統計数理(2006)
第 54 巻 第 2 号 357–373

特集「予測と発見」
[研究詳解]

カーネル法による複数のゲノムデータからのタンパク質間機能ネットワークの推定

山西 芳裕[†]• Jean-Philippe Vert[†]

(受付 2006 年 1 月 4 日；改訂 2006 年 3 月 9 日)

要　　旨

本研究では，様々なゲノム関連データから，高次の生物学的機能を表すタンパク質間ネットワークを予測する手法を開発した．この方法の独自性は，教師付き学習の枠組においてネットワーク推定を行なう点と複数のデータ統合の点にある．カーネル正準相関分析を用いて，遺伝子情報を表すゲノムデータとタンパク質間ネットワークの相関モデルを構築し，それを応用して新規のタンパク質間ネットワークを予測する方法を提案した．実際の適用例として，出芽酵母のタンパク質間の機能ネットワークを，マイクロアレイ遺伝子発現情報，酵母 2 ハイブリッドシステムによる相互作用情報，タンパク質の細胞内局在情報，系統プロファイルの 4 種類のデータから予測した．クロスバリデーションによる性能評価の結果，我々の提案する複数のデータの統合と教師付き学習の効果によって，先行研究の方法よりも予測精度が著しく向上することが確認できた．そこで，全てのタンパク質セットに対して提案手法を適用し，網羅的なネットワークを推定することによって，未知のタンパク質間の機能的な関係を予測した．

キーワード：カーネル法，正準相関分析，グラフ推定，ゲノムデータ，タンパク質間ネットワーク．

1. はじめに

ゲノム情報から遺伝子やタンパク質によって成り立つ生命のはたらきを明らかにすることが，ゲノム解析の最終的な目的の一つである．生命のはたらきとは個々の遺伝子あるいはタンパク質に帰するものではなく，多数の遺伝子あるいはタンパク質が複雑に相互作用したネットワークのシステムで実現されるものである．その意味で，制御および代謝経路などのタンパク質間ネットワークは高次の生物学的な機能を表すため，ゲノム情報から未知のタンパク質のネットワークを予測することは，新しい生物学的発見に直結するため意義がある．近年の生物工学の進歩によって，遺伝子やタンパク質に関するゲノムワイドなデータが蓄積されてきた．例えば，マイクロアレイ遺伝子発現データ(Eisen et al., 1998; Spellman et al., 1998)，酵母 2 ハイブリッドによるタンパク質間相互作用情報(Uetz et al., 2000; Itoh et al., 2001)，タンパク質の局在情報(Huh et al., 2003)，系統プロファイル(Pellegrini et al., 1999)，パスウェイ情報(Kanehisa et al., 2004)などが挙げられる．そこで，これらのゲノムデータや実験データを有効に使って，高次の生物学的な機能を表すタンパク質間ネットワークを予測することが，近年のバイオイン

[†] Ecole des Mines de/ Paris, Center for Computational Biology: 35 rue Saint-Honore 77305 Fontainebleau cedex, France

フォマティクスにおいて重要課題になっている．

これまでに，このようなネットワーク推定問題に対して，様々な理論的な手法が提案されて来た．遺伝子制御ネットワークに関しては，ベイジアンネットワーク(Friedman et al., 2000)やブーリアンネットワーク(Akutsu et al., 2000)，微分方程式系(Chen et al., 1999)，グラフィカルガウシアンモデリング(Toh and Horimoto, 2002)などを用いて，マイクロアレイのデータから遺伝子ネットワークを予測する数理的な方法論が提案されている．タンパク質間物理的相互作用に関しては，遺伝子配列の共進化のパターン(Goh et al., 2000)やオーソログ遺伝子の系統樹の類似度からタンパク質間の物理的相互作用を予測するミラーツリー法(Pazos and Valencia, 2001)やその改良法(Sato et al., 2005)，残基の共進化によるインシリコ 2 ハイブリッドシステム(Pazos and Valencia, 2002)などが提案されている．また複数のゲノム情報をグラフで表し，それを結合することによって，より信頼性のあるタンパク質間の機能的な関連を予測するジョイント法(Marcotte et al., 1999)や，混合モデルのベイジアンネットワーク(Jansen et al., 2003)などが提案されている．これらを含め先行研究のほとんどの手法は，データから遺伝子またはタンパク質間のネットワークを探索的に推定するという意味で，教師無し学習の方法に属する．

本研究では，教師付き学習の枠組で，様々なゲノム情報からタンパク質間ネットワークを予測する手法を開発した(Yamanishi et al., 2004)．この方法の独自性は，教師付き学習の枠組においてネットワーク推定を行なう点にある．ここでいう教師付きとは，これまでに分かっている既知のタンパク質間ネットワークの情報を予測過程の中で用いることを意味する．第一段階として，ネットワークが既知のタンパク質セットから，ゲノムデータとネットワークの相関(ネットワーク構築原理)を，カーネル正準相関分析を用いて数学的に学習させ，モデルを構築する．第二段階として，そのモデルを，ネットワークの分かっていないタンパク質セットに当てはめ，ネットワークを予測する．教師付き学習の概念自体は，フィッシャーの判別分析，決定木，サポートベクターマシンなど，“個々のタンパク質の機能” の分類を目的とする手法として先行研究でたくさんあるが，“タンパク質間の機能的関係” で構成されるネットワークを推定する手法は，これまでに先行研究はない．

実際の適用例として，出芽酵母 *Saccharomyces cerevisiae* のタンパク質間の機能ネットワークを，マイクロアレイ遺伝子発現情報，酵母 2 ハイブリッドシステムによる物理的相互作用情報，タンパク質の細胞内局在情報，系統プロファイルの 4 種類のゲノム関連データから予測した．実験によって判明している既知のタンパク質間ネットワークを用いて，クロスバリデーション実験による性能評価の結果，本研究で提案する複数のデータの統合と教師付き学習の効果によって，先行研究の方法(教師無し学習)よりも予測精度が著しく向上することが確認できた．そこで，全てのタンパク質セットに対して提案手法を適用し，出芽酵母の 6059 個のタンパク質からなる機能的ネットワークを推定した．それを基に，未知のタンパク質間の機能的な関係を予測し，その妥当性について検討することによって，この手法が新しい生物学的な発見に繋がる可能性を示した．

2. データ

2.1 タンパク質間ネットワークの正解データ

出芽酵母 *Saccharomyces cerevisiae* のタンパク質間ネットワークの正解データとして，KEGG/PATHWAY データベース(Kanehisa et al., 2004)で保存されているタンパク質間ネットワークを利用する．KEGG/PATHWAY データベースでは，タンパク質間ネットワークは，頂点(ノード)はタンパク質(またはそれをコードする遺伝子)，辺(エッジ)がタンパク質間の機能的関係で構成される．このタンパク質間ネットワークは，主に，代謝パスウェイにおいて連続的に化学反応を触媒する酵素タンパク質間の関係を表している．最終的に，769 個のノード，3702 個

