Vol.2010-BIO-21 No.19
2010/6/19

複数生物種ネットワークの同時予測：半教師つき学習によるアプローチ

鹿島久嗣[†1]　山西芳裕[†2]　加藤毅[†1]
杉山将[†3]　津田宏治[†4]

従来、生体ネットワークの予測は、遺伝子発現などの個々の生物種のもつ情報をもとに、種別に行われてきた。これに対し、本研究では「リンク伝搬法」と名付けた半教師つき学習法によって、複数の生物種のネットワークを同時に予測する方法を提案する。各生物種のもつ情報として遺伝子発現の類似度を、生物種間をまたぐ情報としてアミノ酸配列の類似度を用いて、*C. elegans*、*H. pylori* および *S. cerevisiae* のネットワークの同時予測を行い、ペアワイズ SVM などの既存手法を精度と速度の両面において上回ることを示す。

Simultaneous Inference of Multiple Biological Networks

Hisashi Kashima,[†1] Yoshihiro Yamanishi,[†2]
Tsuyoshi Kato,[†1] Masashi Sugiyama[†3]
and Koji Tsuda [†4]

The existing supervised methods for biological network inference work on each of the networks individually based only on intra-species information such as gene expression data. We believe that it will be more effective to use genomic data and cross-species evolutionary information from different species simultaneously, rather than to use the genomic data alone. We created a new semi-supervised learning method called *Link Propagation* for inferring biological networks of multiple species based on genome-wide data and evolutionary information. The new method was applied to simultaneous reconstruction of three metabolic networks of *C. elegans*, *H. pylori*, and *S. cerevisiae*, based on gene expression similarities and amino acid sequence similarities. The experimental results proved that the new simultaneous network inference method consistently improves the predictive performance over the individual network inferences, and it also outperforms in accuracy and speed other established methods such as the pairwise support vector machine.

1. はじめに

生物における多くの機能は、細胞中の複数のタンパク質が協調することによって実現され、この相互作用が生体システムの複雑さを生み出している。従って、タンパク質間の関係の分析を通して生体システムを理解することは極めて重要である。これらの関係は、タンパク質をノード、タンパク質間の関係をリンクとするネットワークとして表現することができる。生体ネットワークの例としては、代謝ネットワーク、タンパク質相互作用ネットワーク、遺伝子制御ネットワーク、シグナル伝達ネットワークなどがあるが、様々なゲノム情報や分子情報をもとに、これら生体ネットワークの予測（推定）を行うという試みは近年のバイオインフォマティクスにおけるひとつのトレンドとなっている。近年のバイオテクノロジーの進歩は、トランスクリプトーム／プロテオーム等のハイスループットでの実験を一層後押ししており、これらのデータの出現が計算機による大規模なネットワーク予測に大きく貢献しているといえる。

ある特定の生物種における生体ネットワークを予測するために利用できる情報には大きく分けて 2 種類ある。1 つはゲノム情報等の個々の生物種における情報、もう 1 つは進化的情報等の、異生物種をまたいだ情報である。前者には各生物種内の遺伝子やたんぱく質、バクテリアゲノムの染色体における遺伝子の並び[17]、進化的プロファイル、[19]、遺伝子発現パターン[12] などがある。近年では、次元削減や 2 値分類などの様々の教師つき学習手法を用いて、これらの情報を統合し、生体ネットワークを予測する試みが盛んである。例えば、カーネル正準相関分析[28]、次元削減[27]、em-アルゴリズム[11] などのアプローチや、ペアワイズ SVM と呼ばれる、タンパク質のペアを入力とした 2 値分類として定式化したサポートベクトルマシン（SVM) を用いた 2 値分類アプローチ[2] などがある。教師つき学習によるアプローチは、その適用範囲の広さと良好な予測性能から、生体ネットワーク予測のための

†1 東京大学
The University of Tokyo
†2 パリ国立高等鉱業学校
Mines ParisTech
†3 東京工業大学
Tokyo Institute of Technology
†4 産業総合研究所
National Institute of Advanced Industrial Science and Technology

