修士論文２００４年度（平成１６年度）

Genome-Wide Prediction and Refinement of *E.coli* PPI Data Using Bayesian Approach

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

慶應義塾大学大学院　政策メディア研究科

中村　征良

修士論文要旨　2004 年度（平成 16 年度）

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

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予測手法は，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．これらの予測手法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た．その確率を用いると，実験データにおける個々のデータが相互作用データとしての信憑性があるかどうかを判断することができる．ベイズにおける学習には，独自に構築した既知の PPI 情報(Positive Dataset)及び相互作用しないタンパク質ペアの情報(Negative Dataset)を用いた．

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．このネットワークには，学習に用いた Positive Dataset に含まれていない文献による既知 PPI 情報が含まれていたことや，解析が進んでいる多種において構築された PPI ネットワークと傾向が似ていることからその有用性が示された．これらのデータを用いることで，機能未知タンパク質の機能予測をすることや，昨日既知タンパク質が有する新しい未知機能保持の可能性をみつけることができた．

本修士論文では，本研究を，PPI の現状とともに説明する．

キーワード：

大腸菌， タンパク質間相互作用， ベイズ統計

慶應義塾大学大学院　政策メディア研究科

中村征良

Abstract of Master's Thesis

Academic Year 2004

Genome-Wide Prediction and Refinement of ***E.coli*** PPI Data Using Bayesian Approach

Summary

Construction of genome-wide protein-protein interaction (PPI) network is very important to elucidate global picture of intracellular communications between proteins, particularly for model organisms such as *Escherichia coli*, although number of available PPIs in *E. coli* is limited so far. We constructed reliable *E. coli* PPI network using PPIs collected from the literature and those obtained from our His-tagged Pull-Down method. Characteristics of PPIs reported in the literatures were analyzed and they were used to select reliable PPIs in our experimental data. In particular, we evaluated each interaction using 7 indices, i.e. gene expression pattern, phylogenetic profiles, experimental reproducibilities, motif-motif interactions, gene essentiality, conservation of interaction in *H. pylori*, and participation of protein pairs in the same operon. Indices were integrated into probabilistic framework of Bayesian approach. Using PPIs reported in the literature as positive control, we showed that our method can select reliable interactions from set of PPIs obtained from the experiment.

We integrated the experimental PPIs refined using our method, with PPIs collected from the literature, producing as much as 3,667 interactions connecting 1,277 proteins and it provided insights into the novel functional roles of various proteins including uncharacterized ones.

Key words:

Escherichia coli, Protein-Protein Interaction, Bayesian Approach

Keio University Graduate School of Media and Governance

Seira Nakamura

-日本語版-

目次

-English Version-

Index

1. プロテオミクスにおける PPI の位置づけ

今日，ゲノムプロジェクトによって多くの生物種の遺伝子情報が公開され，解析も自由にできるようになってきたが，遺伝子情報からではわからないことは多い．DNA だけではなく，生体内には様々な分子が含まれており，それらは絶えず消費や生合成，触媒等の反応を行い合いながら生命活動を維持している．しかし，それらの分子の種類数は多大であるため，遺伝子領域（ゲノム），転写産物（トランスクリプトーム），タンパク質（プロテオーム）や代謝産物（メタボローム）に分類してそれぞれ個別の実験手法によるデータマイニングが進んでいる．そのなかでも，転写や翻訳，代謝やシグナル伝達等，多くの生体内現象に関わっているのが「プロテオーム」で，この言葉はオーストラリアマッカーリー大学の Wilkins 氏により提案された用語である．

プロテオーム情報はゲノム情報を基盤として成り立っているため，プロテオーム解析(プロテオミクス)は，ゲノムプロジェクトにより活性化された先端の研究分野であるといえる．そのプロテオーム情報のうち，我々が研究を行っているのはタンパク質が他のタンパク質と結合(相互作用)する情報であり，構造や機能等他のプロテオームの情報全てがタンパク質間相互作用(PPI:Protein－Protein Interaction)と深い関係を有しており，PPI データ取得において有益な情報となる．なお現在において行われているプロテオミクスには PPI を含めて以下のようなものがある．

1-1.タンパク質間相互作用(Protein-Protein Interaction)の抽出

生体内におけるタンパク質は，単体で機能を有しているということは少なく，多くが他のタンパク質と相互作用をすることで初めて機能を有する場合が多いことが知られている．その場合，機能の単位はタンパク質というよりは，それらの相互作用によって形成された複合体であるといってよく，タンパク質間相互作用(PPI: Protein-Protein Interaction)の情報は，機能情報を得るために非常に有益な情報である．また，リン酸化等のシグナルに関わるタンパク質間の物理的な接触も，相互作用によるものであり，それらの情報を得ることは，代謝やシグナル伝達等のシステム解明に非常に有用なことであり，現在ではその抽出作業が広く進められている．PPI に関する実験手法としては，Yeast Two-Hybrid 法や TAP 法といった相互作用の関係抽出法や，NMR や X 線結晶解析のような複合体の立体構造情報を得る，様態を知るための方法があげられる．特に酵母菌においては，PPI の実験データが豊富に揃っている．[3]

1-2.アミノ酸配列からの遺伝子の同定

遺伝子の同定とは，標的のタンパク質をコードしている遺伝子領域を見つける作業をいう．手法としてはペプチドマスフィンガープリンティングや気相シーケンサーによるアミノ酸配列決定が用いられる．ゲノムの塩基情報があれば，翻訳されるアミノ酸配列は予測できるため，それらと照合することで，もととなる遺伝子領域を決定することができる．遺伝子領域を決定するということは，DNA における位置情報を得られることになる．とくにタンパク質ペアにおける転写の際の物理的な距離は

PPI と深い相関があることがわかっており，事実，原核生物において見られる同じオペロンにより転写されるタンパク質グループは互いに相互作用している可能性が高い．[48]

2001 年には大腸菌において，364 スポットの N 末端アミノ酸配列が気相シーケンサーにより決定された[1]．結局 429 本もの配列が決定され，ORF も 90%以上同定された．

1-3.タンパク質構造予測

たとえ，アミノ酸配列の全く違っているタンパク質ペアであっても，それぞれが同じような機能を有するケースが存在することが明らかになってきた．それはタンパク質の機能が必ずしも一次構造によるものではないことを意味しており，今後はタンパク質の機能と三次構造(立体構造)の相関を得る研究が発展していくと考えられる．

立体構造を実験的に直接調べる手法は NMR や X 線結晶解析などが存在する．NMR は，隣り合う原子間の距離や角度を検出する方法で，コストや時間を要することや，高分子を調べることが不可能であることなどが問題視されている．また一方，短波長の電磁波を当てて位相を調べる方法であるX線結晶解析は，対象分子を結晶化させる必要があり，動的な観察ができないうえ，実際の細胞内環境とは全く異なるという問題点が指摘されている．しかしながら，それらの立体構造情報とアミノ酸配列の情報からは，一次構造⇔立体構造の変換方法を得ることができると期待できるため，それらの規則性を検出し，ホモロジー法や 3D-1D 法[2]など，タンパク質の立体予測を行う手法の開発が進んできている．ホモロジー法というのは，既に立体構造情報が得られているタンパク質と相同性の高いアミノ酸配列が他のタンパク質に存在した場合，その両タンパク質は，立体構造的に似ていることがいえるため，立体構造情報が得られているタンパク質のアミノ酸配列と，その他のタンパク質の相同性を調べていく方法である．タンパク質の一次構造の情報は豊富に揃っているため，この手法は広く使用されている．3D-1D 法というのは，立体構造情報とアミノ酸配列情報の適合度を示す評価関数を作成し，構造と配列をアラインメントすることでスコアをつけ，スコアが上位のものを選ぶ方法である．しかし，どの手法を用いたとしても，タンパク質の構造には「揺らぎ」が存在するほか，結晶化された構造ですら未だ明快な規則性が得られていないため，現状では最終的な解明には程遠いといえる．

タンパク質単体や複合体の立体構造を網羅することで，各タンパク質が有する相互作用部位を立体的に把握することができる．もしアミノ酸配列から信頼性の高い立体構造予測ができるようになり，なおかつそこから相互作用に関わる部位を特定することが可能になれば，相互作用を行う予測が容易になると考える．

1-4.タンパク質機能予測

プロテオミクスにおける最終的な目的はタンパク質の機能特定であると考える．転写や翻訳，触媒する代謝反応，シグナル伝達等の機能情報などが明らかになった場合，その情報はゲノムやトランスクリプトーム，メタボローム等の情報の補充にも利用でき，生命理解への大きな一歩となる．現在，タンパク質機能を決定付ける事象(立体構造やタンパク質間相互作用など)を用いたタンパク

質機能予測が頻繁に行われており，その最も主流の機能予測手法は，既知機能タンパク質との相同性を調べる方法や，モチーフを用いたものである．モチーフとは，共通の機能を有しているタンパク質群に共通して存在している配列のことである．網羅的なタンパク質機能検出実験方法は未だ開発されていないが，アフィニティークロマトグラフィーで精製されるタンパク質など，特定のクロマトグラフィー条件で同時に精製されてくるタンパク質の MS/MS ショットガン分析もタンパク質機能解析に利用されている．

1-5. プロテオミクスに必要なデータ

プロテオミクスに必要な情報は，現在オンラインで取得できるものが多い．特に有効な情報は，アミノ酸の配列情報や構造情報，アノテーション，機能を決定付ける配列情報であるモチーフやドメイン情報，そしてタンパク質間相互作用情報などである．以下にそれらのデータベースを紹介する．

・アミノ酸配列情報・アノテーション

SWISS-PROT (http://www.ebi.ac.uk) SIB と EMBL-EBI の共同維持サイト
PIR (http://pir.georgetown.edu)　NBRF
GenPept (http://ncbi.nlm.nih.gov)　NCBI
PRF/SEQDB (http://www.prf.or.jp)　タンパク質研究奨励会

・構造情報

Protein Data Bank(PDB) (http://www.rcsb.org/pdb/)　RCSB
BioMagResBank (http://www.bmrb.wisc.edu/)　University of Wisconsin-Madison

・モチーフやドメイン情報

PROSITE (http://www.expasy.ch/prosite/)　SIB
BLOCKS (http://blocks.fhcrc.org/)　FHCRC
PRINTS (http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/)　UMEBR
ProDom (http://protein.toulouse.inra.fr/prodom.html)　INRA
Pfam (http://www.sanger.ac.uk/Software/Pfam/)　Sanger Institute
InterDom (http://interdom.lit.org.sg/)　IICR

・タンパク質間相互作用情報

MIPS (http://mips.gsf.de)　GSF
DIP (http://dip.doe-mbi.ucla.edu)　UCLA
BIND (http://www.blueprint.org/bind/bind.php)　Blueprint

各データベース間において情報のリンクが存在するなど，連携しているケースがあるため，すでに明らかにされているタンパク質のさまざまな情報を引き出すことができる．立体構造やタンパク質間相互作用のデータベースからは，専用の描写ツール(Viewer)の取得が可能である．また，これらのデータベースにはフラットデータによるデータの取得ができるところが多いため，解析研究者の都合によるオリジナルな解析に利用することが可能である．SWISS-PROT などには，データの重複や，あいまいな情報を登録しないようなシステムもあるため，比較的安心してダウンロードと解析を連続して行うことができる．

2.Protein-Protein Interaction

この章では，タンパク質間相互作用(PPI)の実験やその他解析による検出法や，取得可能なPPI データの詳細(公開データ，フォーマット，使用目的等)を紹介する．

2-1.生体内における PPI の役割

タンパク質の機能予測には，タンパク質複合体情報を取得することが有用であることは前章で述べた．複数のタンパク質が相互作用しあうことで初めて機能を有する原因としては，複合体を構成している各タンパク質がそれぞれ複合体としての機能に必須な機能ドメインを有しており，それぞれが合わさることでその機能ドメインが統合され，初めて複合体として完成した機能を得るためだと考えられる．例えばRNA ポリメラーゼはDNA からmRNA を合成する(転写する)役割を果たしているが，その機能を成し遂げるためにはプロモーターを読み取る部位や，DNA への結合部位や RNA への結合部位(触媒)，そしてそれらをつなぎ合わせる部位がそれぞれ必要となる．そのため，どれか一つのタンパク質が欠けた場合，その機能を失ってしまう．この PPI による複合体機能の調節は，医薬品開発においても利用される．RNA ポリメラーゼは全ての生物の生命活動に必須なタンパク質である．そのため，もし病原性細菌の RNA ポリメラーゼにだけ特異的に結合する化合物をつくれば，それは病原性細菌の生命活動を停止させる抗生物質となる．このような抗生物質だけでなく，活性制御物質の開発など，医療分野においてもPPI という現象は期待されている．

また，細胞運動や細胞分化等に大きく関わっているシグナルは，シグナル伝達系を相互につなぎ合わせたネットワークを形成しており，外部環境の変化に対して柔軟な対応を実現する．そのシグナルの応答にはタンパク質が大きくかかわっており，タンパク質のリン酸化等のシグナル伝達に関わる機能も，複数のタンパク質におけるドメインの組み合わせにより構築されるため，それらの実態の解明には PPI の情報が不可欠である．これらによるシグナル伝達系ネットワークの解明は，ガンや免疫不全，神経変性疾患等の疾患の原因を明らかにすることができるだけでなく，それらにゆうような薬剤の開発にも非常に有用である．