のエッジから構成されるタンパク質間ネットワークを作成した．以下では，これを信頼できるタンパク質間ネットワークとみなし，後で提案するネットワーク予測法の性能を評価するための正解データとして扱う．

2.2 マイクロアレイ遺伝子発現データ

DNA マイクロアレイを利用すれば，発生の様々な段階や異なる組織における細胞の遺伝子発現パターン，経時的な遺伝子発現の変化を系統的に調べることができる(Eisen et al., 1998; Spellman et al., 1998)．同じような発現パターンを持つ遺伝子群は，同じような機能をもつであろう，同じパスウェイ上で働く可能性が高いだろうと考えることが出来る．そこで，本研究の出芽酵母の遺伝子発現データは，Spellman らによる 77 種類の実験(Spellman et al., 1998)，Eisen らによる 80 種類の実験データ(Eisen et al., 1998)を合わせた 157 種類の実験に基づくデータを用いた．前者は，細胞周期の時間変化に対する遺伝子発現パターンを示した時系列データで，後者は，発酵から呼吸への変化，細胞分裂，胞子形成などの状態時における遺伝子発現パターンを示したデータである．各タンパク質をコードする遺伝子が，それぞれ 157 次元の数値ベクトルを持つ多変量データとなる．

2.3 酵母 2 ハイブリッドシステム

酵母 2 ハイブリッドシステムは，転写因子のドメイン構造を巧みに利用し，タンパク質間の物理的な相互作用を検出する方法である(Itoh et al., 2001; Uetz et al., 2000)．相互作用するタンパク質ペアは関連する機能を持つだろうという仮定に基づき，機能既知のタンパク質と機能未知のタンパク質の相互作用から，未知のタンパク質の機能を予測しようとする研究が近年盛んであり，実際に同じ機能を持つタンパク質はくっつきやすい傾向があることが確認されている．ただ，このデータは，ノイズが多く，疑陽性の相互作用が検出され易いという問題点も指摘されている．本研究では，2 種類の酵母 2 ハイブリッドの実験(Itoh et al., 2001; Uetz et al., 2000)に基づく，出芽酵母の 5470 個のタンパク質間物理的相互作用を用いた．これは，タンパク質が物理的に接触するかどうかを表した二項関係のデータである．酵母 2 ハイブリッドシステムで検出される相互作用は，ノイズの多いタンパク質間の関係を表すデータとみなすことができる．

2.4 タンパク質局在データ

タンパク質の細胞内局在情報に関しては，出芽酵母の全タンパク質の局在情報を網羅的に調べたデータが，近年発表された(Huh et al., 2003)．GFP (Green Fluorescent Protein)で目的タンパク質をラベルすることにより，ゴルジ体，細胞質，小胞体，核内などの様々な細胞内局在のうち，出芽酵母のタンパク質が，どこで働いているかという情報を得ることができる．出芽酵母のタンパク質の局在データは，網羅的に細胞内局在情報を調べた実験結果(Huh et al., 2003)から得た．このデータセットからは，酵母の約 6234 個のタンパク質に対して，23 個の細胞内局在のうち，出芽酵母のタンパク質が，どこで働いているかという情報を得ることができる．細胞内局在の例としては，例えば，ミトコンドリア，ゴルジ体，小胞体，核内などがあげられる．各タンパク質に対して，局在プロファイルは，タンパク質が，ある局在に対して観察されれば 1，観察されなければ 0 で表されるビット列である．

2.5 系統プロファイル

系統プロファイルとは，遺伝子がゲノムの中に存在するかどうかを生物種毎に 0，1 で表した文字列であり，各オーソログ遺伝子を様々な生物種が持つかどうかを表した情報と解釈することができる(Pellegrini et al., 1999)．各遺伝子のプロファイルは生物種毎の保存度を表すことから，それを一種の進化のパターンと考え，同じような系統プロファイルを持つ遺伝子ペア

は，共進化の観点から同じような機能を持っていると仮定して，未知の遺伝子の機能予測を行なう方法が提案されている．本研究での出芽酵母の系統プロファイルは，KEGG データベースのオーソログクラスター(Kanehisa et al., 2004)を基に作成した．KEGG データベースでは，全ゲノム配列が解読されている生物種の全遺伝子に対し，全ての組み合わせの配列類似性を Smith-Waterman スコア(Smith and Waterman, 1981)を用い計算している．そこでは，配列類似性ネットワークの中でクリーク構造になっている部分を一つのオーソロググループとみなすことで，全遺伝子に対してオーソログ情報を与えている．この研究では，11 種類の真核生物，16 種類の古細菌，118 種類の真正細菌の合計 145 生物種から構成される系統プロファイルを構築した．ここでの系統プロファイルは，出芽酵母の各タンパク質をコードする遺伝子が，上の生物種に対して存在すれば 1，存在しなければ 0 がコードされる文字列である．

2.6 データ間の関係

タンパク質間ネットワークの予測に使うための 4 つのデータ(ゲノムデータや実験データ)は，それぞれ遺伝子またはタンパク質の情報の一表現である．つまり，それぞれのデータはお互いに別の情報を担っている物であるが，タンパク質間の機能的な関係を推定するのに有効な情報だと考えることができる．本研究では，これらのデータを全て有効に使って，タンパク質間ネットワークを予測する方法を考える．

3. 方法

3.1 カーネルによるデータ表現と統合

様々なゲノムデータを統一的に扱うため，全てのデータをカーネル行列(Schölkopf and Smola, 2002)と呼ばれる類似度行列に変換することを提案する．ここでのカーネルとは，対象全体の集合を $\mathcal{X}$ とすると，対象 $\mathbf{x},\mathbf{x}' \in \mathcal{X}$ に対して，$k:[\mathbf{x},\mathbf{x}'] \to R$ と定義される関数である．またカーネル関数は，$\mathcal{X} \times \mathcal{X}$ を定義域とする実対称関数であり，半正定値性を満たすものである．直感的に，本研究におけるカーネルとは，あるデータに関するタンパク質間(またはタンパク質をコードする遺伝子間)の類似度を表すものだと解釈できる．

例えば，データセットが，遺伝子発現データ，局在情報，系統プロファイルとすれば，ガウシアンカーネル $k(\mathbf{x},\mathbf{y}) = \exp(-\| \mathbf{x}-\mathbf{y} \|^2 / \sigma^2)$ や，線形カーネル $k(\mathbf{x},\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ が自然な候補であると考えられる．データが，タンパク質間ネットワークや，酵母 2 ハイブリッドなどのグラフ構造のときは，拡散カーネル(Kondor and Lafferty, 2002)でカーネルに変換できる．ここで，拡散カーネルとは，グラフのラプラシアンを H としたとき，$K = \exp(\beta H)$ と定義できる．ここで，$\beta > 0$ は正のパラメータであり， H は $H = A - D$ (A は隣接行列， D はノードの次数を対角成分に持つ対角行列)と定義される．全てのデータをカーネルに変換する意義は，それぞれのデータ構造が，ベクトル，グラフ，文字列と異なっていたとしても，同じ数学的な枠組でデータを扱えるというメリットがある．