標準的なツールとなりつつある一方、これらの手法は多くの計算リソースを必要とするため、スケーラビリティの問題に悩まされることが多い。例えば、前出のペアワイズ SVM において、内部で解くべき最適化問題である 2 次計画問題の必要とする計算量は $O(m^6)$、記憶用力は $O(m^4)$ となってしまう（m はネットワーク中のノード数）。

生体ネットワークの予測に用いることのできる、別のタイプの情報としては、インターログ[14),25)] と呼ばれるタンパク質相互作用の保存に関する進化的な情報がある。インターログとは、ある生物種においてタンパク質 α がタンパク質 β と相互作用を持つ場合に、別の種において、オーソログにあたるタンパク質 α' と β' の間にも相互作用がある可能性が高いとする考え方である。この考え方は、物理的な相互作用にのみ適用可能というわけではなく、例えば、完全にゲノム解読された生物種における代謝ネットワークの予測に用いられた事例もある[15)]。しかしながら、インターログに基づくアプローチは、異生物種の間で、配列に十分な相同性が無い場合には適用できず、種特有の相互作用などは予測することができないという欠点がある。現在までのところ、生体ネットワークの予測において、ゲノムデータに基づくアプローチと進化情報に基づくアプローチは別々に研究が行なわれてきた[11),14),25),27),28)] が、両方を組み合わせることによって予測の質を高めることができると期待される。現存する教師つき学習に基づく生体ネットワーク予測法の殆どは、個々の生物種のゲノムデータから、その生物種のネットワークを予測することを想定している。そこで、我々は、教師つき学習の枠組みにおいて、進化情報も併せて用いることで予測を改善することを目指す。（実際、その有効性は、教師なし学習の文脈では示されている[22)]。）

本論文では、**リンク伝播法**と名付けた半教師つき学習によるネットワーク予測法を提案する。この方法は、ゲノム情報と進化情報の両方を用いることによって、複数生物種の生体ネットワークを同時に予測することができる。従来の手法は、各々の生物種におけるゲノム情報を用いることで、各生物種のネットワークを別々に予測する（図 1(a) 参照）のに対し、提案手法では、配列情報等の進化情報も併せて用いることによって、これらを同時に推定する (図 1(b) 参照)。半教師つき学習の代表的な手法の 1 つとしてラベル伝播法[29),30)] と呼ばれる手法があり、バイオインフォマティクスの様々な問題に対して応用されている[7),16),23),26)] が、提案手法のリンク伝播法はこの拡張にあたる。我々の知る限りでは、本研究は半教師つき学習を生体ネットワーク予測に応用した初めての試みといえる（図 3 参照）

提案手法の重要な特長としては、その計算効率の高さを挙げることができる。提案手法と同じ類似度情報を用い、高い予測精度で知られるペアワイズ SVM と比較して、提案手法は遥かに高速で、少ない記憶容量しか必要としない。提案手法は “vec トリック[13),24)]”（図 2 参照）と呼ばれるテクニックを用いることによって、提案手法における巨大な連立方程式を解く際の計算量と記憶容量を大幅に削減することに成功している。提案手法のアルゴリズムは高速化した共役勾配法を用いるが、その計算量は $O(m^5)$ であり、記憶容量は $O(m^2)$ しか必要としない。これはペアワイズ SVM が $O(m^6)$ の計算量と、$O(m^4)$ の記憶容量を必要とするのと比較して大幅な削減といえる。実際、我々の実施した実験では、100 倍の高速化を実現している。さらに、提案手法は非常にシンプルであり、最適化のソフトウェア等を使う必要もなく、実装も容易である。

提案手法により、同一生物種内の情報と異生物種をまたいだ情報を用いて *C. elegans*、*H. pylori* および *S. cerevisiae* の 3 生物種の代謝ネットワークの予測実験を行った。同一生物種内の情報としてはマイクロアレイによって測定された遺伝子発現情報を、異生物種をまたいだ情報としては、タンパク質のアミノ酸配列を用いた。実験の結果、(i) 複数生物種の生体ネットワークの同時推定は予測精度を向上させること、(ii) 提案手法の予測精度は、従来手法のペアワイズ SVM を上回りながらも、その計算量は著しく削減されること、の 2 点を確認した。