このように，PPI データからは多くの生物学的な知見を得ることができる．今日では，以下のような実験手法を用いてPPI データの取得が進んでいる．

2-2.実験による PPI 検出法

・Pull-Down 法[11]

我々が本研究において用いた大腸菌における実験PPI データの検出法である．TAP 法と同じ原理を利用した手法で，FLAG タグや GST タグ，His タグなどを用いて，任意のタンパク質との結合を検出する *in vitro* 手法である．細胞や組織の溶解液とタグ付きタンパク質を混ぜ合わせることで，タグつきタンパク質に任意のタンパク質が相互作用する．それらを免疫沈降で分離し，SDS-PAGE もしくは二次元電気泳動を用いて解析する．検出された未知のバンドに関してはペプチドシーケンスや質量分析器を用いて同定する．なお，我々が本研究において用いたのは大腸菌における 2669

個のタンパク質にHisタグを付けたものを過剰発現させて使用したものであり，それらに相互作用をして得られたタンパク質のバンドの同定にはSDS-PAGEを用いた．これによって得られたPPIデータには804個の機能未知タンパク質が含まれていた．（図2-1）

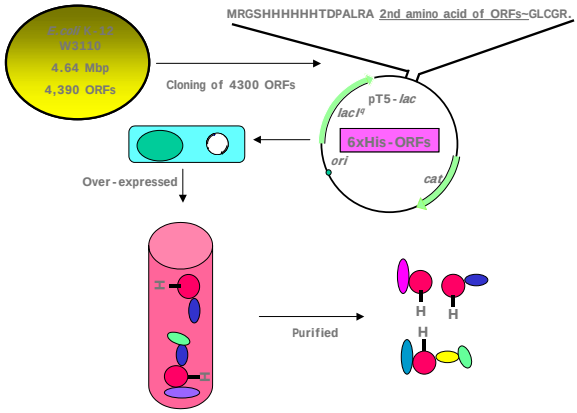


図 2-1: His-tag タンパク質を用いた Pull-Down 法の概念図（提供:伊藤文氏）

PPIにおける結合力としてあげられるものとして，「疎水力」「静電力」「水素結合」等がある．もちろん相互作用しているタンパク質ペアそれぞれの全体の形状による物理的な結合もあると考える．タンパク質間における相互作用の結合力は，生化学反応式と同様な考え方で，解離平衡で取り扱う．タンパク質A(Pa)とタンパク質B(Pb)が相互作用しているときには，

$$[Pa]+[Pb] \Leftrightarrow [PaPb]$$

$$Kd = \frac{[Pa][Pb]}{[PaPb]}$$

と表現する．これによると，複合体が半分生成する点で$[Pa]=[Pb]=[PaPb]$となり，$Kd=[Pa]=[Pb]$となる．シグナル伝達等，他のタンパク質との相互作用時間が相互作用していない時間と比べて非常に少ないものだった場合は，この Kd は非常に小さくなると考えられる．Pull-Down法によって検出されたタンパク質間相互作用は，タグ付タンパク質と結合した後でSDS-PAGEによって確認されるまで常に結合しているようなタンパク質であるため，この Kd は比較的小さい値であると考えられる．

・Yeast Two-Hybrid 法[4][5][6][7]

転写活性タンパク質は DNA 結合ドメインと活性化ドメインの2つのドメインを有している. DNA 結合ドメインはプロモーターに結合するドメインであり, 活性化ドメインはRNAポリメラーゼとの結合ドメインでもある. この転写活性タンパク質を, DNA 結合ドメインを持っている部分のポリペプチド部品と活性化ドメインを持っている方のポリペプチド部品の2つに分断したとき, この2つのポリペプチドは, 互いに物理的につながらなければ転写の機能として働くことができない. 逆を言えば, つながればその複合体は転写機能を取り戻し, mRNA が合成されることになる. よって, DNA 結合ドメインと活性化ドメインそれぞれと, 相互作用を検出したい2つのタンパク質(BaitとPrey)が結合したキメラタンパク2種類(Bait+転写活性タンパク質の DNA 結合ドメインを含む側のポリペプチド)(Prey+転写活性タンパク質の活性化ドメインを含む側のポリペプチド)を翻訳する2つの遺伝子を設計し, 発現させる. もし Bait と Prey が相互作用した場合, DNA結合ドメインと活性化ドメインが同じ複合体に共存していることとなり, 転写が開始される. もし相互作用しない場合は, 転写が開始されることがないため, タンパク質間の相互作用の有無がはっきりと検出できる. (図 2-2) 今現在, Yeast Two-Hybrid 法は, ゲノムワイド PPI 検出法として最も主流な手法であるといえる.

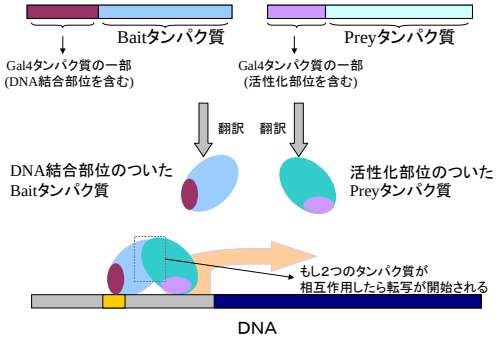


図 2-2: Yeast Two-Hybrid 法の概念図

・TAP(Tandem Affinity Purification)法[8][9][10]

TAP 法はタグをつけたタンパク質(Prey)を発現させ, そのタンパク質に相互作用している全ての標的タンパク質(Bait)を複合体情報として取得することができる方法である. タグ配列特異的なタンパク分解酵素を用いて非特異的なタンパク結合を除去することで標的タンパクの精製度を高めることができるのが特徴である.

・プロテインチップ法[12]

プロテインチップとは, 多種類のタンパク質又はタンパク質と相互作用をする物質を高密度で配列したものであり, タンパク質間相互作用の高感度かつハイスループットな検出法として注目を浴びている. [12]チップ上のスポットにリガンドを固定してタンパク質を含む溶液を加えたのち, 質量分析器で分析する. 検出方法には, 蛍光標識をつける方法などもあるが, それによるタンパク質の失活や検出エラー等の問題点がある. また, 高価格であり, コスト面においても問題視されている.

・免疫沈降法

抗体に対応する抗原タンパク質のみを検出する方法だが, lysate の界面活性剤を調節することで, 抗体に直接結合する抗原タンパク質と一緒に複合体を作っているタンパク質を検出することが可能となる. *in vivo* での PPI を検証する際に使用される方法であるが, 非特異的結合によるノイズを

除去することができない.

・ファージ・ディスプレイ法 [13][14]

バクテリオファージゲノムに, 標的タンパク質をコードする遺伝子を挿入し, ファージタンパク質に融合したタンパク質をファージ分子の表面に位置させ, そのタンパク質に結合する標識分子を試験管内で選出する方法であり, 細胞膜や分泌タンパク質と相互作用する分子の同定に適している.

two-hybrid などにより検出することができる１対１の結合に対し，Pull-Down 法や TAP 法においては細胞内に bait タンパク質を特殊なタグ配列とともに発現させ，結合しているすべての標的タンパクを検出することができ，複合体データとして扱うことができる．また，two-hybrid により検出されたデータは，他のタンパク質による仲介を経て, 間接的に転写因子としての複合体を形成している可能性があるため, 実際に直接相互作用しているとは断言できない．それに対し，Pull-Down 法や TAP 法によるデータは, もともと複合体のデータであるため，そのような考慮をする必要が無い．Pull-Down 法や，TAP 法におけるデータの欠点は，複合体に含まれるタンパク質群のうち，どの分子とどの分子が具体的に相互作用しているのかがわからないところであろう．

図 2-3: Spoke Approach の概念

図 2-4: Matrix Approach の概念

このように，どれも PPI を検出するための実験手法であるが，その出力内容は変わってくる．それぞれの出力データが含む傾向が違えば，解析手法も変える必要があるため，注意する必要がある．複合体データの場合は，含まれるタンパク質それぞれが直接もしくは間接的に相互作用していることがわかる．それらを 1 対１の相互作用にして表現(バイナリー表記)するために用いられる手法には Spoke Approach(図 2-3)と Matrix Approach(図 2-4)の２つが存在する．Spoke Approach は，結合して検出された Prey タンパク質全てが Bait タンパク質と結合していると考える手法[4][5][15]で，Matrix Approach は，複合体に含まれる全てのタンパク質が総当りで互いに相互作用していると考える手法である[4][5][15]．そうして複合体のデータをバイナリーに変換し，その後の解析により実際の相互作用を模索していく．また，これらの実験手法にはどれも多くの偽陽性データが含まれていることが問題視されており，いかにして信頼性の高いデータのみを抽出するのかが模索されている．

2-3.現在公開されている PPI データ

PPI データには大きくわけると 2 種類あり，1 つは相互作用している様子(高次元描写)を表現するデータで，もう一つは相互作用しているタンパク質の関係を表すデータである．例えば，前者の例では，転写に関わっているヒト RNA ポリメラーゼ(図 2-5)や，翻訳に関わっている大腸菌リボソーム(図 2-6)などは，生体内においてタンパク質複合体を形成しており，立体構造や各タンパク質間相互作用の詳細が明らかになっている．後者のほうでは，Y2H 等の実験手法を用いた結果を用いてタンパク質間の相互作用を 1 対 1 で表記しているものである．例えばタンパク質 A(ProteinA)とタンパク質 B(ProteinB)が相互作用していることを表現したい場合，多くのサーバーにおいて，その表記は「ProteinA　ProteinB」のような，1 行につき 1 相互作用のデータが記載されている．そのため，トランスクリプトームやめたボローム等の他のデータとからめたオリジナルの研究に対して非常に柔軟なフォーマットであるといえる．とくに，相互作用しているタンパク質同士を線でつないだネットワークを形成してクラスター認識を行ったり，機能との相関を得たりする解析が主流である．これらの相互作用データをネットワーク表示するためのツールも数多く公開されている．(図 2-7)

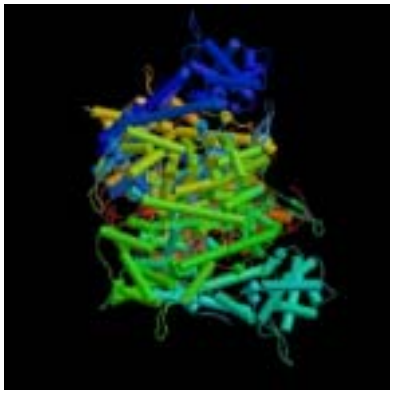


図 2-5: T7 RNA polymerase elongation 複合体の立体構造

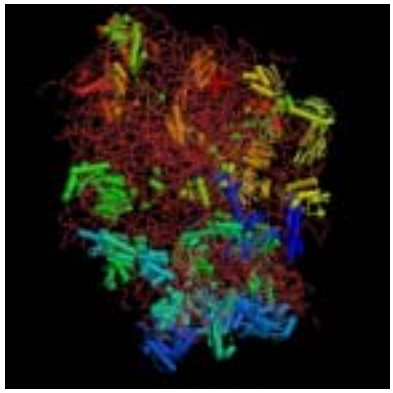
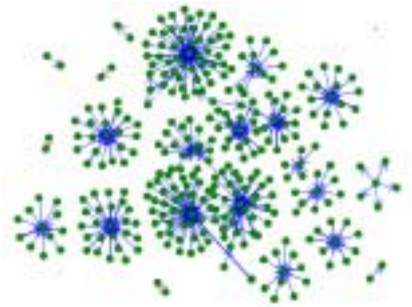


図 2-6: 大腸菌におけるリボソーマルタンパク質の立体構造

(A)



(B)

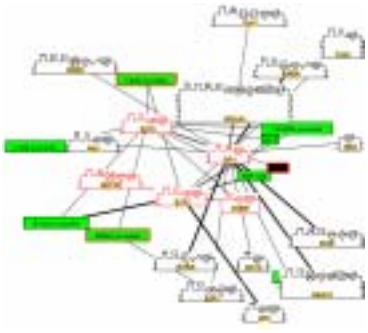


図 2-7: 相互作用ネットワークを表示するツール．左側(A):が BioLayout で右側(B)が BIND における BIND Interaction Viewer v.3.2

現在取得できるタンパク質の立体構造データは，PDB等のデータベースから取得可能である．図 2-5,図 2-6 のような PDB から得られる画像データもあれば，各アミノ酸の座標情報が記載されているテキストデータも入手可能である．タンパク質の基本構造が網羅的に決定されようとしており，PDB に登録されている立体構造のデータ数は 29,000 件を超え(2005/1/10 現在)，登録件数は加速している．また，その中には複合体のデータも揃っており，立体構造情報の幅は広がっている．複合体のデータからは，結合部位の立体的な解釈をすることができる．アミノ酸が関わっている全ての結合様態を捕らえて，表現することができれば，アミノ酸配列からそのタンパク質の相互作用を予測することが可能になる．また今日では，EMBL-EBI によってタンパク質-タンパク質のドッキング構造の予測コンテスト「CAPRI(Critical Assessment of PRediction of Interactions)」(http://capri.ebi.ac.uk/)等も開催されており，予測手法の模索が世界中で行われている．