ここで，$P \geq 1$ 個の異質なゲノムデータが得られており，それぞれ P 個のカーネル $K_1,\ldots,K_P$ で表されているとする．K_p は p 番目のデータセットに関する，タンパク質セットの類似度行列を表す．一つの統合法として， $K^* = \sum_{p=1}^{P} w_p K_p$ と線形和を取ることでデータの統合をすることを提案する．ここでは，簡単のため重みの w_p は 1 と置くことにする．この方法の実際の有効性は確かめられている(Yamanishi et al., 2003)．それぞれのカーネル行列は，各ゲノムデータに基づくタンパク質間の類似度行列を表すので，これらの和を取ることは，多くのデータで高い類似度を示すタンパク質ペアほど，ペア間の強さがより強調される．少数のデータでしか高い類似度を示さない遺伝子ペアは，ペア間の強さは小さく抑えられ，ノイズを抑える効果が期待でき，より信頼性のあるタンパク質間の類似度行列を構築することが期待できる．

3.2 直接的なネットワーク推定法(direct approach)

ここでは，複数のゲノム関連データから，出芽酵母 *Saccharomyces cerevisiae* のタンパク質間ネットワークを予測することを考える．直接的な方法として，カーネル行列の要素自身を用いたタンパク質間のネットワークの予測が考えられる．つまり，機能的に関連のあるタンパク質ペアは，与えられたデータに関して高い類似度を持つと仮定して予測を行なうやり方である．二つのタンパク質 x と y の類似度であるカーネルの値 $K(\mathrm{x},\mathrm{y})$ が，ある閾値よりも大きければ，その 2 つのタンパク質ペアは機能的関係があるとみなす．それぞれのデータに関してカーネルを計算し，共発現するタンパク質ペア(タンパク質をコードする遺伝子ペア)，物理的に相互作用するタンパク質ペア，同じ場所で働くタンパク質ペア，共進化するタンパク質ペア，それら情報を統合したカーネルに基づき，タンパク質間ネットワークを予測する．この離散バージョンは，グラフのジョイント法によるタンパク質間相互作用予測法(Marcotte et al., 1999)に相当する．

3.3 教師無し学習に基づくネットワーク推定法(spectral approach)

クラスタリングの手法の一つとして，スペクトラルクラスタリングという方法が提案されている(Weiss, 1999; Ng et al., 2001)．これは，データのクラスターが検出しやすい特徴空間に，データのオブジェクトをまず射影して，その後に，従来のクラスター分析を行なおうというものである．これは，カーネル主成分分析(kernel principal component analysis (KPCA))(Schölkopf et al., 1998)で得られる小数の主成分で構成される空間でクラスタリングを行なうことに，ほぼ対応する(Bengio et al., 2003)．カーネル主成分分析のアルゴリズムの詳細は，参考文献(Schölkopf et al., 1998)を参照されたい．

本研究での興味は，タンパク質のクラスタリングそのものではないが，ネットワーク推定はタンパク質間の類似度の計算を伴うため，密接な関係がある．そこで，元のデータからタンパク質間の類似度を計算し，それに基づきネットワーク推定を行なうという direct approach に対して，KPCA の主成分で構成される特徴空間に射影して，そこでタンパク質間の類似度を計算し，ネットワーク推定を行なう方法が考えられる．簡単に手順を説明すると，まず各タンパク質 x を，ある特徴空間におけるベクトル $f(\mathbf{x}) = (f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x}))^{\top}$ に射影することを考える．ここで，$L < N$ であり，$f^{(l)}(\mathbf{x})$ は，l 番目の主成分に相当する．その射影された特徴空間において，もう一度タンパク質間の類似度を計算し，再計算されたタンパク質間の類似度を基に，前節で述べた direct approach を実行する．これは教師無し学習に基づくネットワーク推定法に対応し，ここでは，それを spectral approach と呼ぶことにする．

3.4 教師付き学習に基づくネットワーク推定法(supervised approach)

実際に，我々が直面している状況を簡単に説明したのが，図 1 と図 2 である．図 1 は，網羅的に得られたゲノムデータや実験データに基づくタンパク質間の類似度行列を表す．このようなゲノムデータから，高次の生物学的機能を表すタンパク質間ネットワークを予測しようというのが目的である．図 2 は，タンパク質間ネットワークの隣接行列を表す．ここで，黒色はそのタンパク質ペアは相互作用が存在する，白色はその部分はタンパク質間相互作用しない(または確認されていない)，灰色はその部分のタンパク質間の機能的な関係は未知である事を示す．ここでは，$n < N$ 個のタンパク質のネットワークは既知であり，N はタンパク質の総数を示す．

図 1 と図 2 を対比させると，タンパク質間ネットワークの情報の一部に関しては得ることができるので，ゲノムデータとタンパク質間ネットワークの対応関係に関する知識を，一部のタンパク質セットに対しては得られることに気付く．ここで，ゲノムデータからタンパク質間ネットワークができる構築原理を，何らかの形で学習できれば，その構築原理を表すモデルを，

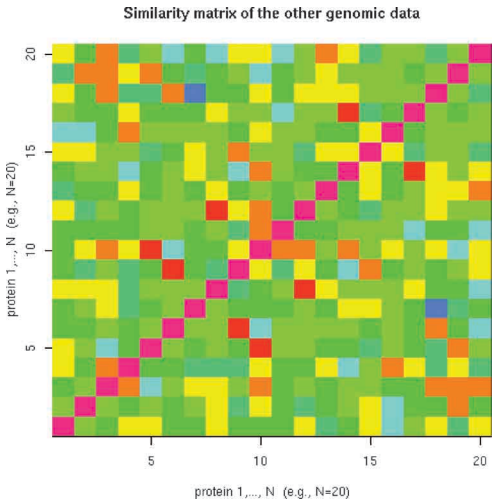

図 1. ゲノムデータに基づいて計算されたタンパク質のカーネル行列の例.

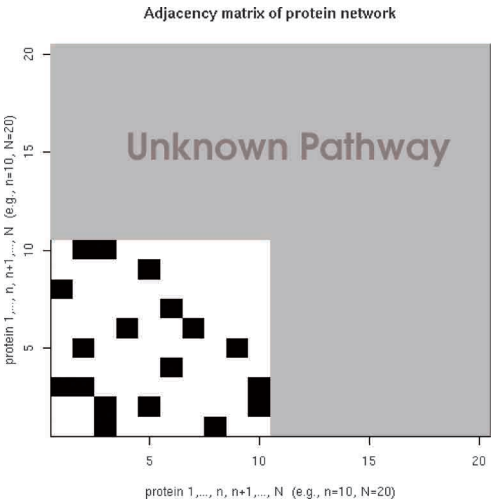

図 2. タンパク質間ネットワークを表す隣接行列の例.