2. デ ー タ

2.1 代謝ネットワークデータ

本研究では、*C. elegans*、*H. pylori* および *S. cerevisiae* の 3 生物種の代謝ネットワークを扱う。代謝ネットワークデータは KEGG PATHWAY[10)] のものを用いた。代謝ネットワークはグラフとして表現でき、各ノードがタンパク質（酵素）を表し、リンクは 2 つの酵素が連続する反応を触媒することを表す。3 つの生物種のネットワークのノード数は、それぞれ 532（*C. elegans*）, 291（*H. pylori*）、722（*S. cerevisiae*）、リンク数はそれぞれ 2,892（*C. elegans*）, 492（*H. pylori*）、2,323（*S. cerevisiae*）であった。なお、リンクは主要な反応だけでなく、種に特有の反応も含んでいる。最終的に、3 つのネットワークの隣接行列 $\mathsf{A}^{(1)}$、$\mathsf{A}^{(2)}$、$\mathsf{A}^{(3)}$ を得た。

2.2 種間の配列類似度データ

3 つの生物種におけるタンパク質のアミノ酸配列は KEGG GENES[10)] から取得した。2 つのタンパク質 g と g' の配列類似度は Smith-Waterman スコアを正規化したもの $s(g,g')/(\sqrt{s(g,g)}\sqrt{s(g',g')})$ を用いた。ここで、$s(\cdot,\cdot)$ は元々の Smith-Waterman スコア[20)] である。最終的に、3 種間の配列類似度行列 $\mathsf{W}^{(1,2)}$（*C. elegans* vs. *H. pylori*）、$\mathsf{W}^{(2,3)}$（*H. pylori* vs. *S. cerevisiae*）、$\mathsf{W}^{(3,1)}$（*S. cerevisiae* vs. *C. elegans*）を得た。な

お、$\mathsf{W}^{(1,2)} = \mathsf{W}^{(2,1)\top}$、$\mathsf{W}^{(2,3)} = \mathsf{W}^{(3,2)\top}$、$\mathsf{W}^{(3,1)} = \mathsf{W}^{(1,3)\top}$ とする。

2.3 遺伝子発現データ

3 生物種における遺伝子発現データは、Gene Expression Omnibus[21] より構築された MSGR[4] から取得した。それぞれ 1,209 個（*C. elegans*）、293 個（*H. pylori*）、753 個（*S. cerevisiae*）のマイクロアレイデータからなっている。従って、各種におけるそれぞれの酵素は、同じだけの次元をもつベクトル $\mathbf{x}$ で表現され、（欠損値を各実験の平均発現量で置き換えたのち）2 つの酵素の遺伝子発現量の類似度は動径基底関数（RBF）カーネル $k(\mathbf{x}, \mathbf{x}') \equiv \exp(-||\mathbf{x} - \mathbf{x}'||^2/(2\gamma^2))$ によって計算した（$\gamma \equiv 2$ とした）。最終的に、3 つの遺伝子発現類似度行列 $\mathsf{W}^{(1,1)}$（*C. elegans*）、$\mathsf{W}^{(2,2)}$（*H. pylori*）、$\mathsf{W}^{(3,3)}$（*S. cerevisiae*）を得た。

3. 提 案 手 法

3.1 問 題 設 定

n 生物種の生体ネットワークを予測したいものとする（我々の実験では $n = 3$ の場合を扱う）。$m^{(k)}$ を k 番目の生物種のネットワークにおけるノード（タンパク質）数、m を最大のネットワークのノード数とする。$\mathsf{A}^{(k)}$ を k 番目のネットワークの隣接行列とし、その (i,j) 要素 $[\mathsf{A}^{(k)}]_{i,j}$ は、i 番目と j 番目のノード間にリンクがある場合に $[\mathsf{A}^{(k)}]_{i,j} \equiv 1$、リンクが無い場合は $[\mathsf{A}^{(k)}]_{i,j} \equiv -1$ とする。リンクの有無が不明の場合には $[\mathsf{A}^{(k)}]_{i,j} \equiv 0$ とする。$A \equiv (\mathsf{A}^{(1)}, \mathsf{A}^{(2)}, \ldots, \mathsf{A}^{(n)})$ を隣接行列の（順序）集合とする。なお、我々の実験では各 $\mathsf{A}^{(k)}$ は対称（無向ネットワーク）であるが、提案手法自体は対称でない場合にも適用可能である。