一方，1 対 1 の相互作用関係を表すデータは DIP や BIND 等から取得することができる．アミノ酸配列とタンパク質の立体構造の相関を網羅的に得るための立体構造データとはちがい，生物種による分類の概念が強い．取得できるテキストファイルも，生物種単位で格納されている．各データベースに格納されている相互作用数(2005/1/11 現在)は，BIND：131,133 個，DIP: 44,482 個と非常に豊富になってきている．特徴としては，BIND は扱っている生物種が非常に豊富であるが，DIP では，扱っている生物種に制限を設けている．DIP における生物種毎の相互作用数は，ショウジョウバエが 20,988 個，酵母菌が 15,675 個，線虫が 4,030 個，ピロリ菌が 1,425 個，ヒトが 1,379 個，大腸菌が 611 個となっており(2005/1/10 現在)，原核生物に比べて真核生物における相互作用データが非常に多い．これらの相互作用関係データからは，相互作用を connection で表現したタンパク質間のネットワーク(PPI ネットワーク)を構築することができる．構築した PPI ネットワークからは，機能間のクロストークや，新しい複合体の発見，各タンパク質の生体内機能における重要性などがわかってくる．たとえば，相互作用相手の数が多いタンパク質は，ネットワークを構築するとその密度から容易に見つけることができ，生体機能内においてそのタンパク質が非常に重要な役割を持っていることを予測することができる．また，そういった密度の大きいネットワークに含まれているタンパク質群は，生体内において何らかの連携した機能を働いているとも考えられる．このように，PPI ネットワークを用いることで，スケールフリーにタンパク質の機能を考察をすることができる．

2-4.種間における PPI の保存性

PPI データは種間によってことなっている. 現在, 完全にゲノム配列が読まれている生物種は 250 種を超えているが, それらの PPI データを網羅することは, 途方にくれる作業である. そこで, 遺伝子機能の同定においても, 生物種間で遺伝子領域における塩基配列のホモロジーを調べてその保存性をもとに機能の予測がされているように, ほとんど PPI 情報の得られていない生物種から PPI 情報を予測する場合は, PPI 情報が豊富に得られている生物種との間における保存性を利用した予測を行う[17]. もちろん, タンパク質単体の保存性と PPI の保存性というのは傾向が異なっている. 真正細菌や古細菌においては, 基本的な代謝物質の生合成やエネルギー代謝等に関わっているハウスキーピング遺伝子群に関して, 水平伝播が頻繁に起こっていることがわかっている[16]. それに対し, 転写や翻訳などに関わっており, 機能を得るために複合体を形成する必要があるタンパク質は水平伝播の率があまり高くない. その差から, 単独で機能するタンパク質の方が, 他種に水平伝播の受容をされやすいことが言われている. ようするに, タンパク質の保存性はその機能に著しく影響を受けることが示唆される. また, 酵母と線虫間に置いて, 相互作用しているタンパク質ペアの共進速度を測ったという報告がある[17]. 各生物種が有しているタンパク質のオーソログ情報を配列の相同性をもとに取得し, それに関わる相互作用を比較解析したものである. その結果, タンパク質単体よりも, 相互作用しているタンパク質ペアの共進速度は, 単体のそれに比べて若干遅いことが考察されている. そのことから, 進化速度は相互作用相手の影響を少なからず受けており, 相互作用相手が多ければ多いほどその進化速度は遅いことが示唆される. 種間において保存されている相互作用は, Ortholog と Interaction という言葉を合わせて Interolog と呼ばれ, この情報を用いた PPI 予測手法も提案されている.

2-5.ゲノムワイド PPI データから取得できる情報

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，そのデータをゲノムワイドに収集することは，生命理解にとって非常に不可欠な要素であるといえる．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相互作用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[18][19]や，代謝[20][21]やシグナル伝達経路[22][23][24][25]の予測を行うことができることを意味する．なお現在，ショウジョウバエや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．

次章以降において，我々が大腸菌における PPI 実験データを用いたゲノムワイド PPI データ構築法を紹介する．

3.信頼性の高いゲノムワイドな大腸菌 PPI データ構築

3-1. 実験 PPI データ

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)[4][5][6][7]や TAP(Tandem Affinity Purification)[8][9][10]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 に関しても「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[26][27]ことが知られている. このようなデータをもとに生物学的な知見を得ようとすると, 誤った解釈につながってしまうため[28], いかにして信頼できるデータセットを構築するかということが重要であり, そのための様々な手法が開発されている[29][30][31][32][33][34]. 本研究における目的も同じであり, 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 データを抽出することで, 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.

DIP や BIND において, PPI が豊富に揃っている生物種は主に真核のモデル生物である線虫やショウジョウバエなどであるが, 原核のモデル生物である大腸菌においてはその数は 600 個程度であり, 豊富であるとはいえない. 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. いかに相互作用の情報だけが揃っていたとしても, 遺伝子の同定や機能同定, 代謝の情報等が無ければ, 各相互作用の意味を得ることができないからである. 原核のモデル生物である大腸菌はその点に関する条件を大きく備えているため, 本研究における対象生物とした.

本研究において, 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) [11]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[29]). 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.

3-2. 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じ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．

3.2.1 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．

・Positive Dataset(既知 PPI 情報)

PPI の公的データベース DIP (http://dip.doe-mbi.ucla.edu/)[35][36]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(図 3-1)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 のデータ及び EcoCyc (http://ecocyc.org/)[37][38]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．

図 3-1: 作成した Curation CGI

・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/)[39]から取得し，発現場所の情報は，PSORT[40]を用いて調べた．PSORT は，各タンパク質のアミノ酸配列から細胞内の局在情報を予測することを使用目的としたツールである．酵母における予測結果の信頼

度は 57%であるが，大腸菌における予測結果の信頼度は 86%にものぼる．その信頼性は十分であると考え，大腸菌において全タンパク質の局在情報を予測した．その結果，2,277,085 個の Negative Data を取得することができた．

以上のプロセスを経て取得したこれらの PPI データセットの分布を表す．(図 3.1)

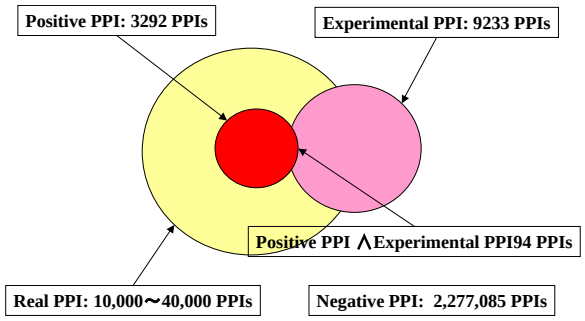


図 3-2: 得られた PPI データセットの分布　Positive PPI∧Experimental PPI は，実験データに含まれている Positive Data の数である．

このように取得した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

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

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

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 の値になる確率である．

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．

3.2.2 Phylogenetic Profiling(系統プロファイル)

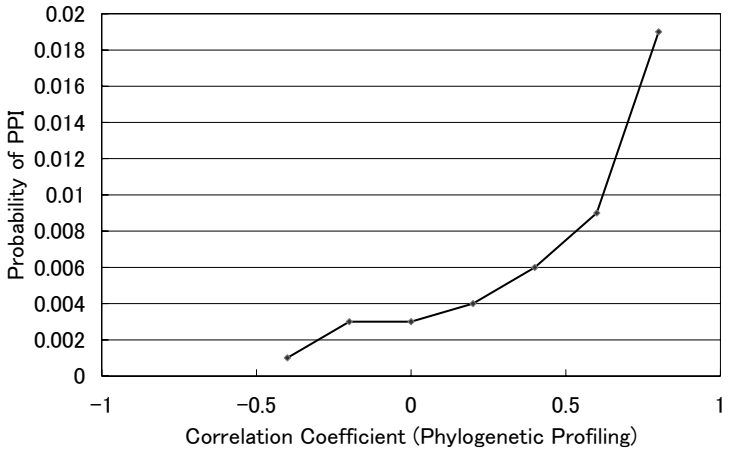
機能的につながりを有するタンパク質ペアは，他生物種における存在情報も似通っている[41]．ようするに，ほぼ同じ進化過程をたどっている傾向が高いということになる．大腸菌の各タンパク質の進化過程パターンを調べるために，大腸菌における各翻訳産物のアミノ酸配列と，全ゲノム配列が解読された 42 生物種(古細菌 6 種，真正細菌 36 種)における(表 3-1)各翻訳産物のアミノ酸配列の相同性を，BLAST を用いて調べ，大腸菌のタンパク質と系統的につながりを持つものを BLAST の出力である E-value を用いて表現した．例えば，大腸菌におけるタンパク質 A,B,C があったときに，それらを他の生物種 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 による配列の相関係数を算出することで調べた．(図 3-3) ここで気をつける必要があるのは，他生物種の選択である．各配列に対して同じ傾向を持った要素が存在した場合，各配列間の相関のばらつきが少なくなる．これは当初我々が 100 種程度の生物種を用いて行った場合に見られた傾向であるが，選択した生物種が近縁種を多く含んだ場合に，各タンパク質におけるその進化的情報の類似度があがる．相関係数は，数的配列の推移の類似度を測るための係数であるから，もちろんその値も上昇する．そのため，1 属につき 1 種を選ぶこととし，近縁種を省く作業を行い，最終的に 42 生物種にした．以上から，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 3-4)

各タンパク質における他生物種間における存在情報をE-valueの配列にして表示

	生物種A	生物種B	生物種C	生物種D
タンパク質A	5.00E-05	2.00E-22	1.00E-22	1.4
タンパク質B	2.00E-27	5.00E-33	0.81	3.00E-36
タンパク質C	1.00E-43	5.00E-42	3.00E-28	2.00E-62
タンパク質D	0.62	8.00E-26	7.7	3.00E-48
タンパク質E	5.00E-05	2.00E-22	3.00E-42	3.2

相関係数を算出し、タンパク質ペアのスコアとする

図 3-3: Phylogenetic Profiling によるタンパク質ペアのスコア算出方法



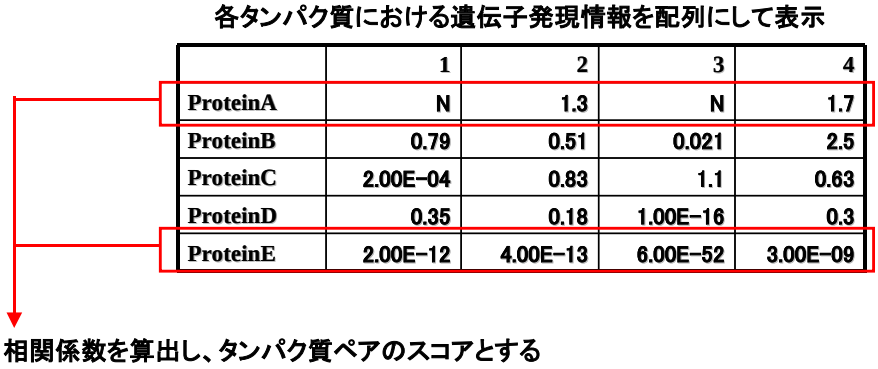
(図 3-4)
Phylogenetic Profilig におけるスコア(横軸)と, Probability(縦軸)の相関グラフ

表 3-1: Phylogenetic Profiling において使用した生物種(A は古細菌, E は真正細菌を表す)

Aeropyrum_pernix	A
Agrobacterium_tumefaciens_C58_Cereon	E
Aquifex_aeolicus	E
Archaeoglobus_fulgidus	A
Bacillus_subtilis	E
Borrelia_burgdorferi	E
Buchnera_sp	E
Campylobacter_jejuni	E
Caulobacter_crescentus	E
Chlamydia_muridarum	E
Chlamydophila_pneumoniae_AR39	E
Clostridium_acetobutylicum	E
Deinococcus_radiodurans	E
Haemophilus_influenzae	E
Halobacterium_sp	E
Helicobacter_pylori_J99	E
Lactococcus_lactis	E
Listeria_innocua	E
Methanobacterium_thermoautotrophicum	E
Methanococcus_jannaschii	A
Mycobacterium_leprae	E
Mycoplasma_genitalium	E
Neisseria_meningitidis_MC58	E
Nostoc_sp	E
Pasteurella_multocida	E
Pseudomonas_aeruginosa	E
Pyrococcus_furiosus	A
Rickettsia_conorii	E
Salmonella_typhi	E
Sinorhizobium_meliloti	E
Staphylococcus_aureus_Mu50	E
Streptococcus_pneumoniae_TIGR4	E
Sulfolobus_solfataricus	A
Synechocystis_PCC6803	E
Thermoplasma_acidophilum	A
Thermotoga_maritima	E
Treponema_pallidum	E
Ureaplasma_urealyticum	E
Vibrio_cholerae	E
Xylella_fastidiosa	E
Yersinia_pestis_CO92	E