ネットワーク情報が未知のタンパク質のセットに対して当てはめ，その未知の部分のタンパク質間の相互作用の関係を予測できるのではないか？と考えた．前節で示した，direct approach と spectral approach は基本的に教師無し学習なので，その意味で，図 2 に示されているような事前知識を予測に用いておらず，図 1 に示されたゲノムデータだけを用いて，タンパク質間ネットワークを探索的に予測していることに注意されたい．

本研究では，教師付き学習の枠組で，ゲノムデータとネットワークの事前知識の両方を用いて，タンパク質間ネットワークを推定することを提案する．ここで，前節で述べた spectral approach を，教師付き学習になるように修正したアルゴリズムを考える．まず，各タンパク質 x を，ある特徴空間におけるベクトル $f(\mathrm{x}) = (f^{(1)}(\mathrm{x}), \ldots, f^{(L)}(\mathrm{x}))^{\top}$ に射影することを考える．ここで，$L < N$ であり，spectral approach では，$f^{(l)}(\mathrm{x})$ は，l 番目の成分に相当する．ここでの射影の目的は，相互作用するタンパク質が，近くにいるような特徴空間を定義することである．それゆえ，x_i が x_j と相互作用するときは，$f(\mathrm{x}_i)$ は，$f(\mathrm{x}_j)$ と同じような特徴量であってほしいわけである．理想的には，$l = 1, \ldots, L$ に対して $f^{(l)}(\mathrm{x}_i)$ が $f^{(l)}(\mathrm{x}_j)$ に近ければよい．逆に，もしタンパク質間ネットワークが事前に分かるのであれば，理想の特徴空間とは $f^{(l)}$ $(l = 1, \ldots, L)$で構成される部分空間であり，タンパク質間ネットワークを表すグラフ上で隣接するノード間で滑らかに変化するものに相当すると考えられる．ここで，そのグラフに基づく拡散カーネルに関連するノルム $||f||$ は，その滑らかさの度合いを定量化したものになることが知られている(Vert and Kanehisa, 2003)．つまり，f が滑らかであればあるほど，$||f||$ の値は小さくなる．従って，もしタンパク質間ネットワークが既知であると仮定すれば，理想の特徴空間とは，グラフの拡散カーネルで主成分分析したときの主成分で構成される特徴空間となる．

実際には，真のタンパク質間ネットワークの全ての情報は事前には知ることはできないので，その理想的な特徴空間への射影は求めることはできない．しかしながら，部分的には真のネットワークの情報を知ることができるので，その部分的な既知のネットワークに適合するような理想的な特徴空間を構築し，spectral approach によって作られる特徴空間を改良することを考える．ここでは，全てのタンパク質の数を N とすると，n 個のタンパク質 $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ がネットワーク情報が分かっているタンパク質のセットであり，残りの $\{\mathrm{x}_{n+1}, \ldots, \mathrm{x}_N\}$ がネット

ワーク情報が分かっておらず，推定すべきタンパク質のセットとする．ここで，K_1 をネットワーク情報が既知のタンパク質に関するゲノムデータに基づき計算されたカーネル，K_2 をネットワーク情報が既知のタンパク質間ネットワークから計算された拡散カーネルと定義する．K_1 と K_2 は両方とも $n \times n$ の行列であり，f_1 と f_2 を $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ に基づき定義された関数，$||f_1||$ と $||f_2||$ をそれぞれに対応するノルムとする．spectral approach の枠組の中で，ノルム $||f_1||$ と $||f_2||$ が同時に小さくなるような特徴量を見つけたい．ここで，$k = 1, 2$ に対して，$\sum_{i=1}^{n} f_k(\mathbf{x}_i)^2 = 1$ を満たし，次のような量を最大にするような f_1 と f_2 を見つけたい．

$$\mathrm{corr}(f_1, f_2) \times \frac{1}{\sqrt{1 + \lambda_1 ||f_1||^2}} \times \frac{1}{\sqrt{1 + \lambda_2 ||f_2||^2}}. \tag{3.1}$$

ここで，λ_1 と λ_2 は，正の正則化パラメータを表し，$\mathrm{corr}(f_1, f_2)$ は，f_1 と f_2 の標本相関係数を表す．この式の第一項は，f_1 の事前情報のネットワークに基づく f_2 への適合を示しており，第二項と第三項は $||f_1||$ と $||f_2||$ を小さく抑えることを意味している．ここで，直交条件 $\sum_{i=1}^{n} f_k^{(l)}(\mathbf{x}_i) f_k^{(m)}(\mathbf{x}_i) = 0 \; (k = 1, 2, \quad l > m)$ を追加し，上の相関係数の最大化問題を解くことによって，逐次的に複数の関数 $f_k^{(l)}$（$k = 1, 2, \quad l = 1, \ldots, L$）を求めることを考える．実際にこのペナルティ付き相関の最大化問題は，以下のような一般化固有値問題に帰着する（Bach and Jordan, 2002）．

$$\begin{pmatrix} \mathbf{0} & K_1 K_2 \\ K_2 K_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + \lambda_1 I)^2 & \mathbf{0} \\ \mathbf{0} & (K_2 + \lambda_2 I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{3.2}$$

ここで，I は単位行列を示す．実際に，逐次的な解は $f_1 = K_1 \alpha_1$ ，$f_2 = K_2 \alpha_2$ と求めることができる．ここで，α_1 と α_2 は，式(3.2)の固有ベクトルである．この問題は，実際には，カーネル正準相関分析（Akaho，2001）の問題に帰着させることができる．もし，式(3.2)の L 個の解 $\alpha_1^{(1)}, \ldots, \alpha_1^{(L)}$ に注目するなら，それらは興味のある L 個の特徴量を $f_1^{(l)} = K_1 \alpha_1^{(l)}$ （$l = 1, \ldots, L$）と定義することになる．これらの特徴量は，既知のネットワーク情報に基づく理想的な特徴量と適合していることが期待される．これらの特徴量は，任意のタンパク質 x に対して，

$$f_1^{(l)}(\mathbf{x}) = \sum_{k=1}^{n} \alpha_1^{(l)}(\mathbf{x}_k) K_1(\mathbf{x}_k, \mathbf{x}) \tag{3.3}$$

