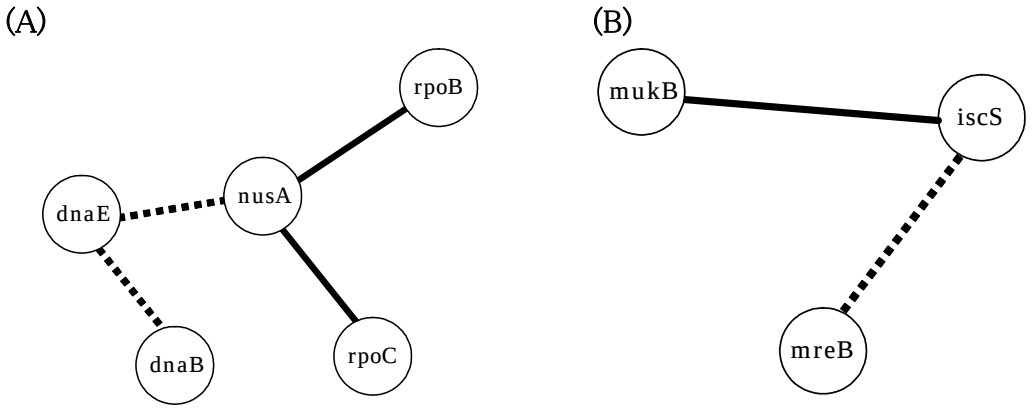


図 5: 構築した PPI データから作られる相互作用ネットワークの例。直線は Positive Data に含まれている相互作用データを意味し、点線は精製された実験データに含まれているデータであることを意味する。(A)は NusA 周辺のタンパク質のネットワークであり、(B)は IscS 周辺のタンパク質である。

5.3 まとめ

今回の我々の実験 PPI 精製手法は、使用した Positive Data と７つの予測手法の傾向に大きく偏っている。実際には、そのような偏りを持たない PPI データが存在している可能性もあり、われわれの手法ではそれらを検出することは難しい。新しい傾向を有する相互作用を検出するためには、学習セット(Positive Data, Negative Data)を更新する必要がある。ただし、それらのデータは信頼性の高い PPI 検出法を用いたものであり、個々のデータが偏りを持っていないものでなくてはならない[56]。

最終的に、我々はゲノムワイドレベルに大腸菌の PPI ネットワークを構築した。そしてそのネットワークを用いてさまざまなタンパク質の新規機能を提案し、構築したネットワークの有用性を示すことができた。最後に、我々が構築した信頼性の高い PPI ネットワークと Protein-DNA 相互作用データのようなゲノムワイドデータを統合することにより、更なるタンパク質の機能に関する情報を得られ、新規な生化学経路を発見することにもつながることが期待できると考える。

6.謝辞

本研究において、実験 PPI データの提供及び助言を頂いた多くの方々、特に 2 年半に渡りご指導いただいた斎藤輪太郎専任講師、伊藤文氏、荒武氏、金井昭夫助教授、和田千恵子氏、MD Arifuzzaman 氏、そして奈良先端科学技術大学院大学森研究室の方々には深くお礼を申し上げます。また、慶應義塾大学先端生命科学研究所に在籍する全ての関係者の方々に感謝の意を表します。最後に、このようなすばらしい研究環境と機会を与えてくださった冨田勝環境情報学部教授に多大なる感謝を致します。

7.参考文献

[1] Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18:** 1257-1261.
[2] Vazquez, A., A. Flammini, A. Maritan, and A. Vespignani. 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* **21:** 697-700.
[3] Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929-934.
[4] Segal, E., H. Wang, and D. Koller. 2003. Discovering molecular pathways from protein interaction and gene expression data.*Bioinformatics* **19 Suppl 1:** i264-271.
[5] Bar-Joseph, Z., G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaakkola, R.A. Young, and D.K. Gifford. 2003. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21:** 1337-1342.
[6] Ideker, T., O. Ozier, B. Schwikowski, and A.F. Siegel. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1:** S233-240.
[7] Liu, Y. and H. Zhao. 2004. A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. *BMC Bioinformatics* **5:** 158.
[8] Steffen, M., A. Petti, J. Aach, P. D'Haeseleer, and G. Church. 2002. Automated modelling of signal transduction networks. *BMC Bioinformatics* **3:** 34.
[9] Sprinzak, E., S. Sattath, and H. Margalit. 2003. How reliable are experimental protein-protein interaction data? *J Mol Biol* **327:** 919-923.
[10] Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141-147.
[11] Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, and M. Tyers. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415:** 180-183.
[12] Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98:** 4569-4574.
[13] Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403:** 623-627.
[14] Gully, D., D. Moinier, L. Loiseau, and E. Bouveret. 2003. New partners of acyl carrier protein detected in Escherichia coli by tandem affinity purification. *FEBS Lett* **548:** 90-96.
[15] Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B. 2001. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods.* Jul;24(3):218-29.
[16] Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. 1999 A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol.* Oct;17(10):1030-2.
[17] Michael Cornell, Norman W. Paton, Stephen G. Oliver. 2004 A critical and integrated view of the yeast interactome. *Comparative and Functional Genomics* Volume 5, Issue 5, 2004. 382-402
[18] Deng, M., F. Sun, and T. Chen. 2003. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput***:** 140-151.
[19] von Mering, C., R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417:** 399-403.
[20] Bader, J.S. 2003. Greedily building protein networks with confidence. *Bioinformatics* **19:** 1869-1874.
[21] Deane, C.M., L. Salwinski, I. Xenarios, and D. Eisenberg. 2002. Protein interactions: two methods for assessment of the