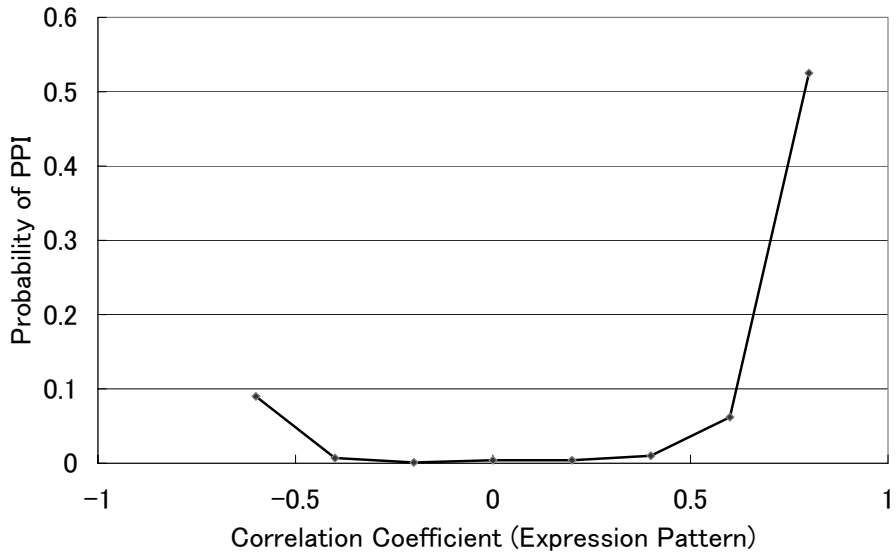
3.2.3 Expression Pattern(遺伝子発現相関)

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[13][42]. 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/)と 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. (図 3-5)そのことから，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 3-6)



	1	2	3	4
ProteinA	N	1.3	N	1.7
ProteinB	0.79	0.51	0.021	2.5
ProteinC	2.00E-04	0.83	1.1	0.63
ProteinD	0.35	0.18	1.00E-16	0.3
ProteinE	2.00E-12	4.00E-13	6.00E-52	3.00E-09

図 3-5:Expression Pattern を用いたタンパク質ペアのスコア算出方法



(図 3-6)タンパク質ペアの発現の相関係数(横軸)と Probability(縦軸)の相関グラフ

3.2.4 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[33][43]. 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. その出力スコアの大きさにより，0 を「0」，1 もしくは2を「1」，2 以上を「2」と分類して Probability の算出に使用した. なお，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[15]. 各分類における Probability を算出した結果，IG と Probability には正の相関がみられ，PPI 予測手法としての有効性が示された. (表 3-2)

3.2.5 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[44]. 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. 各分類における Probability を算出した結果，MMI の PPI 予測手法としての有効性が示された. (表 3-2).

3.2.6 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[45]. Essential protein 周辺の PPI ネットワークを Type A: 周りのタンパク質がすべて essential protein であるもの，Type B: 周りのタンパク質が全て non-essential protein であるもの，Type C: 周りのタンパク質が essential protein と non-essential protein 両方混ざっているもの3つに分類した場合，Type A と Type C が 75.02%と全体の大半を占める. とくに Type A の代表となるものが，巨大な複合体を形成しているといわれている ribosomal protein 群である. (図 3-7)

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. この場合においても，各分類における Probability を算出したところ，Essentiality と PPI には大きな相関があることが示された. (表 3-2)

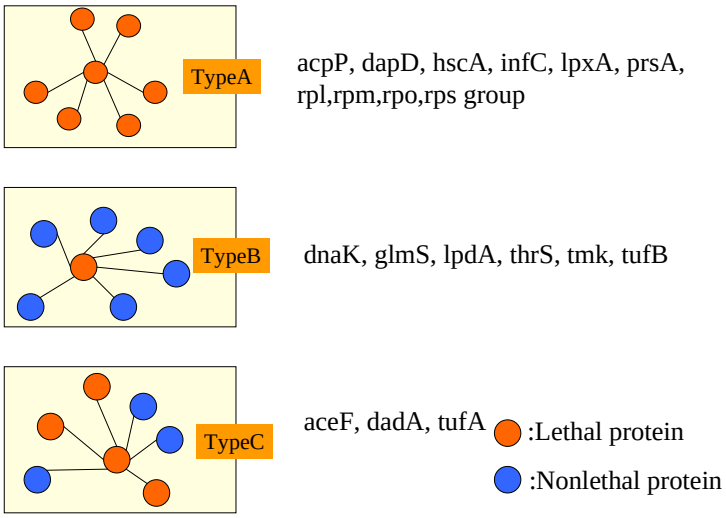


図 3-7 Essential protein の，周辺 PPI ネットワークによる 3 分類と，各分類を代表する遺伝子名及び遺伝子群.

3.2.7 Interolog (他生物種 PPI データの参照)

相互作用を行うタンパク質ペアには，進化における共進性がみられる[17]. そのような種間における相互作用の関係は Interolog と呼ばれている. この Interolog 情報を他生物種における PPI データから相互作用を行うタンパク質ペアのオーソログを取得することで取得し，目標生物種(ここでは大腸菌)の PPI を予測した [46]. 他生物種における PPI データは，ゲノムワイドに取得されたピロリ菌の PPI データ[47]を使用した. オーソログ情報は，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. 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. (表 3-2)

3.2.8 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[48]. 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している. 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. 結果，高い PPI との相関が得られた. (表 3-2)

表 3-2: 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

3.3 PPI 予測手法の統合と評価

3.3.1 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(図 3-8), 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.

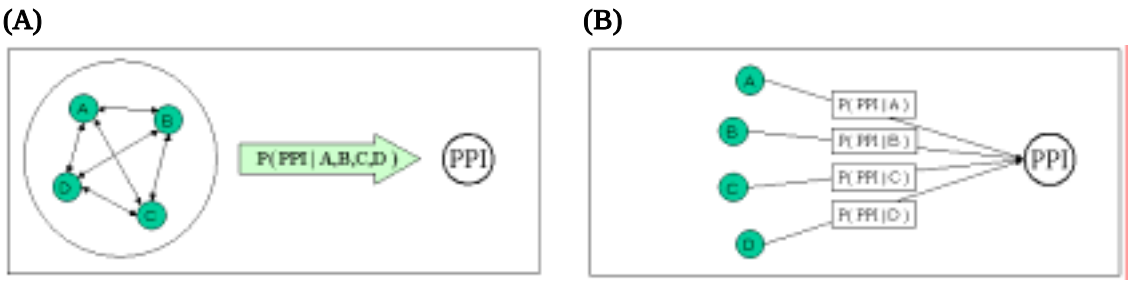


図 3-8: Fully Connected Bayes(A)と Naïve Bayes(B)の概念図

統合した予測手法における出力の確率変数 Probability(P(*positive* | *ES*1,*ES*2,*ES*3…))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.

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

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

3.3.2 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 3-9)，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.

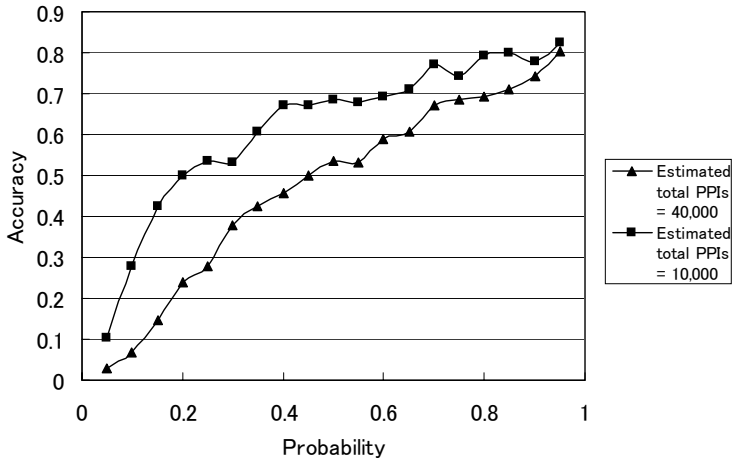


図 3-9: Probability(横軸)とその Accuracy(縦軸)の相関
(大腸菌における全ての PPI データ数を 10,000 とした場合と 40,000 とした場合)

3.4 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 3-10 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.

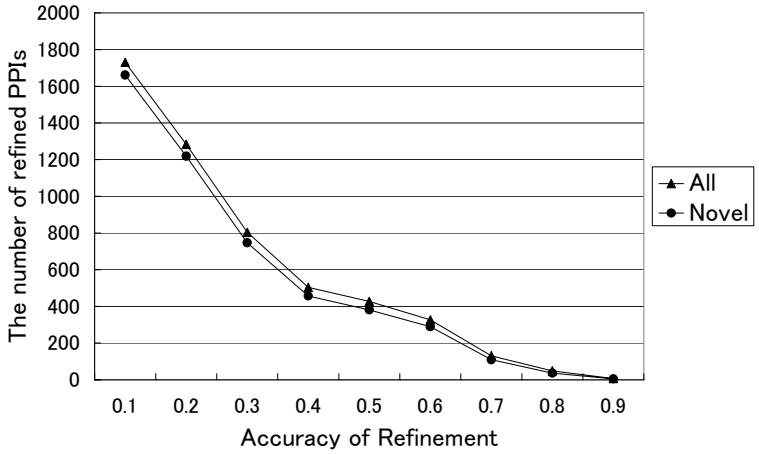


図 3-10: PPI 精製における Accuracy(横軸)と, 取得できる PPI 数(縦軸)の推移(ALL:取得できる PPI 数 Novel:取得した PPI データに含まれ，Positive Data に含まれない新規 PPI データ)

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 3-3 である.

表 3-3：実験 PPI データと精製後 PPI データの内訳

	総数	既知データ	新規 PPI データ
実験 PPI データ	9233	94	
精製後 PPI 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 Probability が 0.5 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.

また，精製の対象とするデータを実験 PPI データではなく，大腸菌における全てのタンパク質ペアとした場合，ゲノムワイドな PPI 予測をすることができる．(表 3-4)

表 3-4：大腸菌における全てのタンパク質ペアと，それを対象にした PPI 予測結果の内訳

	総数	既知データ	新規PPIデータ
大腸菌における全てのタンパク質ペア	9290205	3589	
予測後PPIデータ	1407	330	1077

対象とするデータは, 9,290,205 個であり, 実験 PPI データ(9233 個)と比べておよそ 1000 倍とはるかに大きくなっているが，予測結果数は 3.5 倍程度である．これは実験 PPI データが無作為にえらんだタンパク質ペアではなく, PPI としての集合の特性を兼ね備えていることを裏付ける結果であるとも考える．これらによって得られたデータは，実験による裏付けが得られていないため，考察に用いるデータセットとしては扱わないこととした．逆に，実験による裏付けが得られている，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, 我々は最終的に, 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した．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 3-11 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[45].

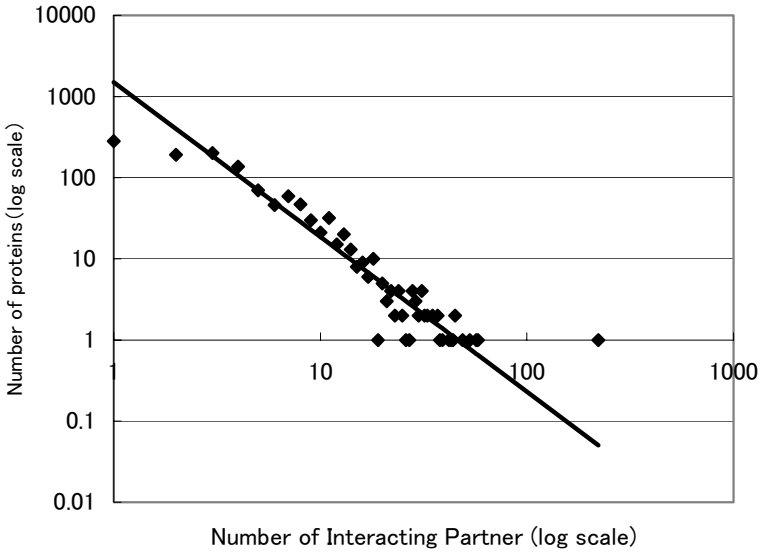


図 3-11: 相互作用相手タンパク質の数(横軸)と，相互作用相手をその数所有しているタンパク質の数(縦軸)の相関

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[45]．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．

3.5. 考察

以上より，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[49])を新たに文献により見つけることができた．

3.5.1 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[27][50][51][52]．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[33]．(図 3-12)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[34]．表 3-5 が，9 個の機能未知タンパク質の機能予測結果である．

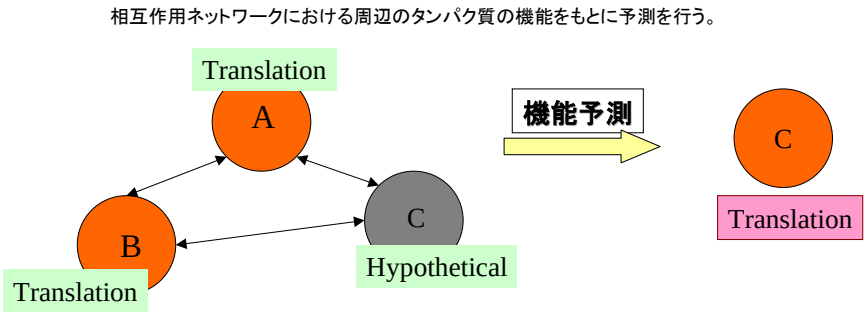


図 3-12: PPI ネットワークを用いた機能未知タンパク質の機能予測の概念図．各タンパク質を丸で表現している．双方向の矢印は，各タンパク質が相互作用していることを示している．図は，タンパク質 A,B の機能が翻訳関係で，タンパク質 C が機能未知タンパク質であった場合の例

表3-5: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)

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3-13-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[53]．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

染色体分配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3-13-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．いわゆる実験データから精製されて得られたデータである．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[8]．

MreBとIscS間の相互作用は実験データから精製されて取得した相互作用である．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配にとって不可欠なタンパク質である[54]. 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, 原核生物において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配を調整する役割を担っていることが知られている[55][56][57]. 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[58], 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[60]. 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[61]. 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. 近年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配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配に関係する未知の機能を有していることが示唆された.