と計算することができる．この特徴量のセットが，タンパク質間ネットワークを予測する前に射影を実行したときのタンパク質のセットである．

教師付きネットワーク推定法（supervised approach）の視覚的なイメージを図 3 と図 4 にまとめた．ステップ 1 として，ネットワーク情報が既知であるタンパク質セットをトレーニングセットとして用い，機能的な相互作用するタンパク質ペアが近くにあるような特徴空間を構築する．ステップ 2 として，相互作用が検出され易い特徴空間において，ネットワーク情報が未知であるテストセットのタンパク質の相互作用ペアを direct approach によって予測する．つまり，特徴空間において距離が近い（類似度が高い）タンパク質ペアにエッジを結ぶ．

spectral approach と supervised approach によって射影された各タンパク質 x は，L-次元のベクトルで，$\mathbf{u} = (u_1, \ldots, u_L)^\top = (f_1^{(1)}(\mathbf{x}), \ldots, f_1^{(L)}(\mathbf{x}))^\top$ と表される．射影後の特徴空間におけるタンパク質 x とタンパク質 y は，$\mathbf{u} = (u_1, \ldots, u_L)^\top$ and $\mathbf{v} = (v_1, \ldots, v_L)^\top$ と表され，そのネットワーク上におけるエッジとしての強さとして，ピアソンの相関係数のような以下の尺度を用いることにする．

$$\widehat{\mathrm{corr}}(\mathbf{u}, \mathbf{v}) = \frac{\widehat{\mathrm{cov}}(\mathbf{u}, \mathbf{v})}{\sqrt{\widehat{\mathrm{var}}(\mathbf{u})} \sqrt{\widehat{\mathrm{var}}(\mathbf{v})}} = \frac{\frac{1}{L} \sum_{l=1}^{L} (u_l - \bar{\mathbf{u}})(v_l - \bar{\mathbf{v}})}{\sqrt{\frac{1}{L} \sum_{l=1}^{L} (u_l - \bar{\mathbf{u}})^2} \sqrt{\frac{1}{L} \sum_{l=1}^{L} (v_l - \bar{\mathbf{v}})^2}}. \tag{3.4}$$

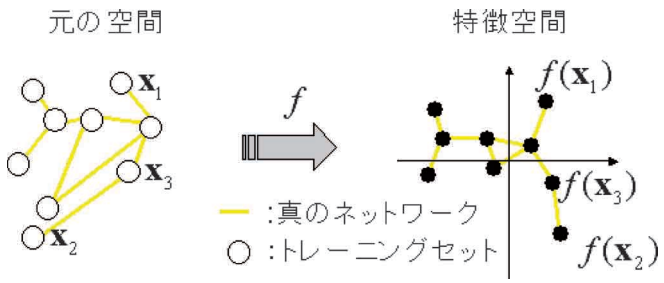

図 3. 教師付きネットワーク推定法(supervised approach)のステップ 1．ネットワーク情報が既知であるタンパク質をトレーニングデータセットとして用い，機能的な相互作用するタンパク質ペアが近くにあるような特徴空間を構築．

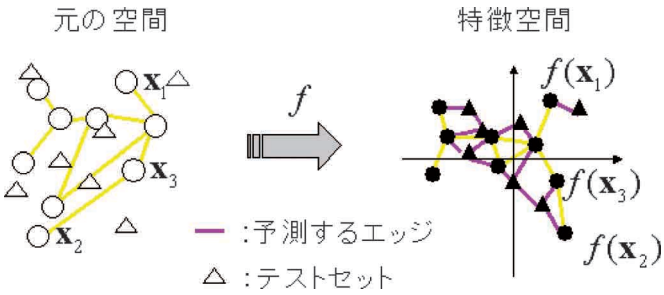

図 4. 教師付きネットワーク推定法(supervised approach)のステップ 2．相互作用が検出され易い特徴空間において，ネットワーク情報が未知であるテストセットのタンパク質の相互作用ペアを予測．

ここで，$\bar{\mathrm{u}}$ と $\bar{\mathrm{v}}$ は u と v の平均を表す．この値がある閾値よりも高ければ，タンパク質 x とタンパク質 y は，ネットワーク上で相互作用するとみなし，この値がある閾値よりも低ければ，ネットワーク上で相互作用しないであろうとみなす．この過程を全タンパク質ペアに行うことによって，網羅的なネットワークを予測する．

4. 結果

4.1 ゲノムデータの変換

全てのゲノムデータを，まずカーネルに変換した．正解データのタンパク質間ネットワークと酵母 2 ハイブリッドのデータは，グラフ構造なので，拡散カーネルを用いて，K_{gold}，K_{y2h} とそれぞれカーネルの形に変換した．ここで，パラメータは $\beta=1$ とした．ここでは，我々の過去の先行研究で拡散カーネルを利用した際と同じパラメータ値を用いた(Yamanishi et al., 2003)．遺伝子発現データは，実数値を値に取る数値ベクトルなので，ガウシアンカーネルを用いて $K_{\exp}$ と変換した．ここで，パラメータは，$\sigma=5$ とした．局在データと系統プロファイルは，ビット列なので，線形カーネルを用いて K_{loc}, K_{phy} と変換した．最終的に，全てのカーネルは，対角成分が 1 になるように基準化した(Schölkopf and Smola, 2002)．

結果として，遺伝子発現データ，酵母 2 ハイブリッド，細胞内局在情報，系統プロファイル，正解のタンパク質間ネットワークデータを，$K_{\exp}$，K_{y2h}，K_{loc}，K_{phy}，K_{gold} と，それぞれカーネル行列に変換したことになる．

4.2 タンパク質間ネットワークの予測法としての性能評価

実際に，タンパク質間の機能的なネットワーク予測としての性能を見るため，direct approach と spectral approach の性能を，個々のゲノムデータと，全てのゲノムデータを統合したカーネルの両方に対して適用した．全ての実行手順およびデータの組み合わせのリストを，表 1 の上段，中段に示す．spectral approach に対しては，最初の $L=50$ 個の主成分を，特徴空間を構成するために用いた．予測精度は，正解データのタンパク質間ネットワークをどれだけ復元できるかで評価した．ある閾値を設定し，タンパク質ペアの類似度が，閾値よりも大きい時そのタンパク質ペアは機能的な相互作用があると予測し，その閾値よりも小さい時そのタンパク質ペアは機能的な相互作用がないと予測する．閾値の値を，小さい値から少しずつ大きくしていき，それぞれの閾値の値でエッジの有無を予測したときの，true positives（予測したエッジが実際に正解データの中にあるとき）の数と，false positives（予測したエッジが正解データに無いとき）の数を記録していった．

図 5 と図 6 は，変化させていった閾値の値に対し，false positives の割合に対して true positives の割合をプロットした ROC カーブ（Gribskov and Robinson, 1996）を示している．ROC カーブでは，45 度の対角線はランダムな予測精度に相当し，左上に行けば行くほど，true positives の割合が増え予測精度が良いことを表し，対角線に近づくようだと予測精度は悪いことを表す．両方の場合とも，45 度の対角線より少し上にプロットされているが，全体的な予測精度は，あまり良くないことが図 5 と図 6 から読み取れる．ほとんどランダムな予測に近く，実用的に使える予測精度のレベルではないのが分かる．direct approach に比べると，spectral approach は，特にデータを統合したとき，少し精度が改善していることが分かる．

次に，教師付き学習に基づくネットワーク推定法を適用した．アルゴリズムの正則化パラメータ λ_1 と λ_2 はそれぞれ 0.1 とおき，特徴空間の次元数として，$L=50$ 個の特徴量を使って特徴空間を構築した．また，個々のゲノムデータの重要性，全てのデータを統合した時の効果の両方を見るために，それぞれの場合に対して性能の検証を行なった．全ての実行手順および

表 1. Direct approach, spectral approach, supervised approach に対して行なった数値実験の例.