我々の目的は、0（リンクの有無が未知）であるような A の要素に対し、リンクの有無を予測することである。そのために、提案手法のアルゴリズムは行列の（順序）集合 $F = (\mathsf{F}^{(1)}, \mathsf{F}^{(2)}, \ldots, \mathsf{F}^{(n)})$（それぞれの要素が $m^{(k)} \times m^{(k)}$ 行列）を出力する。$\mathsf{F}^{(k)}$ の (i,j) 要素はリンク強度 、すなわち i 番目のノードと j 番目のノードの間にどのくらいの確信度でリンクがありそうかを示す。（大きいほど確信度が高いものとする。）

ネットワークの既知部分 A に加え、個々のノード（タンパク質）の情報、例えばタンパク質のアミノ酸配列や遺伝子発現情報なども利用できる。これらは n^2 個の非負の類似度行列 $\{\mathsf{W}^{(k,\ell)}\}_{k,\ell=1}^{n}$ として与えられるとする。ここで、$\mathsf{W}^{(k,\ell)}$ は k 番目のネットワークと ℓ 番目のネットワークのノード間の類似度を表す行列であり、$\mathsf{W}^{(k,\ell)}$ の (i,j) 要素 $[\mathsf{W}^{(k,\ell)}]_{i,j}$ は、k 番目のネットワークの i 番目のノードと、ℓ 番目のネットワークの j 番目のノードの間の（非負の）類似度である。なお、$\mathsf{W}^{(k,\ell)} = \mathsf{W}^{(\ell,k)\top}$ であることに注意する。我々の実験では、同一種内の類似度行列 $\mathsf{W}^{(k,\ell)}$（$k = \ell$）は遺伝子発現類似度によって、種間の類似度行列は $\mathsf{W}^{(k,\ell)}$（$k \neq \ell$）はアミノ酸配列の類似度によって定義する。

以上の問題をまとめると以下のようになる：

入力：

· n 生物種のネットワークの既知部分を表した隣接行列 $A = (\{\mathsf{A}^{(k)}\}_{k=1}^{n})$

· ノード間の類似度を表した、n^2 個の類似度行列 $\{\mathsf{W}^{(k,\ell)}\}_{k,\ell=1}^{n}$

出力： 各ノード間のリンク強度（予測結果）を表す n 個の行列 $F = (\{\mathsf{F}^{(k)}\}_{k=1}^{n})$

3.2 定 式 化

我々の目的は、ネットワーク構造の既知部分とノード間の類似度情報を用いて、ネットワーク構造の未知部分に対するリンク強度を推定することである。そのための推論原則として“リンク伝播原則”、すなわち「2 組のノードペアがお互いに似ているならば、近いリンク強度をもつ」という仮説を用いる（図 3）。これは、良く知られた半教師つき学習の手法であるラベル伝播法[29),30)] が用いる推論原則を、ノードペアに対して拡張したものである。ラベル伝播はもともと本研究の問題とは異なり、ノードの分類を行うために提案された手法であり、ラベル伝播は「ネットワーク上で隣り合うノードは、同じラベル（分類）に属する可能性が高い」という仮説を用いて推論を行う。例えば、タンパク質の相互作用ネットワーク上において、タンパク質（ノード）のそれぞれがどのような機能をもつかを予測するなどの目的に利用できる。我々が予測したいのは 2 つのノードの間の関係であるため、ラベル伝播の原則をノードペアに対して用いる。例えば、k 番目のネットワークにおける $(i^{(k)}, j^{(k)})$ と ℓ 番目のネットワークにおける $(i^{(\ell)}, j^{(\ell)})$ の 2 組のノードペアがあるものとする。リンク伝播の原則を用いると、もし 2 組のペアがお互いに似ていれば、これらに対するリンク強度 $[F^{(k)}]_{i^{(k)},j^{(k)}}$ と $[F^{(\ell)}]_{i^{(\ell)},j^{(\ell)}}$ が近い値を持つべきということを意味する。

