ベイズ統計を用いた
大腸菌ＰＰＩの精製と予測

政策・メディア研究科　中村征良

Abstract

Escherichia Coli におけるタンパク質間相互作用 (以下 PPI:Protein-Protein Interaction) の既知情報は、約 3 万という予想される情報量に程遠く、DIP(Database of Interacting Proteins) におけるその情報は、バイナリーデータにしておよそ 500 程度である。PPI データが系統プロファイルや発現相関などと相関を有することがわかっていることから、それらの相関を用いて実験データの精製や、ゲノムワイドな PPI 予測を行うことを試みた。使用したデータは系統プロファイル, 発現相関,IG(Interaction Generality),MMI(Motif-Motif Interaction),Essentiality, そして実験結果 (Pull-Down 法) である。以上の手法を統合し、DIP と論文サーベイから得た Positive Data を用いてトレーニングし、大腸菌における全遺伝子の翻訳産物であるタンパク質同士の組み合わせを対象にしてベイズにより統合した。その結果、信頼性と共に新規 PPI の精製、ゲノムワイドな予測を行うことに成功した。

1　はじめに

PPI と相関を有すると考えられる各事象は、その相関を検証し、Pull-Down 法 [5] を用いた *Escherichia coli*(大腸菌) における PPI 実験結果データから擬陽性を除くといった精製過程に利用することができる。しかし、単に条件の整った和集合や積集合では、それらの精製結果の信頼性は同機能率（タンパク質ペアが同じ機能カテゴリに属している率）に依存せざるを得ない曖昧なものであり、定量的にその信頼性を示すことが出来ない。また、系統プロファイルや発現相関などが PPI と相関を有しているのではなく、直に同機能率と相関を有しているという考察も出来たため、同機能率に依存しない新たな PPI 予測手法を必要とした。

現在、*Saccharomyces cerevisiae* において、実験データやその他の独立した PPI 予測手法をベイズ統合した手法が提案されており [1]、その精度の高さは Y2H 法による実験結果を上回っていることで注目を集めている。そのことから、大腸菌における PPI 予測手法のベイズ統合の成績も期待することができ、実験結果や大腸菌内全タンパク質ペアをベイズで統合した手法の対象にすることで、さらに精密な実験データの精製、及びゲノムワイドな PPI 予測を行った。

2　統合する各手法の紹介

統合予定の PPI 予測手法は以下のものであり、PPI との相関が確認されているものを使用する。

2.1　系統プロファイリング

系統プロファイリング (Phylogenetic Profiling)[2] は、類似した進化経路を有するタンパク質ペアは相互作用している可能性が強いという考察のもと使用される手法である。大腸菌における約 4000 個の遺伝子の翻訳産物である全てのタンパク質が、複数の異種（原核生物に限る）においてどのような存在パターンを有しているのか、翻訳するアミノ酸配列の相同性から測る。この研究において、それぞれのタンパク質の情報の類似度は相関係数をスコアとして利用している。

2.2　発現相関

相互作用を有するタンパク質ペアは、類似した発現パターンを持つ傾向が高いとされている [3]。使用した発現データは奈良先端技術大学院大学森研究室より頂いた。大腸菌における全遺伝子の発現パターンを配列化し、類似度を相関係数でスコア化して利用した。

2.3　Interaction Generality

斎藤氏らにより提唱された指標 [4] で、構築された PPI ネットワークをもとに各エッジの信頼度を定量的に計測したものである。スコアは自然数で算出され、大きければ大きいほど信頼性が高い。今回使用したスコアは、実験結果（森研究室提供）を Matrix Approach でバイナリー化したものをもとにしている。

2.4　Motif-Motif Interaction

Motif-Motif Interaction(以下 MMI) とは、相互作用しているタンパク質ペアからモチーフペアの情報を抽出し、相互作用を決定付ける一般情報として得るものである。そのモチーフの組み合わせを新規 PPI 予測に利用することができ、O/E 値をスコアとしたものを使用した（木村曜氏提供)。