Approach	カーネル (データ)
Direct	K_{exp} (発現データ)
Direct	K_{y2h} (酵母 2 ハイブリッド)
Direct	K_{loc} (細胞内局在情報)
Direct	K_{phy} (系統プロファイル)
Direct	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)

Approach	カーネル (データ)
Spectral	K_{exp} (発現データ)
Spectral	K_{y2h} (酵母 2 ハイブリッド)
Spectral	K_{loc} (細胞内局在情報)
Spectral	K_{phy} (系統プロファイル)
Spectral	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)

Approach	カーネル 1 (データ)	カーネル 2 (ターゲット)
Supervised	K_{exp} (発現データ)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{y2h} (酵母 2 ハイブリッド)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{loc} (細胞内局在情報)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{phy} (系統プロファイル)	K_{gold} (タンパク質ネットワーク)
Supervised	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)	K_{gold} (タンパク質ネットワーク)

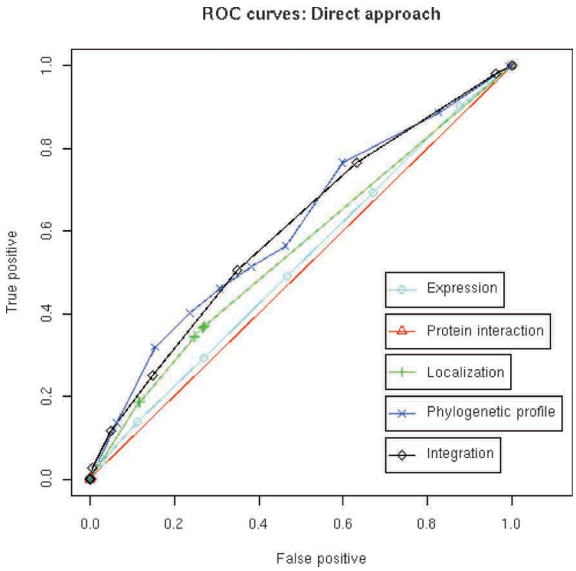

図 5. ROC **カーブ**: Direct approach**．水色は発現データのみを使った結果，赤色は酵母** 2 **ハイブリッドのみを使った結果，緑色は細胞内局在情報だけを使った結果，紺色は系統プロファイルだけを使った結果，黒色は全てのデータを統合した結果を表す．**

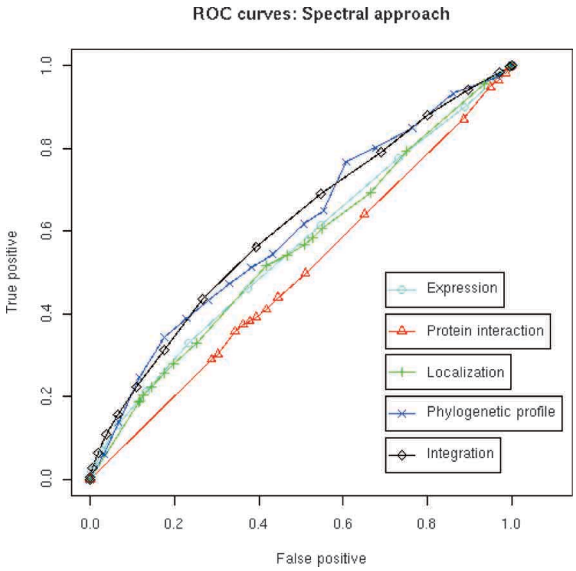

図 6. ROC **カーブ**: Spectral approach**．水色は発現データのみを使った結果，赤色は酵母** 2 **ハイブリッドのみを使った結果，緑色は細胞内局在情報だけを使った結果，紺色は系統プロファイルだけを使った結果，黒色は全てのデータを統合した結果を表す．**

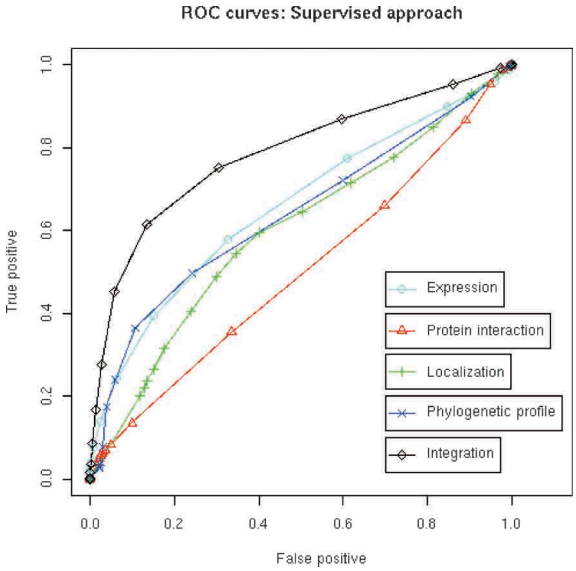

図 7. ROC **カーブ**: Supervised approach**．水色は発現データのみを使った結果，赤色は酵母** 2 **ハイブリッドのみを使った結果，緑色は細胞内局在情報だけを使った結果，紺色は系統プロファイルだけを使った結果，黒色は全てのデータを統合した結果を表す．**

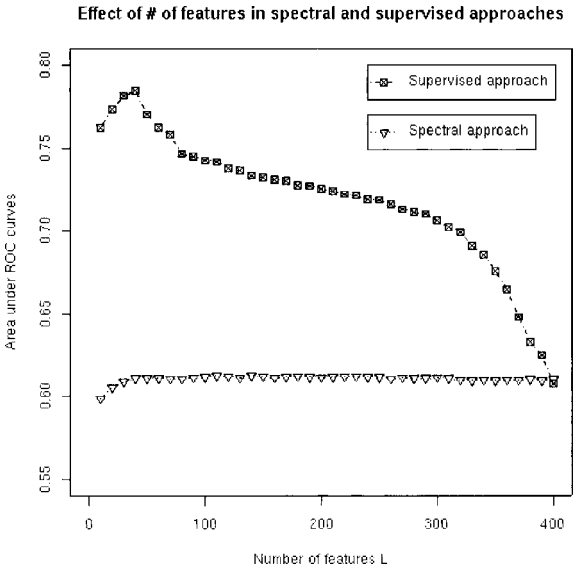

図 8. Spectral approach**，** supervised approach **における特徴量の個数の影響．特徴空間の次元数** L **を，**10 **から** 400 **まで少しずつ変化させていったときの** ROC **カーブの下の面積の変化．**x **軸は特徴空間の次元数，**y **軸は** ROC **カーブ下の面積を表す．三角のマークは** spectral approach **の結果を示し，菱形のマークは** supervised approach **の結果を示す．**