以上を最小化問題として定式化するためには、2 組のノードペア間の類似度を定義することが必要になる。そこで、k 番目のネットワークにおけるノードペアと ℓ 番目のネットワークにおけるノードペアの間の類似度を $m^{(k)^2} \times m^{(\ell)^2}$ 行列 $\tilde{\mathsf{W}}^{(k,\ell)}$ として定義する。この行列の $(i^{(k)} + m^{(k)}j^{(k)}, i^{(\ell)} + m^{(\ell)}j^{(\ell)})$ 要素 $[\tilde{\mathsf{W}}^{(k,\ell)}]_{i^{(k)}+m^{(k)}j^{(k)},i^{(\ell)}+m^{(\ell)}j^{(\ell)}}$ はノードペア $(i^{(k)}, j^{(k)})$ とノードペア $(i^{(\ell)}, j^{(\ell)})$ の間の類似度を表す。この類似度の定義として、2 組のペアのそれぞれから取ってきたノードが互いに似ているときに、これらは似ていると定義するのは自然であると思われる（図 3 参照）。本論文では、2 組のノードペア間の類似度行列

をノード類似度行列のクロネッカー積として次のように定義する。

$$\tilde{\mathsf{W}}^{(k,\ell)} \equiv \mathsf{W}^{(k,\ell)} \otimes \mathsf{W}^{(k,\ell)} \tag{1}$$

ここで、$\otimes$ は行列のクロネッカー積を表すものとし、式 (1) を要素ごとに書けば、

$$[\tilde{\mathsf{W}}^{(k,\ell)}]_{i^{(k)}+m^{(k)}j^{(k)},i^{(\ell)}+m^{(\ell)}j^{(\ell)}} \equiv [\mathsf{W}^{(k,\ell)}]_{i^{(k)},i^{(\ell)}} [\mathsf{W}^{(k,\ell)}]_{j^{(k)},j^{(\ell)}}$$

となる。これは、カーネル法などで用いられる定義[1),2),18)] と基本的には同じものである。

リンク伝播原則を表現するために、以下の目的関数を定義する。

$$J(F) \equiv \frac{\sigma}{2} \mathbf{vec}\,(F)^{\top} \mathsf{L} \mathbf{vec}\,(F) + \frac{1}{2} \parallel \mathbf{vec}\,(F) - \mathbf{vec}\,(A^*) \parallel_2^2 \tag{2}$$

ここで、$A^* \equiv (\mathsf{A}^{(1)^*}, \mathsf{A}^{(2)^*}, \ldots, \mathsf{A}^{(n)^*})$ とする。また $\mathsf{A}^{(k)^*}$ は $m^{(k)} \times m^{(k)}$ 行列で、以下で定義されるような予測の目標値を表すものとする。

$$[\mathsf{A}^{(k)^*}]_{i^{(k)},j^{(k)}} \equiv \begin{cases} \frac{|A^{(k)^+}|+|A^{(k)^-}|}{|A^{(k)^+}|} & \text{if } [\mathsf{A}^{(k)}]_{i^{(k)},j^{(k)}} = 1 \\ -\frac{|A^{(k)^+}|+|A^{(k)^-}|}{|A^{(k)^-}|} & \text{if } [\mathsf{A}^{(k)}]_{i^{(k)},j^{(k)}} = -1 \\ 0 & \text{それ以外} \end{cases}$$

ここで、**vec** は、ある行列に対して、その行列の列を縦に並べてベクトルとするような作用を表すものとする。**vec** を行列の（順序）集合 F について適用したときには、$\mathbf{vec}\,(F) \equiv \mathbf{vec}\left(\left[\mathbf{vec}\left(\mathsf{F}^{(1)}\right), \mathbf{vec}\left(\mathsf{F}^{(2)}\right), \ldots, \mathbf{vec}\left(\mathsf{F}^{(n)}\right)\right]\right)$ となるものとする。また、式 (2) において、L は以下で定義されるラプラシアン行列とする。

$$\mathsf{L} \equiv \begin{bmatrix} \tilde{\mathsf{D}}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \tilde{\mathsf{D}}^{(n)} \end{bmatrix} - \begin{bmatrix} \tilde{\mathsf{W}}^{(1,1)} & \cdots & \tilde{\mathsf{W}}^{(1,n)} \\ \vdots & \ddots & \vdots \\ \tilde{\mathsf{W}}^{(n,1)} & \cdots & \tilde{\mathsf{W}}^{(n,n)} \end{bmatrix} \tag{3}$$