2.5　Essentiality

PPI と Essentiality 間には高い相関が見られる。Essential Protein のデータを PEC データベース (http:www.grs.nig.ac.jp/ecoli/pec/index.jsp) より取得したところ、Pull-Down 法による実験結果を Matrix Approach によってバイナリー化した PPI データには Essential Protein が 341 個、Nonessential Protein が 1922 個出現する。その中で Essential Protein 同士の相互作用は 4086 個 (O/E:11.6)、片方が Essential Protein である相互作用は 7578 個 (O/E:3.8)、Nonessential Protein 同士の相互作用は 3883 個 (O/E:0.3) となっており、相互作用するタンパク質ペアにおいて少なくとも片方が Essential Protein である可能性が高いことを意味している。

3　統合手法

3.1　ベイズとは

ベイズとは、ある命題から目標命題への推移確率を統計する手法であり、現在取得している不完全な情報のみで目的事象の発生確率を予測することができる。複数の命題の目標命題への推移確率をベイズで統合する際、各々の予測手法が独立であるならば Naive Bayes（図１）を用い、互いに相関を有していれば Fully Connected Bayes（図２）を用いる必要がある。Naive Bayes は、各命題から目標命題への推移確率をオッズ比に換算して積算統合するものであり、Fully Connected Bayes は各命題の組み合わせ毎に推移確率を計算する。

図 1: Fully Connected Bayes

図 2: Naive Bayes

3.2　各手法間の相関解析

ベイズ統合手法を確立するために各手法（命題から目標命題への推移）同士の相関解析を行った。大腸菌における全タンパク質ペアを対象に、各手法のスコアを算出してその相関係数を検定したものが以下である。（表１）

表 1：各 PPI 予測手法間の相関係数検定 (各手法間の相関係数を t 値に変換したもの)

系統プロファイル					
164.277	発現相関				
0.192	53.0185	IG			
0.033	51.155	7959.295	実験結果		
73.293	64.630	378.305	434.542	MMI	
45.860	219.780	83.422	83.552	147.713	Essentiality

有意水準 0.05 の場合 t 値の臨界値は 1.96 である（自由度 9152779）。これを基にした各手法間の相関ネットワークは以下のようになる（図３）。各手法は互いに独立ではなく、互いに相関を有していることから、Fully Connected Bayes を用いて PPI 予測手法の統合を行うことを決定した。

図 3: 各手法間の相関図

3.3 確率の計算

Fully Connected Bayes による統合は、各手法における予測結果スコア $(f_1, f_2, f_3 \cdots)$ の組み合わせ毎に発生確率 $P\,(positive\ data|f_1, f_2, f_3 \cdots)$ を算出する必要がある。しかし、大腸菌において存在する全ての Positive Data は未だ明確になっていないため、その確率を直接算出することが出来ない。そのため、以下の数式変換を用いて行う。

$$P'\,(positive\ data|f_1, f_2, f_3 \cdots) = \frac{P(positive\ \ data)\ \ P(f_1,f_2,f_3 \cdots|positive\ \ data)}{P(f_1,f_2,f_3 \cdots)}$$
$$P'\,(negative\ data|f_1, f_2, f_3 \cdots) = \frac{P(negative\ \ data)\ \ P(f_1,f_2,f_3 \cdots|negative\ \ data)}{P(f_1,f_2,f_3 \cdots)}$$

$$P\,(positive\ data) = \frac{30000(\text{予想されている } PPI \text{ の総数})}{9152781(\text{大腸菌内において考えられる全タンパク質ペア})}$$
$$P\,(negative\ data) = 1 - P\,(positive\ data)$$