データの組み合わせのリストを，表1の下段に示す．予測精度を測るために，以下のようなクロスバリデーション実験を行なった．まず，769個のタンパク質のセットを，ランダムに9対1の割合で，トレーニングデータとテストデータに分割する．次に，トレーニングデータを基に特徴空間を学習し，テストデータのタンパク質が持つ可能性のあるペア(図2の灰色部分)のタンパク質相互作用について予測を行なった．

このクロスバリデーションの過程を10回行ない，そこで生成されるROCカーブの平均をプロットしたのが，図7である．図7を見るとこの教師付き学習のネットワーク推定法によって，精度が向上していることが分かる．個々のゲノムデータの中では，発現データと系統プロファイルが，高い寄与を与えているので，代謝パスウェイ上の機能的なタンパク質間相互作用の予測には，遺伝子の共発現の情報とタンパク質の進化的な情報が重要であることがわかる．次にタンパク質の細胞内局在情報が重要で，酵母2ハイブリッドのデータは代謝パスウェイ上のタンパク質間相互作用には，ほとんど関連が無いことを示唆している．全ての結果を比較した結果，全てのゲノムデータを統合し，かつ教師付き学習に基づくネットワーク推定を行なった結果が一番良いことが分かる．つまり，様々なゲノムデータの統合，教師付き学習の二つの効果によって，予測精度が向上していることが確認できた．

spectral approach と supervised approach では，低次元の特徴空間に射影をしてから，タンパク質間の類似度を測っているわけであるが，その特徴空間の次元である L が，予測精度に及ぼす影響を調べた．ここでは，両方の場合とも，全てのデータを統合したカーネルを用い，特徴量の次元数である L を10から400まで少しずつ変えていき，その性能を調べた．図8は，L を少しずつ変化させていったときの，ROCカーブの下の面積の変化をプロットしたものである．このスコアは1に近ければ性能が良い，0.5に近ければ性能が悪い(ランダムな予測である)ことを示す．supervised approach は特徴量の次元数に敏感であり，$L=40$ ぐらいのときに最大となり，その後，次元数の増加に伴い徐々に落ちていった．対照的に，spectral approach は，次元数が変わってもそれほど性能の変化は見られなかった．この結果は，supervised approach を実際に適用するときには，適切な次元数を設定する必要があることを示唆している．

4.3 全タンパク質に対する網羅的なネットワーク予測

クロスバリデーション実験によって，本研究で提案するネットワーク推定法の妥当性が確認できたので，次に全タンパク質を使って網羅的なタンパク質間ネットワークの予測を行なった．ここでは，酵母2ハイブリッドのデータは使わず，遺伝子発現データ，タンパク質局在情報，系統プロファイルの3種類のデータから，出芽酵母の6059個のタンパク質をノードとするネットワークを予測した．この予測したタンパク質間ネットワークによって，未知のタンパク質間相互作用だけでなく，missing 酵素の遺伝子の同定や未知のタンパク質の生化学的な機能の予測など，様々な新しい生物学的な考察が可能になることが期待できる．

一例として，機能未知であるタンパク質の生物学的機能の予測への応用を考える．ここでは，機能がよく分かっていないタンパク質YJR137Cに注目した．2003年9月の段階では，その酵素としての機能は未知であり，それが触媒する化学反応のクラスを表すEC番号は分かっていなかった．つまり，学習するためのタンパク質間ネットワークの既知ネットワークには，入っていなかったタンパク質である．我々の予測したネットワークを見てやると，YJR137Cは，酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 に繋がっていた．この二つの酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 は，硫黄の代謝パスウェイで働くことが知られているので，このYJR137Cは，硫黄に関連するような生物学的な機能を持つのではないかと推測できる．図9は，出芽酵母の硫黄の代謝パスウェイを示している．また，このパスウェイで，ターゲットのタンパク質YJR137Cは，酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 に，連続して化学反応を触媒する機能があ

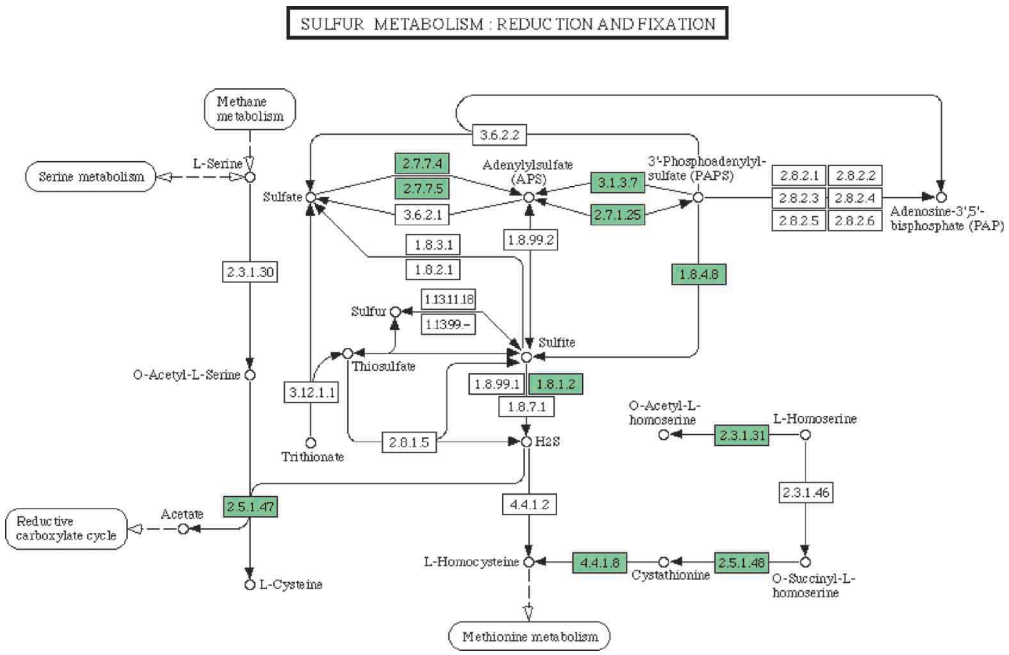

図 9. 硫黄の代謝パスウェイ．丸は化合物を表し，長方形の箱は酵素タンパク質を意味する．

るのではないかと推測でき，KEGG/PATHWAY **データベースのリファレンスパスウェイにある** EC **番号** EC:1.8.1.2 **に相当するのではないかと予測できる．近年，出芽酵母のコミュニティデータベースである** MIPS **データベースにおいて，**YJR137C **は，** EC **番号** EC:1.8.1.2 **に対応する酵素であるという報告がされていた．つまり，予測したネットワークに基づいた，このタンパク質の機能予測は当たっていたことを意味する．これは，本研究で提案する教師付き学習に基づくネットワーク推定法の有効性を支持する結果といえるであろう．**