reliability of high throughput observations. *Mol Cell Proteomics* **1:** 349-356.
[22] Goldberg, D.S. and F.P. Roth. 2003. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* **100:** 4372-4376.
[23] Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449-453.
[24] Saito, R., H. Suzuki, and Y. Hayashizaki. 2002. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res* **30:** 1163-1168.
[25] Saito, R., H. Suzuki, and Y. Hayashizaki. 2003. Global insights into protein complexes through integrated analysis of the reliable interactome and knockout lethality. *Biochem Biophys Res Commun* **301:** 633-640.
[26] Zhu, H., M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R.A. Dean, M. Gerstein, and M. Snyder. 2001. Global analysis of protein activities using proteome chips. *Science* **293:** 2101-2105.
[27] Bader, G.D. and C.W. Hogue. 2002. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20:** 991-997.
[28] Salwinski, L., C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32 Database issue:** D449-451.
[29] Xenarios, I., L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30:** 303-305.
[30] Karp, P.D., S. Paley, C.J. Krieger, and P. Zhang. 2004. An evidence ontology for use in pathway/genome databases. *Pac Symp Biocomput***:** 190-201.
[31] Karp, P.D., M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. 2002. The EcoCyc Database. *Nucleic Acids Res* **30:** 56-58.
[32] Oshima, T., H. Aiba, Y. Masuda, S. Kanaya, M. Sugiura, B.L. Wanner, H. Mori, and T. Mizuno. 2002. Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. *Mol Microbiol* **46:** 281-291.
[33] Nakai, K. and P. Horton. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-36.
[34] Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-4288.
[35] Ge, H., Z. Liu, G.M. Church, and M. Vidal. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29:** 482-486.
[36] Grigoriev, A. 2001. 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* **29:** 3513-3519.
[37] Saito, R., H. Suzuki, and Y. Hayashizaki. 2003a. Construction of protein-protein interaction networks with a new interaction generality measure, In *Bioinformatics*, pp. 756-763
[38] Deng, M., S. Mehta, F. Sun, and T. Chen. 2002. Inferring domain-domain interactions from protein-protein interactions. *Genome Res* **12:** 1540-1548.
[39] Walhout, A.J., R. Sordella, X. Lu, J.L. Hartley, G.F. Temple, M.A. Brasch, N. Thierry-Mieg, and M. Vidal. 2000. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287:** 116-122.
[40] Rain, J.C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. 2001. The protein-protein interaction map of Helicobacter pylori. *Nature* **409:** 211-215.
[41] Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23:** 324-328.
[42] Jeong, H., S.P. Mason, A.L. Barabasi, and Z.N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* **411:** 41-42.
[43] Fraser, G.M., J.C. Bennett, and C. Hughes. 1999. Substrate-specific binding of hook-associated proteins by FlgN and FliT, putative chaperones for flagellum assembly. *Mol Microbiol* **32:** 569-580.
[44] Brun, C., C. Herrmann, and A. Guenoche. 2004. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* **5:** 95.
[45] Hishigaki, H., K. Nakai, T. Ono, A. Tanigami, and T. Takagi. 2001. Assessment of prediction accuracy of protein function from protein--protein interaction data. *Yeast* **18:** 523-531.
[46] Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.
[47] Muhlberger, R., R. Robelek, W. Eisenreich, C. Ettenhuber, E.K. Sinner, H. Kessler, A. Bacher, and G. Richter. 2003. RNA DNA discrimination by the antitermination protein NusB. *J Mol Biol* **327:** 973-983.
[48] Niki, H., A. Jaffe, R. Imamura, T. Ogura, and S. Hiraga. 1991. The new gene mukB codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of E. coli. *Embo J* **10:** 183-193.
[49] Carballido-Lopez, R. and J. Errington. 2003. A dynamic bacterial cytoskeleton. *Trends Cell Biol* **13:** 577-583.
[50] Kruse, T., J. Moller-Jensen, A. Lobner-Olesen, and K. Gerdes. 2003. Dysfunctional MreB inhibits chromosome segregation in Escherichia coli. *Embo J* **22:** 5283-5292.
[51] Norris, V., C. Woldringh, and E. Mileykovskaya. 2004. A hypothesis to explain division site selection in Escherichia coli by combining nucleoid occlusion and Min. *FEBS Lett* **561:** 3-10.
[52] Lauhon, C.T., E. Skovran, H.D. Urbina, D.M. Downs, and L.E. Vickery. 2004. Substitutions in an active site loop of Escherichia coli IscS result in specific defects in Fe-S cluster and thionucleoside biosynthesis in vivo. *J Biol Chem* **279:** 19551-19558.
[53] Kato, S., H. Mihara, T. Kurihara, Y. Takahashi, U. Tokumoto, T. Yoshimura, and N. Esaki. 2002. Cys-328 of IscS and Cys-63 of IscU are the sites of disulfide bridge formation in a covalently bound IscS/IscU complex: implications for the mechanism of iron-sulfur cluster assembly. *Proc Natl Acad Sci U S A* **99:** 5948-5952.
[54] Ding, H., R.J. Clark, and B. Ding. 2004. IscA mediates iron delivery for assembly of iron-sulfur clusters in IscU under the limited accessible free iron conditions. *J Biol Chem* **279:** 37499-37504.
[55] Janakiraman, A. and M.B. Goldberg. 2004. Evidence for polar positional information independent of cell division and nucleoid occlusion. *Proc Natl Acad Sci U S A* **101:** 835-840.
[56] Mrowka, R., A. Patzak, and H. Herzel. 2001. Is there a bias in proteome research? *Genome Res* **11:** 1971-1973.