$P\,(positive\ data)$、及び $P\,(negative\ data)$ は、PPI の総数を過程して利用する。*Saccharomyces cerevisiae* における PPI 総数予測の幅は 10,000〜40,000 個とされている [9]。仮にタンパク質一つあたりの相互作用相手数を固定した場合、遺伝子数がおよそ 3 分の 2 である大腸菌においては、PPI 総数は 7,000〜27,000 個と予想されるが、大腸菌においては各タンパク質が巨大な複合体を形成しているため、PPI 総数を 30,000 個と仮定した。positive data と negative data の両方を加味した最終的な $P\,(positive\ data|f_1, f_2, f_3 \cdots)$ は、$P'\,(positive\ data|f_1, f_2, f_3 \cdots)$ と $P'\,(negative\ data|f_1, f_2, f_3 \cdots)$ から以下の様に算出する。

$$Ratio\ Value = \frac{P'(positive\ \ data|f_1,f_2,f_3 \cdots)}{P'(negative\ \ data|f_1,f_2,f_3 \cdots)}$$
$$P\,(positive\ data|f_1, f_2, f_3 \cdots) = \frac{Ratio\ \ Value}{1+Ratio\ \ Value}$$

以上の導出法を使うと、現在判明している positive data と negative data のみを利用して全体の positive data を予測することが可能となる。ここで、判明している positive data は PPI のデータベースである DIP[6] と及び論文サーベイによるものを使用した (データ数 716)。また、negative data は、PSORT より得た各タンパク質の localization が異なるものと定義した（データ数 4514881)。

3.4 統合した手法のト レーニング、及びテスト

大腸菌において考えられる全タンパク質ペアから PPI を予測するに際して、まず PPI 予測手法における出力結果全ての組み合わせにおいて、$P\,(positive\ data|f_1, f_2, f_3 \cdots)$ を算出し、その成績を把握必要がある。よって、positive data をトレーニング集合とテスト集合に分断し、その成果の照合を行った。各手法を組み合わせた条件はスコアにより順列に並べることが出来ず、トレーニング集合とテスト集合のどちらにおいても断続的な推移になってしまうが、条件の総数は約 4200 存在するのに対し、positive data はトレーニング、テストにおいて 350 程度であるが、トレーニングとテストにおいて成績が重なるものが多く、相関係数の検定を行ったところ、有意な相関が得られ (相関係数の t 値 7.525, 臨界値 1.961)、ベイズにより統合した手法による予測結果の再現性を確認することができた。

4　結果

ベイズによる予測は、分かっている情報が多ければ多いほどその精密度が上がる。そのため positive data(716 個) 全てを利用してトレーニングした確率を利用した。

4.1　精製結果

統合した手法を用い (表 3)、実験データを matrix approach によりバイナリー化したもの (約 13000 個から信頼性の高いもの (信頼度 50 %以上) を抽出して精製を行った。その結果 904 個の PPI データを得ることが出来た。

表 3：精製に利用した PPI 予測信頼度 (一部)

系統プロファイル	発現相関	IG	MMI	Essentiality	オッズ比	確率
0.1	0.1	2	1	2	10.349	0.912
0.4	0	1	0	2	10.349	0.912
0.4	0	2	2	1	10.349	0.912
0.5	0.7	0	2	1	10.349	0.912
0.6	0.7	0	2	1	10.349	0.912
0	0.4	1	4	3	20.697	0.954
0	0.6	2	4	1	20.697	0.954
0.2	0.4	2	3	2	20.697	0.954
0.4	0.9	0	0	1	41.394	0.976
0.2	0.9	0	0	1	62.091	0.984

4.2　ゲノムワイドな PPI 予測

また、実験データを PPI 予測の条件として加えて精密化し、ゲノムワイドな PPI 予測を行ったところ (表 4)、新規に PPI を 936 個得る事に成功した。

表 4：ゲノムワイドな PPI 予測に使用した様々な条件下における PPI 予測信頼度 (一部)