5. 考察

本研究では，高次の生物学的機能を表すタンパク質間ネットワークを，複数のゲノム関連データから予測する手法を提案した．提案手法では，カーネル正準相関分析のモデルを用い，教師付き学習の枠組においてネットワーク推定を行なっている点と，異質なデータの統合を行なっている点が独自的である．クロスバリデーション実験において，先行研究(教師無し学習の方法論)に比べて，より良い精度を達成できた．実際の応用例では，出芽酵母の代謝パスウェイを中心とするタンパク質の機能ネットワークを，遺伝子発現データ，酵母 2 **ハイブリッド，細胞内局在情報，系統プロファイルの** 4 **つのゲノム関連データを用いて予測し，新しい生物学的な知見を得るための可能性を示した．**

予測したタンパク質間ネットワークを基に，新しい生物学的な発見を得るための応用として，未知のタンパク質間相互作用の検出，missing **酵素遺伝子の同定，機能未知のタンパク質の機能予測などが挙げられる．タンパク質がどのような機能を持つか?という意味の機能予測だけでなく，どのパスウェイのどの辺りで働くタンパク質であるか，どの酵素タンパク質と連続して化学反応に関わるのか?といった，タンパク質間の機能的な関係を含めた機能予測を可能に**

したのが，この手法の特長である．これは，従来の配列類似性に頼ったバイオインフォマティクスの方法では不可能であり，本研究で提案する手法の独自の利点である．一例として，タンパク質の機能予測の例をあげたが，同じような予測は他のパスウェイや他のタンパク質に対しても行なうことができる．ただし，新しい生物学的な発見ができたという確証を得るには，実際に生化学的な実験をして確認する必要があるだろう．

アルゴリズムの観点からみると，本研究で提案する手法は教師付き学習であるのに対し，今まで提案されている先行研究の手法は全て教師無し学習に属する．教師付き学習では，アルゴリズムの中で，既知のネットワークとそれに対応するゲノムデータの相関を自動的に学習できる点が特長である．それゆえ，生化学的な代謝パスウェイに限らず，遺伝子制御ネットワークや，シグナリングパスウェイ，物理的なタンパク質間相互作用ネットワークなど，学習過程で使うターゲットのネットワークを替えるだけで，様々な種類のネットワーク推定に利用することができる．もう一つの長所として，異質なデータを同時に統合できるという点がある．データ構造に適したカーネル関数を使って，タンパク質間の類似度行列にさえ変換できれば，どのようなデータでも統一的な枠組で扱うことができる．それゆえ，バイオインフォマティクスの分野では，ベクトル，グラフ，木構造，文字列など様々な構造を持つゲノムデータへのカーネル関数の開発が盛んである(Schölkopf et al., 2004)．実際の適用では，より良いカーネル関数を用いてゲノムデータを変換することが重要であろう．より最適なカーネル関数やそのパラメータの選択といった問題は，今後の課題である．

参 考 文 献

Akaho, S. (2001). A kernel method for canonical correlation analysis, *Proceedings of International Meeting of Psychometric Society* (*IMPS*), Springer Verlag, Tokyo.

Akutsu, T., Miyano, S. and Kuhara, S. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function, *Journal of Computational Biology*, **7**, 331–343.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis, *Journal of Machine Learning Research*, **3**, 1–48.

Bengio, Y., Vincent, P., Paiement, J.-F., Delalleau, O., Ouimet, M. and Le Roux, N. (2003). Spectral clustering and kernel PCA are learning eigenfunctions, Tech. Report, No. 1239, Département d'informatique et recherche opérationnelle, Université de Montreal.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling gene expression with differential equations, *Proceedings of Pacific Symposium on Biocomputing*, 29–40.

Eisen, M. B., Spellman, P. T., Patrick, O. B. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyze expression data, *Journal of Computational Biology*, **7**, 601–620.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners, *Journal of Molecular Biology*, **299**, 283–293.

Gribskov, M. and Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, *Computers and Chemistry*, **20**(1), 25–33.

Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast, *Nature*, **425**, 686–691.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two-

hybrid analysis to explore the yeast protein interactome, *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, **302**, 449–453.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004). The KEGG resources for deciphering the genome, *Nucleic Acids Research*, **32**, D277–D280.

Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input, *Proceedings of the International Conference on Machine Learning*, 315–322.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83–86.

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems*, **14**, 849–856.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering*, **14**, 609–614.

Pazos, F. and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs, *Proteins*, **47**, 219–227.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288.

Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding phylogenetic relationships, *Bioinformatics*, **21**(17), 3482–3489.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*, MIT Press, Cambridge, Massachusetts.

Schölkopf, B., Smola, A. J. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299–1319.

Schölkopf, B., Tsuda, K. and Vert, J.-P. (2004). *Kernel Methods in Computational Biology*, MIT Press, Cambridge, Massachusetts.

Smith, T. F. and Waterman, M. S. (1981). Identification of common modlecular subsequences, *Journal of Molecular Biology*, **147**(1), 195–197.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**(12), 3273–3297.

Toh, H. and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, **18**, 287–297.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, **10**, 601–603.

Vert, J.-P. and Kanehisa, M. (2003). Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA, *Advances in Neural Information Processing Systems*, **15**, 1425–1432.

Weiss, Y. (1999). Segmentation using eigenvectors: A unifying view, *Proceeding of the International Conference on Computer Vision*, 975–982.

Yamanishi, Y., Vert, J.-P., Nakaya, A. and Kanehisa, M. (2003). Extraction of correlated gene clusters

from multiple genomic data by generalized kernel canonical correlation analysis, *Bioinformatics*, **19**, i323–i330.

Yamanishi, Y., Vert, J.-P. and Kanehisa, M. (2004). Protein network inference from multiple genomic data: A supervised approach, *Bioinformatics*, **20**, i363–i370.

Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, 357–373 (2006) 373

Estimating Protein Network from Multiple Genomic Data by Kernel Methods

Yoshihiro Yamanishi and Jean-Philippe Vert

Ecole des Mines de Paris, Center for Computational Biology

This paper presents a new method for infering protein networks from multiple types of genomic data. Based on a variant of kernel canonical correlation analysis, the originality is in the formalization of the protein network inference problem as a supervised graph learning problem, and in the integration of heterogeneous genomic data within this framework. Promissing results are presented on prediction of the protein network for yeast *Saccharomyces cerevisiae* from four types of available data: gene expressions, protein interaction data from yeast two-hybrid systems, protein localization data, and phylogenetic profiles. It is shown that the proposed method outperforms other unsupervised network inference methods. The comprehensive prediction of a global protein network enables estimation of unknown functional relationship between proteins.

Key words: Kernel methods, canonical correlation analysis, graph inference, genomic data, protein network.