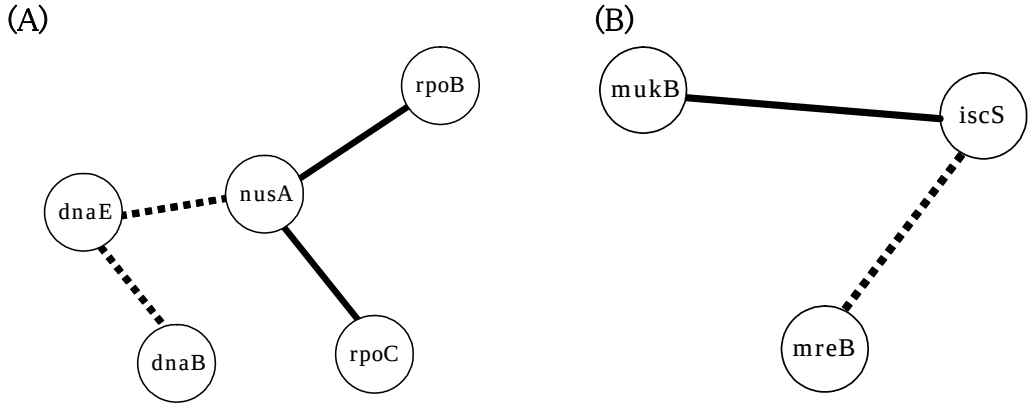


図 3-13 構築した PPI データから作られる相互作用ネットワークの例. 直線は Positive Data に含まれている相互作用データを意味し, 点線は精製された実験データに含まれているデータであることを意味する. (A)は NusA 周辺のタンパク質のネットワークであり, (B)は IscS 周辺のタンパク質である.

3.6 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[62].

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.

4.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や，有益な議論をしてくださった伊藤文氏，荒武氏，金井昭夫助教授，そして実験データの提供をしてくださった和田千恵子氏，MD Arifuzzaman 氏，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．

参考文献

[1] Link, AJ., Robison, K., Church, GM. 1997. Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. *Electrophoresis.* **18**:1259-313.

[2] Bowie, JU,. Luthy, R,. Eisenberg, D,. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991 **253**:164-70.

[3] Sprinzak, E., S. Sattath, and H. Margalit. 2003. How reliable are experimental protein-protein interaction data? *J Mol Biol* **327:** 919-923.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.* **24**:218-229.

[10] 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.* **17**:1030-1032.

[11] Kang, JS,. Kim, SH,. Hwang, MS,. Han, SJ,. Lee, YC,. Kim, YJ,. 2000 The structural and functional organization of the yeast mediator complex. J Biol Chem.**276**:42003-42010

[12] 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.

[13] 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.

[14] Smith, GP,. Schultz, DA,. Ladbury, JE,. 1993. A ribonuclease S-peptide antagonist discovered with a bacteriophage display library. *Gene*. **128**:37-42.

[15] Bader, G.D. and C.W. Hogue. 2002. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20:** 991-997.

[16] Rivera, MC,. Jain, R,. Moore, JE,. Lake, JA,. 1998 Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A*. **95**:6239-44.

[17] Fraser, HB,. Hirsh, AE,. Steinmetz, LM,. Scharfe, C,. Feldman, MW,. 2002 Evolutionary rate in the protein interaction network. *Science*. **296**:750-2.

[18] Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18:** 1257-1261.

[19] Vazquez, A., A. Flammini, A. Maritan, and A. Vespignani. 2003. Global protein function prediction from

protein-protein interaction networks. *Nat Biotechnol* **21:** 697-700.
[20] 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.
[21] Segal, E., H. Wang, and D. Koller. 2003. Discovering molecular pathways from protein interaction and gene expression data.*Bioinformatics* **19 Suppl 1:** i264-271.
[22] 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.
[23] 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.
[24] Liu, Y. and H. Zhao. 2004. A computational approach for ordering transduction pathway components from genomics and proteomics Data. *BMC Bioinformatics* **5:** 158.
[25] Steffen, M., A. Petti, J. Aach, P. D'Haeseleer, and G. Church. 2002. Automated modelling of signal transduction networks. *BMC Bioinformatics* **3:** 34.
[26] 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
[27] 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.
[28] 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.
[29] Bader, J.S. 2003. Greedily building protein networks with confidence. *Bioinformatics* **19:** 1869-1874.
[30] 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.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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
[44] Deng, M., S. Mehta, F. Sun, and T. Chen. 2002. Inferring domain-domain interactions from protein-protein interactions. *Genome Res* **12:** 1540-1548.
[45] Jeong, H., S.P. Mason, A.L. Barabasi, and Z.N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* **411:** 41-42.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] Brun, C., C. Herrmann, and A. Guenoche. 2004. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* **5:** 95.
[51] 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.
[52] 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.
[53] 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.
[54] 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.
[55] Carballido-Lopez, R. and J. Errington. 2003. A dynamic bacterial cytoskeleton. *Trends Cell Biol* **13:** 577-583.
[56] 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.
[57] 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.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] Mrowka, R., A. Patzak, and H. Herzel. 2001. Is there a bias in proteome research? *Genome Res* **11:** 1971-1973.

Web site references

http://www.ebi.ac.uk; SWISS-PROT
http://pir.georgetown.edu; PIR
http://ncbi.nlm.nih.gov; GenPept
http://www.prf.or.jp; PRF/SEQDB
http://www.rcsb.org/pdb/; PDB, RCSB Protein Data Bank
http://www.bmrb.wisc.edu/; Bio Mag Res Bank
http://www.expasy.ch/prosite/; PROSITE
http://blocks.fhcrc.org/; BLOCKS
http://pfam.wustl.edu/; The Pfam database of protein families and HMMs
http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/; PRINTS
http://interdom.lit.org.sg/; InterDom
http://mips.gsf.de; MIPS
http://dip.doe-mbi.ucla.edu/; DIP, Database of Interacting Proteins
http://www.blueprint.org/bind/bind.php; BIND
http://capri.ebi.ac.uk/ CAPRI Critical Assessment of PRediction of Interactions
http://ecoli.aist-nara.ac.jp;GenoBase
http://www.grs.nig.ac.jp/ecoli/pec/index.jsp; PEC database
http://ecocyc.org/; EcoCyc, Encyclopedia of *Escherichia coli* K12 Genes and Metabolism

参考書籍:

「タンパク質機能解析のためのバイオインフォマティクス」　藤博幸　講談社

「プロテオミクスの基礎」　著者:綱澤進, 平野久　講談社

「科学と生物　実験ライン　タンパク質　その本質と研究法」　著者:井本泰治　廣川書店

付録：学会発表(2003 年 4 月～2005 年 1 月 11 日現在)

★：学会口頭発表 ◎：学会ポスター発表
○：学会内出版

1. "Development of method to predict protein function of *Escherichia coli* using protein complex data."
Seira Nakamura, Rintaro Saito, Takeshi Ara, Aya Itoh, MD Arifuzzaman, Maki Maeda, Taku Oshima, Shigehiko Kanaya, Chieko Wada, Hirotada Mori, Masaru Tomita
◎First International E.coli Alliance Conference on Systems Biology of E.coli, 2003
○First International E.coli Alliance Conference on Systems Biology of E.coli, P80, 2003

2."Genome-wide analysis of protein-protein interactions in *Escherichia coli*"
Md. Arifuzzaman, Maki Maeda, Takita Chiharu, Taku Oshima, Aya Itoh, Shigehiko Kanaya, Altaf-Ul-Amin, Hikaru Kimura, Seira Nakamura, Rintato Saito, Masaru Tomita, Chieko wada, Hirotada Mori
★First International E.coli Alliance Conference on Systems Biology of E.coli, 2003
○First International E.coli Alliance Conference on Systems Biology of E.coli, P57, 2003

3."Prediction of Interacting Motif from Protein-Protein Interaction Data of *Escherichia coli*.“Hikaru Kimura, Seira Nakamura, Rintaro Saito, Takeshi Ara, Aya Itoh, Md. Arifuzzaman, Maki Maeda, Taku Oshima, Shigehiko Kanaya, Chieko Wada, Hirotada Mori, Masaru Tomita
◎First International E.coli Alliance Conference on Systems Biology of E.coli, 2003
○First International E.coli Alliance Conference on Systems Biology of E.coli, P81, 2003

4.「大腸菌タンパク質間相互作用データの精製とタンパク質の機能予測」
中村征良 斎藤輪太郎 荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
◎情報計算化学生物学会 Chem-Bio Informatics Society 2003 年大会
○New Frontiers for Chem-Bio Informatics, P238, 2003

5."Prediction of Motif-Motif Interaction and Protein-Protein Interaction in *Escherichia coli*.“Hikaru Kimura, Seira Nakamura, Rintaro Saito, Takeshi Ara, Aya Itoh, Md. Arifuzzaman, Maki Maeda, Taku Oshima, Chieko Wada, Hirotada Mori, Masaru Tomita
◎情報計算化学生物学会 Chem-Bio Informatics Society 2003 年大会
○New Frontiers for Chem-Bio Informatics, P239, 2003

6."Development of Method to Predict Protein Function of *Escherichia coli* Using Protein Complex Data."
Seira Nakamura, Rintaro Saito, Takeshi Ara, Aya Itoh, Md. Arifuzzaman, Maki Maeda, Taku Oshima, Chieko Wada, Hirotada Mori, Masaru Tomita
◎International Workshop for *Escherichia coli* Towards New Biology in the 21st Century
○Systematic functional analysis for *Escherichia coli* genome - resorces, systems approach and towards modeling -, P.91, 2003

7."Prediction of Motif-Motif Interaction and Protein-Protein Interaction in *Escherichia coli*."
Hikaru Kimura, Seira Nakamura, Rintaro Saito, Takeshi Ara, Aya Itoh, Md. Arifuzzaman, Maki Maeda, Taku Oshima, Chieko Wada, Hirotada Mori, Masaru Tomita
◎International Workshop for *Escherichia coli* Towards New Biology in the 21st Century
○Systematic functional analysis for *Escherichia coli* genome - resorces, systems approach and towards modeling -, P.96, 2003

8."Refinement and Mining of Protein-Protein Interaction Network in *Escherichia coli*."
Rintaro Saito, Seira Nakamura, Hikaru Kimura, Kohei Tsuzuki, Tomoko Takeda, Daisuke Kyuma, Yusuke Kobayashi, Shigeo Fujimori, Takeshi Ara, Aya Itoh, Md. Arifuzzaman, Maki Maeda, Taku Oshima, Chieko Wada, Hirotada Mori, Masaru Tomita
★International Workshop for *Escherichia coli* Towards New Biology in the 21st Century
○Systematic functional analysis for *Escherichia coli* genome - resorces, systems approach and towards modeling -, P.64, 2003

9.「大腸菌タンパク質間相互作用データの精製とタンパク質の機能予測」
中村征良　斎藤輪太郎　荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
◎第 26 回日本分子生物学会年会 2003
○第 26 回日本分子生物学会年会プログラム・講演要旨集 P.685

10.「大腸菌におけるモチーフ間相互作用とタンパク質間相互作用の予測」
木村曜 中村征良　斎藤輪太郎　荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
◎第 26 回日本分子生物学会年会 2003
○第 26 回日本分子生物学会年会プログラム・講演要旨集 P.686

11.「大腸菌タンパク質間相互作用ネットワークのマイニング」
斎藤輪太郎　中村征良 続木恒平 木村曜 荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
★微生物ゲノム研究のフロンティア, 2004
○Frontier of Microbial Genome Research, P13, 2004

12.「ゲノムワイドデータの統合による信頼性の高い PPI ネットワークの構築」
中村征良 木村曜 斎藤輪太郎 荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
◎微生物ゲノム研究のフロンティア, 2004
○Frontier of Microbial Genome Research, P54, 2004

13.「大腸菌タンパク質間相互作用ネットワークのマイニング」
続木恒平 中村征良 斎藤輪太郎 木村曜 武田朋子 久間大輔 小林雄輔 藤森茂雄 荒武 伊藤文 Md.Arifuzzaman 前田真希 大島拓 和田千恵子 森浩禎 冨田勝
◎微生物ゲノム研究のフロンティア, 2004
○Frontier of Microbial Genome Research, P57, 2004

14."Construction of Reliable Protein-Protein Interaction Network in Escherichia coli"
Rintaro Saito, Seira Nakamura, Hikaru Kimura, Kohei Tsuzuki, Takeshi Ara, Aya Ito, Md. Arifuzzaman, Maki Maeda, Masanari Kitagawa, Aki Hirai, Chieko Wada, Taku Oshima, Hirotada Mori, Masaru Tomita
★2nd International E.Coli Alliance Conference on Systems Biology Information
○2nd International E.Coli Alliance Conference on Systems Biology - Project Gemini

15.「再現性の高い大腸菌タンパク質間相互作用ネットワークの構築と機能未知タンパク質の機能予測」
斎藤輪太郎 中村征良 荒武 伊藤文 Md.Arifuzzaman 前田真希 北川正成 平井晶 和田千恵子 大島拓 森浩禎 冨田勝
◎第 27 回日本分子生物学会年会 2004
○第 27 回日本分子生物学会年会プログラム・講演要旨集 P.789

16.「ゲノムワイドデータの統合による信頼性の高いＰＰＩネットワークの構築」
中村征良 木村曜 続木恒平 斎藤輪太郎 荒武 伊藤文 Md.Arifuzzaman 前田真希 北川正成 平井晶 大島拓 和田千恵子 森浩禎 冨田勝◎第 27 回日本分子生物学会年会 2004
○第 27 回日本分子生物学会年会プログラム・講演要旨集 P.789

17.「選択的スプライシングによるタンパク質間相互作用制御のコンピュータ解析」
加来聡子 中村征良 斎藤輪太郎 冨田勝
◎第 27 回日本分子生物学会年会 2004
○第 27 回日本分子生物学会年会プログラム・講演要旨 P.789

Chapter 1. Introduction

1.1 Protein-Protein Interaction

Various phenomena *in vivo* are caused by nucleic acids, proteins, metabolites and the other molecules. These molecules interact with each other to keep life of an organism. For example, enzymes, which are proteins catalyzing the reactions among the metabolites, form metabolic pathway. Each of them is translated from RNA by ribosome and RNA is transcribed by RNA polymerase. In this way, proteins are closely linked to every essential intracellular activity.