系統プロファイル	発現相関	IG	実験データ	MMI	Essentiality	オッズ比	確率
0.6	0.7	0	0	2	1	10.368	0.912
0.9	0	2	1	0	2	10.368	0.912
0	0.8	2	1	0	0	20.736	0.954
0.2	0	1	1	4	0	20.736	0.954
0.2	0.4	2	1	3	2	20.736	0.954
0.4	0	1	1	1	2	20.736	0.954
0.8	0.8	0	0	0	1	20.736	0.954
0.4	0.9	0	0	0	1	41.472	0.976
0.2	0.9	0	0	0	1	62.208	0.984

5　考察

これらの予測結果から、未知機能タンパク質の相互作用データを得る事ができる。相互作用を有するタンパク質ペアは、同機能カテゴリに属する可能性が高い。そのため、未知機能タンパク質の機能は、相互作用相手のタンパク質の属する機能カテゴリに高い確率で等しいと考えられる。精製結果やゲノムワイドな予測結果をもとにして、55 個の未知機能タンパク質の機能予測をすることができた (表３)。その信頼性は各ベイズにおける条件によって変わる。PPI 予測精度が 50 %以上の時は、相互作用を有するタンパク質が同機能カテゴリに属する率 (同機能率) が 31 % (1706 個中 534 個) であり、90 %以上の時は 50 % (109 個中 54 個) になる。

表5：未知機能タンパク質の機能予測結果

遺伝子名	機能カテゴリ	PPI 予測精度	遺伝子名	機能カテゴリ	PPI 予測精度
asmA	Biosynthesis of cofacters,prosthetic groups,carriers	0.53	bglJ	Translation	0.60
gph	Transport/binding protein	0.84	emrR	Replication	0.72
gutQ	Transport/binding protein	0.61	bolA	Energy metabolism	0.53
hrpA	Translation	0.60	lar	Translation	0.75
lasT	Biosynthesis of cofacters,prosthetic groups,carriers	0.72	mcrD	Other categories	0.95
rem	Biosynthesis of cofacters,prosthetic groups,carriers	0.78	phnA	Translation	0.56
nfrA	Translation	0.81	pdxK	Other categories	0.53
rhsA	Translation	0.67	rhsB	Energy metabolism	0.91
rhsC	Translation	0.67	rhsD	Translation	0.81
rtcB	Celluler process	0.81	rtcB	Energy metabolism	0.75
smp	Biosynthesis of cofacters,prosthetic groups,carriers	0.76	smg	Transcription	0.75
sprT	Translation	0.56	sufI	Other categories	0.53
vacB	Transcription	0.67	yafL	Nucleotide metabolism	0.65
yafN	Nucleotide metabolism	0.75	yafO	Regulatory functions	0.95
yagR	Nucleotide metabolism	0.81	yagY	Translation	0.78
yahB	Transcription	0.81	yahK	Translation	0.60
yajC	Energy metabolism	0.79	yajO	Regulatory functions	0.75
ybcM	Central intermediary metabolism	0.95	ybdD	Energy metabolism	0.95
ybdF	Translation	0.75	ybdQ	Translation	0.95
ybfD	Transport/binding protein	0.65	ybgA	Translation	0.78
ybgL	Biosynthesis of cofacters,prosthetic groups,carriers	0.65	ybgF	Cell envelope	0.76
ybhA	Transport/binding protein	0.65	ybhE	Other categories	0.53
ybhJ	Amino acid metabolism	0.72	ybiA	Translation	0.78
ybiA	Translation	0.75	ybiA	Translation	0.78
ybiB	Energy metabolism	0.76	ybiR	Other categories	0.53
ycaK	Translation	0.95	yceD	Translation	0.95
ychK	Central intermediary metabolism	0.99	ychB	Translation	0.51
yciL	Biosynthesis of cofacters,prosthetic groups,carriers	0.78	yciG	Translation	0.75
yciV	Biosynthesis of cofacters,prosthetic groups,carriers	0.91	yciO	Other categories	0.65
ydaE	Biosynthesis of cofacters,prosthetic groups,carriers	0.75	ycjG	Regulatory functions	0.81
ydbC	Transport/binding protein	0.67	ydaW	Translation	0.75
ydcF	Translation	0.78	yddM	Translation	0.67
yedU	Central intermediary metabolism	0.95	ydeV	Regulatory functions	0.65
yeeE	Central intermediary metabolism	0.73	ydfE	Transcription	0.53
yefJ	Replication	0.65	yegD	Transcription	0.60
yegH	Energy metabolism	0.72	yegN	Transcription	0.67
yegO	Energy metabolism	0.67	yehU	Replication	0.72
yehW	Transport/binding protein	0.87	yeiA	Other categories	0.61
yfcF	Replication	0.81	yfcG	Transcription	0.81
yffH	Translation	0.56	ygdE	Translation	0.59
yggF	Regulatory functions	0.95	ygiE	Translation	0.79
ygjD	Fatty acid/Phospholipid metabolism	0.78	ygiN	Other categories	0.79
yhaK	Translation	0.95	yhbZ	Translation	0.61
yhgF	Cell envelope	0.51	yhgI	Transcription	0.87
yiaY	Biosynthesis of cofacters,prosthetic groups,carriers	0.95	yhjJ	Nucleotide metabolism	0.72
yieE	Amino acid metabolism	0.76	yieH	Transcription	0.67
yjbC	Central intermediary metabolism	0.81	yihA	Celluler process	0.75
yjgA	Central intermediary metabolism	0.53	ykgF	Energy metabolism	0.65
ykgK	Biosynthesis of cofacters,prosthetic groups,carriers	0.75	yliG	Nucleotide metabolism	0.81
ynaE	Translation	0.75	yneH	Other categories	0.53
yneJ	Nucleotide metabolism	0.60	yohF	Amino acid metabolism	0.78
yohI	Energy metabolism	0.61	yqcD	Energy metabolism	0.76
yqiB	Translation	0.60	yqiE	Nucleotide metabolism	0.65
yrfG	Other categories	0.99			