Although proteins used to be believed that they are isolated entity and acting independently of surrounding proteins, today we know that most proteins *in vivo* are interacting with other ones and form complexes. This phenomenon, i.e., interaction among proteins, is called "PPI: Protein-Protein Interaction". PPI studies have potentials to elucidate various intracellular activities and they can also be applied to industries such as antibody medicines. Therefore the researches based on interactions are focus of constant attention.

The function of the complex is determined by its structure. Today, they can be described by 3-dimentional graph (**Figure1-A,B**) and stored in PDB (http://www.rcsb.org/pdb/). However, these structure data are of complex after crystallization, which do not always reflect the actual structure. And the number of solved 3 –dimentional structures are limited compared to number of protein sequences determined due to time and cost to solve such structures.

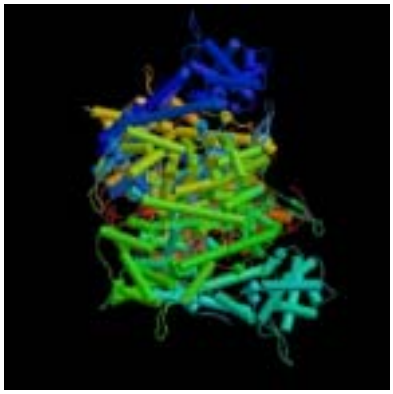


Figure 1-A: Structure of a T7 RNA polymerase elongation complex at 2.9A resolution

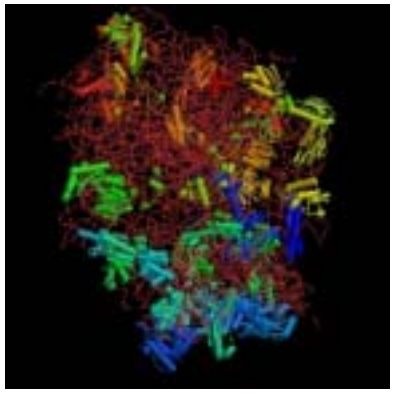


Figure 1-B: Structure of the *E. Coli* ribosomal termination complex with release factor 2

The most popular technique for high throughput screening of PPIs is Y2H (Yeast Two Hybrid Assay), which can identify many novel protein-protein interactions. Although the method can be used to find possible interacting partners there are many important interactions which cannot be detected using this method. In addition to this restriction, the low reproducibility and the low accuracy of its products are one of the most critical issues. Therefore in this thesis, we seek effective ways to construct genome-wide PPI network.

1.2 Construction of reliable Protein-Protein Interaction Network

Genome-wide protein-protein interaction (PPI) network contains functional complexes related to various reactions such as transcription, translation, metabolisms and signal transductions. Therefore experimental screenings of PPIs at the genome-wide level is important, particularly for well studies organisms such as *Saccharomyces cerevisiae* and *Escherichia coli*.

In yeast, PPIs were screened at genome-wide level using various methods including yeast two-hybrid assay (Gavin et al. 2002; Ho et al. 2002; Ito et al. 2001; Uetz et al. 2000). Thousands of PPIs are now available in yeast (Sprinzak et al. 2003). The usefulness of these data was shown by many previous researches. For example, Schwikowski et al. (Schwikowski et al. 2000) and Vazquez et al. (Vazquez et al. 2003) as well as other researchers have shown that functions of uncharacterized proteins can be predicted using PPI network. Discovery of novel biological pathways from the PPI network is one of the most important issues and has been conducted by many researchers (Bar-Joseph et al. 2003; Ideker et al. 2002; Liu and Zhao 2004; Segal et al. 2003; Steffen et al. 2002). For example, Ideker et al. identified PPIs which respond to perturbations in yeast galactose utilization pathway by combining PPI data, expression data and other proteomic data with biological experiments (Ideker et al. 2001). Also building principles of PPI network have been discovered which states that there are some “hub” proteins in the PPI network, which have many interacting partners and usually essential for cell to survive (Han et al. 2004; Jeong et al. 2001).

One of the greatest problem in dealing with these PPI data is that the experimental PPI data contain large amount of false-positive PPIs which may lead one to incorrect interpretations of PPI network (von Mering et al. 2002). Therefore various methods have been developed to eliminate these false positive PPIs (Bader 2003; Deane et al. 2002; Goldberg and Roth 2003;

Jansen et al. 2003; Saito et al. 2002; Saito et al. 2003b).

Compared to the case of yeast, the number of available PPIs in *Escherichia coli* is limited. For example, in one of the most known database DIP (Salwinski et al. 2004; Xenarios et al. 2002), the number of deposited *E. coli* PPIs (heterodimers) is only 265 (as of Aug. 30, 2004). Although known protein complexes are stored in EcoCyc (Karp et al. 2004; Karp et al. 2002) it is mainly focused on enzymes. Still it is very important to construct PPI network at the genome-wide level in *E. coli*, because it contains functional units or regulatory pathways which are specific to eubacteria.

In this work, we constructed genome-wide PPI network in *E. coli* using Bayesian approach (Jansen et al. 2003). We used 9,233 PPIs obtained from our Pull-Down assay (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) (Zhu et al. 2001) and collected 3,285 PPIs reported in the literature. Characteristics of PPIs in literature data were analyzed and Bayesian approach was used to assess each interaction obtained from the experiments according to these characteristics. Subsequently we removed suspicious PPIs and constructed reliable PPI network by integrating the remaining experimental PPIs with those reported in the literature, which finally contained 3,667 interactions connecting 1,277 proteins. Constructed network enabled us to predict functions of uncharacterized proteins and gave new insights into biological reactions which occur in *E. coli* cell, showing usefulness of our PPI network.

Chapter 2. Results

2.1 Collection of PPIs

We first collected 18,870 interacting partners screened by our genome-wide Pull-down assay using 2,669 bait proteins (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). Interacting partners having high affinities to the columns were discarded from them. Some of the detected interacting partners were ambiguous – the detected weights of proteins corresponded to multiple proteins. These interacting partners were also discarded. We assumed that proteins detected by a bait interacted with each other (The “matrix approach” as discussed in (Bader and Hogue 2002)). As a result, we

obtained 9,233 PPIs from our experiment.

In addition, we collected PPIs reported in the literatures; we extracted abstracts related to PPI in PubMed using keyword search. We manually curated all the abstracts and extracted protein names which interact. Then we integrated these documented PPIs with those in DIP (http://dip.doe-mbi.ucla.edu/) and EcoCyc (http://ecocyc.org/). Finally we obtained 3,285 documented PPIs (referred to as positive PPIs hereafter).

For training of Bayesian network which we use to model PPIs, set of protein pairs which do not interact (referred to as negative PPIs hereafter) was required. We followed similar procedure used by Jansen et al. to prepare negative PPIs (Jansen et al. 2003); Protein pairs which interact are considered to share same functions and co-localized in a cell (Sprinzak et al. 2003). We therefore defined negative PPIs as protein pairs which do not share same function and are not co-localized. Protein functions were assigned using GenoBase (Ara et al. in preparation). Protein localizations were predicted by PSORT (Nakai and Horton 1999). We looked through all possible protein pairs in *E. coli* and found that 2,277,085 protein pairs satisfy the above criteria. We defined them as negative PPIs. The datasets we collected are summarized in **Table 1**.

Table 1: PPI datasets collected or constructed in this study.

PPI Dataset	Description	The number of PPIs
Experimental PPIs	PPIs obtained from our Pull-down experiment.	9,233
Positive PPIs	PPIs reported in the literatures and/or stored in public databases.	3,285
Negative PPIs	Protein pairs which does not interact according to computational prediction.	2,277,085
Refined PPIs	Experimental PPIs which are suggested to be reliable using our method.	427
Predicted PPIs	Protein pairs in *E. coli* proteome which are predicted to interact using to our method.	1,407
Finally constructed PPIs	Integration of positive PPIs with refined PPIs.	3,667

2.2 Construction of Reliable PPI Network

PPIs obtained from genome-wide screening assay have been known for their low accuracy (Mrowka et al. 2001). To eliminate false positive PPIs, we assessed each interaction according to 7 indices, i.e., gene expressional correlations (Grigoriev 2001), correlations of phylogenetic profiles (Pellegrini et al. 1999), Interaction Generality (Saito et al. 2002; Saito et al. 2003a), motif-motif interactions (Deng et al. 2002), gene essentialities (Jeong et al. 2001), interologs (Walhout et al. 2000), and participation of genes in the same transcriptional unit (Dandekar et al. 1998).

First, we analyzed whether these 7 indices can be used as evidences for protein-protein interactions. Each probability $P(positive \mid Indice)$ was calculated from positive data and negative data. We found that pairs of proteins having certain value of 7 indices are likely to interact suggesting that all the indices can be used to assess PPIs. The description is given below.

2.2.1 Gene Expression Pattern

It is shown in yeast that protein pairs encoded by co-expressed genes interact with each other more frequently than with randomly selected protein pairs (Ge et al. 2001; Grigoriev 2001). We checked whether this phenomenon can be observed in *E. coli*. We downloaded gene expression data from KEGG and GenoBase (Oshima et al. 2002). Using correlation coefficient as a measure of expressional similarity, we found that protein pairs having similar expression profiles in *E. coli* are also likely to interact **(Figure 2A)**.

2.2.2 Phylogenetic Profiling

If two proteins are functionally linked, they are likely to show similar phylogenetic profiles, i.e., they are likely to be in the same subset of completely sequenced genomes (Pellegrini et al. 1999). To identify pairs of proteins showing similar phylogenetic profiles, each protein sequence of *E. coli* was subjected to BLAST search against other completely sequenced genomes (6 archaeal genomes and 36 eurobacterial genomes: **Table 2**). Then pairs of proteins showing similar E-values among other species were selected as phylogenetically linked. For example, consider the case where we deal with three *E. coli* proteins, A, B and C. Each of three proteins is subjected to BLAST search against genomes of other species X, Y and Z. E-values of protein A to species X, Y and Z is 1.0×10^{-4}, 2.0, 1.0×10^{-6}, and those of protein B is 1.0×10^{-10}, 1.0×10^{-20}, 2.0, and those of protein C is 2.0 x

10^{-4}, 4.0, 2.0 x 10^{-6}. In this case, protein A and C are assumed to be phylogenetically linked, because their E-values to other species correlate. Correlation coefficient was used to assess similarity of E-values. We found that as shown in **Figure 2B**, pairs of proteins having similar phylogenetic profiles are likely to interact.

Table 2: The list of genomes that we used in phylogenetic profiling. "A" and "E" stands for "Archaea" and "Eurobacteria".

Aeropyrum_pernix	A
Agrobacterium_tumefaciens_C58_Cereon	E
Aquifex_aeolicus	E
Archaeoglobus_fulgidus	A
Bacillus_subtilis	E
Borrelia_burgdorferi	E
Buchnera_sp	E
Campylobacter_jejuni	E
Caulobacter_crescentus	E
Chlamydia_muridarum	E
Chlamydophila_pneumoniae_AR39	E
Clostridium_acetobutylicum	E
Deinococcus_radiodurans	E
Haemophilus_influenzae	E
Halobacterium_sp	E
Helicobacter_pylori_J99	E
Lactococcus_lactis	E
Listeria_innocua	E
Methanobacterium_thermoautotrophicum	E
Methanococcus_jannaschii	A
Mycobacterium_leprae	E
Mycoplasma_genitalium	E
Neisseria_meningitidis_MC58	E
Nostoc_sp	E
Pasteurella_multocida	E
Pseudomonas_aeruginosa	E
Pyrococcus_furiosus	A
Rickettsia_conorii	E
Salmonella_typhi	E
Sinorhizobium_meliloti	E
Staphylococcus_aureus_Mu50	E
Streptococcus_pneumoniae_TIGR4	E
Sulfolobus_solfataricus	A
Synechocystis_PCC6803	E
Thermoplasma_acidophilum	A
Thermotoga_maritima	E
Treponema_pallidum	E
Ureaplasma_urealyticum	E
Vibrio_cholerae	E
Xylella_fastidiosa	E
Yersinia_pestis_CO92	E

2.2.3 Interaction Generality

IG (Interaction Generality) is designed to measure the experimental reproducibility of each interaction according to the topology of PPI network (Saito et al. 2002; Saito et al. 2003a). We used the modified version of IG to assess each interaction obtained from our experiment. The IG value is calculated by the following procedure. First our experimental data were converted into binary PPI data by "spoke approach" which assumes that all detected proteins interact with bait protein (Bader and Hogue 2002). Let *r* be the number of times protein X and Y are detected using the common baits, and a1 be the number of proteins which interact with both X and Y. Then IG for interacting protein X and Y is max(*r*, a1). As shown in **Table 3**, protein pairs having higher IG values are likely to interact.

2.2.4 Motif-Motif Interaction

Motif-Motif Interaction (MMI) is the interaction of motifs between interacting proteins. To find potential MMIs from experimental PPI data, we first searched for known protein motifs in each protein sequence using Pfam database (http://pfam.wustl.edu/). Then all the pairs of motifs found in one protein and in its interacting partners were statistically assessed to identify those which appear frequently compared to the expected frequency based on number of times that each protein appears in the protein interaction network. In particular, we calculated Observed / Expected (O/E) ratio for each MMI. MMI score for each interaction was defined as O/E ratio of MMI corresponding to pair of motifs interacting proteins have. If no corresponding pair is found, MMI score is defined 0. We found that protein pairs having high MMI scores are likely to interact **(Table 3)**.

2.2.5 Gene Essentiality

It is shown that essential proteins interact more frequently than expected by chance (Saito et al. 2003b). We classified each protein into essential and non-essential protein according to PEC database (http://www.grs.nig.ac.jp/ecoli/pec/index.jsp). Experimental PPI data contained 341 essential proteins and 1,922 non-essential proteins. We classified interactions into 1: Non-essential - Non-essential interactions, 2: Essential - Non-essential interactions, and 3: Essential - Essential interactions. For those where the phenotype of one of interacting proteins is

unknown, they were classified to 0: unknown. We found that pairs of essential proteins are likely to interact **(Table 3)**.

2.2.6 Interolog

Interolog is the interaction among the orthologous proteins (Walhout et al. 2000). We collected PPIs of *H. pylori* obtained from genome-wide screening assay (Rain et al. 2001) and predicted their interologs in *E. coli*. In particular, we subjected each pair of interacting proteins in *H. pylori* to ssearch (threshold was set to E-value < e^{-10}) against protein sequences of *E. coli* and assumed pairs of *E. coli* proteins as potential interologs of *H. pylori*. We found that potential interologs are likely to interact **(Table 3)**.

2.2.7 The Transcription Units (Operon)

Bacterial mRNAs often encode multiple genes. Proteins encoded in the same transcriptional units tend to be functionally related (Dandekar et al. 1998). To assess interacting potentials of proteins in the same transcriptional unit, we obtained the transcription units from EcoCyc and correlated them with positive and negative PPIs. We found that protein pairs encoded in the same transcription unit are often reported in the literatures as interacting pairs **(Table 3)**. Therefore we suggest that experimental PPIs are likely to be true positives if interacting proteins are encoded in the same operon.

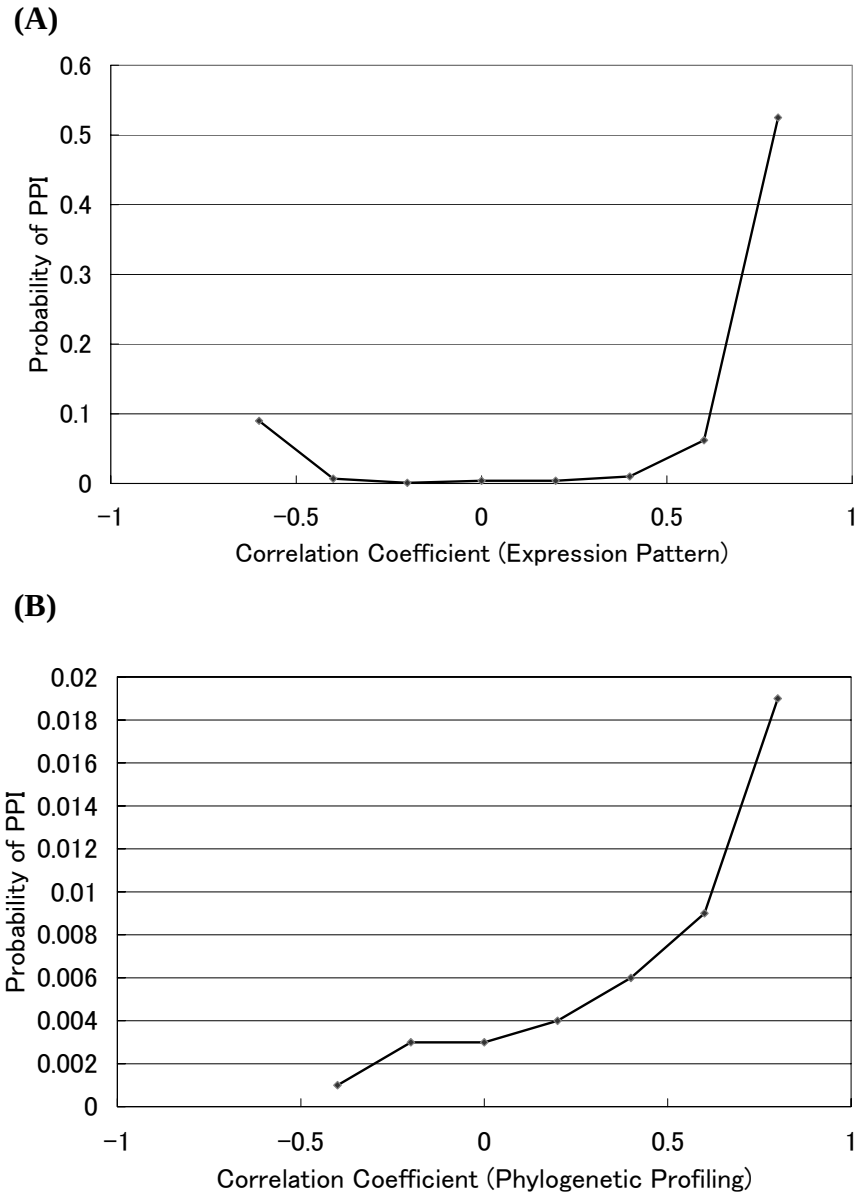


Figure.2: Probability that interaction occur between two proteins having certain level of (A) expressional correlation and (B) phylogenetic correlation

Table 3: Probability that interaction occur between protein pairs having specified value for given index (Interaction Generality, Essentiality Pattern, MMI Score, Interolog and Operon) as described in text. The procedure to calculate probability is based on Bayesian approach and described in methods.

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

2.2.8 Integration of 7 Evidences

As shown, we confirmed that all 7 indices can be used as evidence sources for PPIs. We integrated these 7 indices using Bayesian approach to calculate interacting potentials between given protein pairs. First we investigated the correlations among these indices. We scored the interacting protein pairs using 7 indices and calculated the correlations coefficients between them, and tested the statistical significances by t-test. There were significant correlations among 7 indices. Therefore we integrated the 7 indices by Fully connected Bayes and calculated probability of interactions between given protein pairs (See methods). To calculate them using Bayesian approach, we needed to estimate number of PPIs exist in *E. coli* cell. The number of PPIs in *S. cerevisiae* is estimated to be between 16,000 and 26,000 (Grigoriev 2003). The number of coding regions in *E. coli* is ~4,000, which is 2/3 of Yeast. If the number of interaction partner per protein in Yeast (5 ~ 8) and *E. coli* is similar, the number of real *E. coli* PPIs will be estimated to be between 10,000 (4,000 x 5 / 2) and 16,000 (4,000 x 8 / 2). If the density of PPI (number of interactions over number of all the pairs of proteins in the set of proteins) is similar in positive PPIs and the whole proteome of *E. coli*, the number is estimated to be ~40,000 (The average degree of the 3,285 positive PPIs which consist of 1123 proteins, is $3285/{}_{1123}C_2$, which is ~0.005. As *E. coli* have ~4,000 ORFs, the estimated total number of PPIs is ${}_{4000}C_2$ x 0.005, which is ~40,000). We tried the above 2 estimates and investigated the performance of our method.

2.3 Validation of the Method

Our method calculates probability of interaction between given protein pair according to the 7 indices using Bayesian approach. To validate our method, we calculated the accuracy of our method using cross validation. We chose one positive or negative PPI from collection of positive and negative dataset. Then using the rest of the data, we trained the probability function P(positive | ES) and P(negative | ES) where ES (Evidence source(s)) indicates 7 indices. Then we tested whether the function correctly predicts the chosen sample as interacting proteins or not. The prediction is counted as true positive if the function correctly predicts interacting proteins as interacting. The prediction is counted as false positive if the function incorrectly predicts non-interacting proteins as interacting. The accuracy of the prediction is measured as *True Positives* / (*True Positives* + *False Positives*). We found the positive correlation between the probability P(positive | ES) and accuracy. **(Figure 3)** Therefore we suggest that our method is valid for estimating probability of interactions.

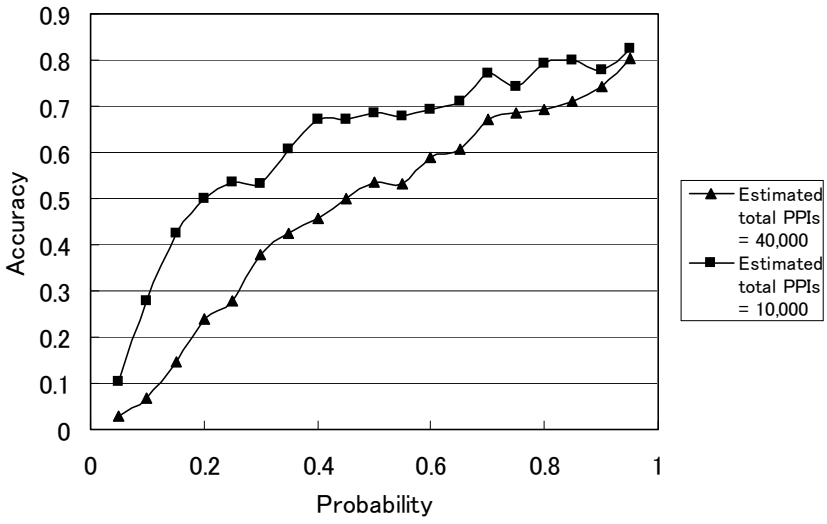


Figure 3 Correlation between probability of interaction calculated according to Bayesian approach and accuracy calculated as rate of true positives among protein pairs predicted as positive. We estimated number of PPIs in *E. coli* as 10,000 and 40,000 to calculate probability.

We can obtain reliable PPIs by selecting experimental PPIs having interaction probability higher than the threshold. There is a trade-off between the threshold and the number of reliable PPIs obtained after refinement. If we set threshold high, number of reliable interactions we obtain will be few. **(Figure 4)**

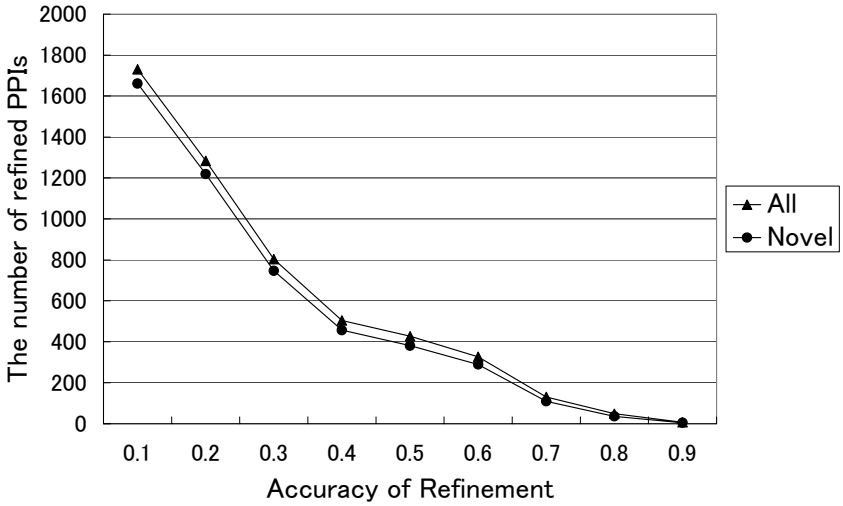


Figure 4 The accuracy of refinement and the number of PPIs left in the refined PPI network having corresponding accuracy. "All" indicates all the interactions left in the refined PPI network and "Novel" indicates those not previously reported.