酵母菌だけでなく大腸菌における PPI 予測においても複数手法のベイズによる統合が有効である。加えて、系統プロファイルを実装させた今回のオリジナル手法も効果があることがわかる。しかし、今回使用した Positive Data 数はおよそ 700 である。手法のトレーニングの際してその数の少なさには不安を覚えるが、この手法の有効性を確認するには十分であったと感じる。ベイズによる予測は、判断材料が多ければ多いほど信頼性の高い予測となる。よって今後の新規 PPI 予想は、今後の Positive Data の増加に準ずるだろう。

6 謝辞

首長である斎藤輪太郎氏には多くの議論の場を設けて頂き、環境情報学部古川康一教授にはベイズに関して様々な助言を頂きました。また冨田勝教授には研究の場を提供して頂きました。この場を借りて感謝の意を表します。

参考文献

[1] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M.**A Bayesian networks approach for predicting protein-protein interactions from genomic data.**Science. 2003 Oct 17;302(5644):449-53.

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

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

[4] Saito R, Suzuki H, Hayashizaki Y. **Global insights into protein complexes through integrated analysis of the reliable interactome and knockout lethality.** Biochem Biophys Res Commun. 2003 Feb 14;301(3):633-40.

[5] Kumar A, Snyder M. **Protein complexes take the bait.** Nature. 2002 Jan 10;415(6868):123-4. No abstract available.

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

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

[8] Grigoriev PS, Lobocka MB. **Determinants of segregational stability of the linear plasmid-prophage N15 of Escherichia coli.** Mol Microbiol. 2001 Oct;42(2):355-68.

[9] Grigoriev A. **On the number of protein-protein interactions in the yeast proteome.** Nucleic Acids Res. 2003 Jul 15;31(14):4157-61.

[10] Grigoriev A. **On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res. 2003 Jul 15;31(14):4157-61.**