We set the probability threshold 0.5 (accuracy would also be ~0.5) and selected interactions having higher probability than the threshold. There are 9,233 PPIs in our experimental data, which contains 94 positive PPIs (1% of experimental data). After refinement using the threshold, we obtained 427 reliable PPIs containing 46 positive PPIs (10.8% of reliable PPIs).

We also predicted 1,407 PPIs from all the pairs of proteins in *E. coli* (9,290,205 pairs, 3.9×10^{-4} of which are positive PPIs). Predicted PPIs contained 330 positive PPIs (23% of predicted PPIs) and 1,077 novel PPIs. By integrating PPIs reported in the literature with our refined experimental PPIs, we finally obtained 3,667 reliable PPIs. This dataset contains 1,277 proteins (>1/4 of the number of *E. coli* coding regions), and the average of the number of interaction partners per protein is about 6.

Figure 5 shows relationship between number of interacting partners and number of proteins having corresponding number of interacting partners. The distribution follows a power law which is observed in many interaction networks (Jeong et al. 2001). Furthermore there were positive correlation between number of interacting partners and rate of essential proteins having corresponding number of interacting partners; For example rate of essential proteins having more than 5 interacting partners were 20.6%, whereas rate of those having 5 interacting partners or less were 11.6% ($P < 0.002$). These observations are similar to that observed in *S. cerevisiae* (Jeong et al. 2001). Therefore we suggest that global structure of our PPI network reflects the actual PPI network in cells.

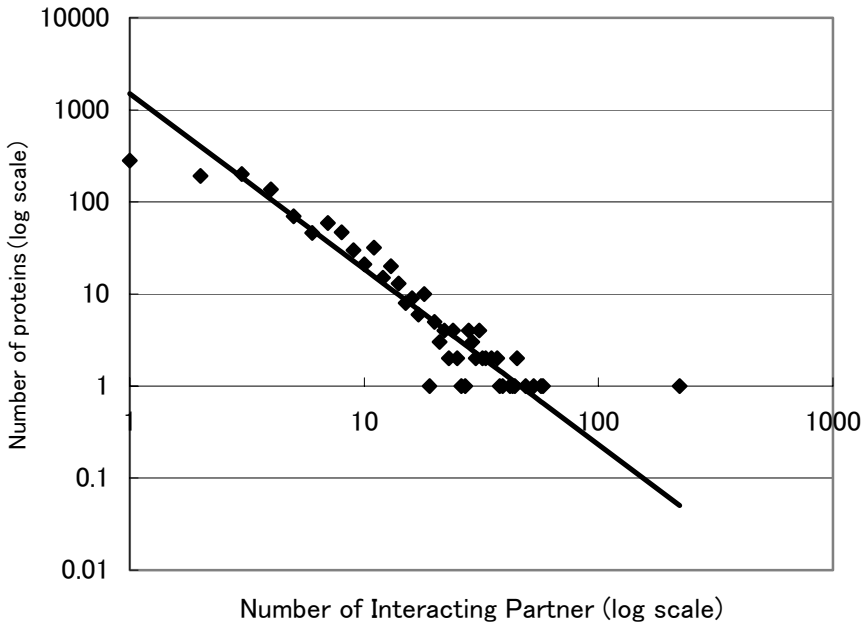


Figure 5 Relationship between number of interacting partners and number of proteins having corresponding number of interacting partners

Chapter 3. Discussion

We have developed the method to refine PPI data obtained from our experimental method and constructed reliable PPI network at the genome-wide level in *E. coli*. Our method was justified using cross validation as discussed. Further validations were performed by curating each interaction in the constructed PPI network. We found that some of the interactions can be supported by literatures which are not found in sources of our positive PPIs. For example, FlgN, known as a chaperon for flagellum interacted with FlgL, suggesting the possibility that FlgN transports FlgL. We found a literature relevant to this interaction (Fraser et al. 1999).

E. coli genome data contain many uncharacterized and functionally unannotated ORFs; the functions of thousands of predicted proteins still remain unknown. However it is possible to predict functions of uncharacterized proteins according to our PPI network (Brun et al. 2004; Deng et al. 2003; Hishigaki et al. 2001; Marcotte et al. 1999). Interacting proteins tend to share common functional roles. Therefore, functions of uncharacterized proteins can be predicted according to the functions of the interacting partners. The prediction in this way would be successful using our reliable PPI network (Saito et al. 2002). The constructed network contained 98 uncharacterized proteins. Fifty-four of them had interacting partner(s) having only one function and therefore one putative function can be assigned for each protein. Nine of them had more than one interacting partners having common function **(Table 4)** and accuracy of functional predictions for these proteins are supposed to be much higher than the other ones (Saito et al. 2003b).

Table 4: Prediction of protein function of uncharacterized proteins from PPI network. Protein functions are obtained from GenoBase. Numbers in parentheses denote the numbers of interaction partners of known function.

Gene	Functional prediction
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)

There are many multifunctional proteins in all cellular organisms. Some of our PPIs connect two or more protein complex having different functional categories. We looked for such type of interactions using functional classification information and considered the novel functional roles of each interaction. Interaction between protein complex of transcription and replication mediated by NusA protein was observed **(Figure 6A)**. NusA binds directly to RNA polymerase and controls the anti-termination of transcription by forming the complex with other factors such as NusB, NusG, and Rho (Muhlberger et al. 2003). In fact, there were interactions between NusA and subunits of transcription complex, RpoB-RpoC, in our positive PPIs. On the other hand, we found novel interaction between NusA and DnaE protein (DNA polymerase III), which interacted with DnaB DNA helicase. DnaB and DnaE are components of DNA replication complex. Therefore we suggest that NusA can also regulate anti-termination of DNA replication as well.

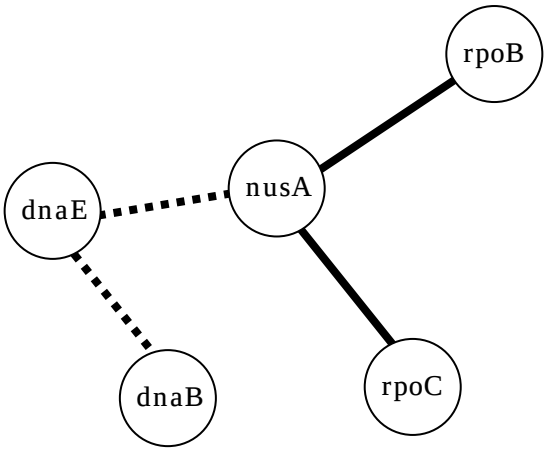
Two protein complexes of chromosome segregation mechanism and bacterial cytoskelton may be connected by a protein complex of sulfur metabolism **(Figure 6B)**. We found interaction between MukB and IscS in the refined PPI network. Although it was not included in the positive PPIs, we found a literature reporting this interaction (Gully et al. 2003). However interaction between MreB and IscS is not previously reported. MukB is one of the SMC (structural maintenance of chromosomes) protein and essential for chromosome segregation coordinated with cell division (Niki et al. 1991). MreB is thought as component of cytoskelton network which is a structural and functional homologue of actin in bacteria. MreB forms helical filaments that extended along the long axis of the cell. It also has a regulatory role of cell shape and chromosome segregation (Carballido-Lopez and Errington 2003; Kruse et al. 2003; Norris et al. 2004). IscS, initially identified as the third cysteine desulfurase, encodes tRNA thiouridine modification enzyme (Lauhon et al. 2004). IscS and IscU forms protein complex for the biosynthesis of iron-sulfur clusters (Kato et al. 2002). IscA may be also included in a component of iron-sulfur cluster assembly (Ding et al. 2004). Gene cluster of iscSUA is highly conserved in bacteria. Recently it was reported that IscA was localized in polar or near potential cell division site in the cell, independent of reported positioning factors FtsZ and MinCDE (Janakiraman and Goldberg 2004). However, relationship between IscA and

MreB protein was not mentioned in this report. Thus interactions between MukB, IscS and MreB may suggest new possible candidates of factors in chromosome segregation and suggests that IscS has the other unknown functions related to chromosome segregation in bacterial cell division.

Our method relies heavily on the tendencies of 7 indices in known PPIs (positive PPIs). However there may be interactions which do not follow tendencies in currently known interactions and our method may not detect such interactions. To detect interactions having novel tendencies using our method, we need appropriate training set; those which were confirmed by reliable assay systems having no bias to particular proteins.

In summary, we have constructed reliable PPI network at the genome-wide level in *E. coli*. The network suggested the novel roles for various proteins participating in the network, showing the usefulness of our network. We finally suggest that by integrating our reliable PPI network with other types of genome-wide data such as protein-DNA interaction data, more information on roles of each protein may be extracted which may lead us to the discovery of novel biological pathways.

(A)



(B)

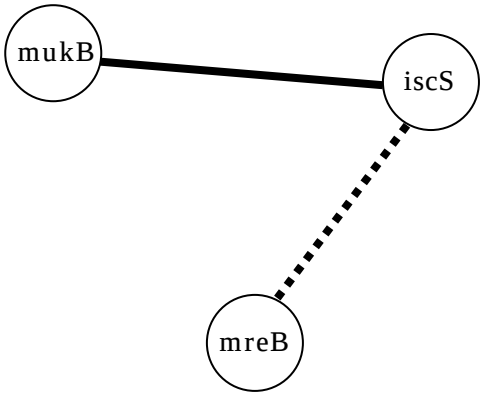


Figure 6: Examples of the constructed PPI network. Solid lines indicate interactions included in positive PPI data or reported in the literature, and dotted lines indicate refined interactions which are not previously reported. (A) Local PPI network around nusA. (B) Local PPI network around iscS.

Chapter. 4 Methods

Bayes' theorem is a simple mathematical formula used for calculating conditional probabilities. By using Bayesian approach, we can predict the probability of the object (it is PPI, in this paper) even when the part of information is missing. The matters which are related to the object are called evidence sources (ES). The probability of the object is calculated by integrating these ESs. These are 2 kinds of Bayesian approaches to integrate them. One is “Fully Connected Bayes”, which is used when there are correlations among ESs and the other is “Naïve Bayes” which is used when there are no correlations among ESs. As we observed correlation among each ES, we integrated ESs by Fully Connected Bayes.

To calculate the probability that given protein pair interacts using Bayesian approach, at first, we calculated the score of every ES by the pair of *E.coli* proteins. For example, the ES scores of “sucA” and “sucC” pair is “Phylogenetic Profiling: 0.2, Gene Expression Pattern: 0.6, IG: 2, MMI: 2, Gene Essentiality: 1, Operon: 1”.

We counted positive PPIs $\{positive \mid ES1, ES2, ES3,,,\}$, negative PPIs $\{negative \mid ES1, ES2, ES3,,,\}$, and all pairs of proteins within *E.coli* $\{ES1, ES2, ES3,,,\}$ by combination of these scores as the training set.

The probability that interaction occur between given protein pair having ES1, ES2, ES3, … is designated as $P'(positive \mid ES1,\ ES2,\ ES3,\ldots)$ and calculated using the following formula.

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

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

$$P(positive) = \frac{\text{The number of actual interactions in a cell}}{\text{The number of the all pairs of proteins}}$$

$$P(negative) = 1 - P(positive)$$

$P(positive)$ and $P(negative)$ are the probabilities that interaction occur and no interaction occur between given protein pair respectively.

$P(ES1, ES2, ES3,\ldots \mid positive)$ is the percentage of the number of the protein pairs that have scores of ES1, ES2,E S3,,, in positive data, which

is calculated as $\{positive \mid ES1, ES2, ES3, \ldots\}/\{positive\}$. In a same way, $P(ES1, ES2, ES3, \ldots \mid positive)$ is {negative | ES1,ES2,ES3,...}/{negative}. P'(positive | ES1, ES2, ES3,...), P'(negative | ES1, ES2, ES3,...) are biased probabilities from each dataset (positive dataset, negative dataset). Finally the probability $P(positive \mid ES1, ES2, ES3, \cdots)$, coupled with positive data and negative data is calculated according to the following formula.

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

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

We searched the protein pairs of P(positive | ES1,ES2,ES3,...)>=0.5 in refining of experimental PPIs and genome-wide predicting, because of its ratio values are more than 1. The ratio value is used for integrating ESs by "Naïve Bayes". If we could found novel ES for PPI which is independent of the other ESs, the integration of that is easily calculated as the following formula.

$$RatioValue(ES1, ES2, ES3, \cdots, NovelES) = RatioValue(ES1, ES2, ES3, \cdots) RatioValue(NovelES)$$

Chapter. 5 Acknowledgements

The following people kindly provided me experimental data of PPI, which is the basis of my work; Md.Arifuzzaman, Maki Maeda, Masanari Kitagawa, Aki Hirai, Taku Oshima and Chieko Wada. We had fruitful discussions with many people on the topics covered in this thesis. Especially, I thank Dr. Rintaro Saito, Dr. Aya Itoh, Dr. Takeshi Ara, Dr. Akio Kanai and Professor Hirotada Mori. I also thank faculties and staffs of Institute for Advanced Biosciences, Keio University. Finally, I am very grateful to Professor Masaru Tomita for offering me the opportunity and environment for my study.