なお、$\tilde{\mathsf{D}}^{(k)}$ は ${m^{(k)}}^2 \times {m^{(k)}}^2$ の対角行列であり、その対角成分は $\tilde{\mathsf{D}}^{(k)} = \sum_{p=1}^{n} \mathsf{D}^{(k,p)} \otimes \mathsf{D}^{(k,p)}$ と定義されるものとする。$\mathsf{D}^{(k,p)}$ は $m^{(k)} \times m^{(k)}$ 行列で、その対角成分は、$[\mathsf{D}^{(k,p)}]_{i,i} = \sum_{j=1}^{m^{(p)}} [\mathsf{W}^{(k,p)}]_{i,j}$ と表されるものとする。

目的関数 (2) を最小化する F を求めるために $\frac{\partial J}{\partial \mathbf{vec}(F)} = \mathbf{0}$ とおくと、以下の連立方程式を得る。

$$(\sigma \mathsf{L} + \mathsf{I})\, \mathbf{vec}\,(F) = \mathbf{vec}\,(A^*) \tag{4}$$

連立方程式 (4) の解を得るために、我々は共役勾配法[5)] を高速化したものを用いる。共役勾配法の素朴な適用は、$O(m^4)$ の記憶要領と $O(m^6)$ の計算量を必要としてしまうが、“vec トリック[13),24)]”と呼ばれるテクニック（図 2 参照）を用いることによって、これらを $O(m^2)$ の記憶容量と $O(m^5)$ の計算量に減らすことができる[⋆1]。

4. 結　　果

4.1 実験の設定

複数の生物種のネットワークを同時予測することで、これらを個別に予測するよりも高い精度が得られることを確認するための実験を行った。また、提案手法を、高い性能を持つことが確認されているペアワイズ SVM[2)] と比較し、提案手法が精度と速度の両面でこれを上回ることを示した。

我々の文脈では、それぞれのネットワークを個別に予測することは、各種における配列類似度のみを用いて予測を行うことに等しい。そこで、個別予測の場合、 $k = \ell$ については $\mathsf{W}^{(k,\ell)}$ を用い、$k \neq \ell$ については $\mathsf{W}^{(k,\ell)}$ の全ての要素を 0 と置いた。同時予測の場合には、全ての (k,ℓ) について $\mathsf{W}^{(k,\ell)}$ を用いた。なお、2 節で述べたように、$k = \ell$（同一種内）については、$\mathsf{W}^{(k,\ell)}$ としてガウスカーネル（$\gamma \equiv 2$ としたもの）を用いた。一方、$k \neq \ell$（異種間）については正規化した Smith-Waterman スコアを用いた。全ての類似度行列は次数 2 の多項式カーネルと組み合わせて用いた。

2 つ目の実験結果については、提案手法をペアワイズ SVM (P-SVM)[2)] と比較した。P-SVM において用いられる（直積）ペアワイズカーネル[1),2),18)] は、我々の用いる類似度行列と基本的に同一であるため[⋆2]、P-SVM もまた同時予測に用いることができる。ペアワイズカーネルの基本となるカーネルとしては、提案手法と同じ $\{\mathsf{W}^{(k,\ell)}\}_{k,\ell=1}^{n}$ を用いた[⋆3]。提案手法と同じく、ペアワイズカーネルの定義を変えることによって個別予測と同時予測を切り替えることができる。また、この実験においてはネットワークは対称であるためカーネル関数も対称化したものを用いた[⋆4]。ペアワイズカーネル行列は記憶領域に明示的に構成するには大きすぎるため、SVM^{light} [9)] などの SVM の標準的な実装をそのまま用いることはできない。そこで、各ステップで 1 つずつ訓練データを処理するオンライン学習アルゴリズム

⋆1 アルゴリズムの詳細は `http://cbio.ensmp.fr/~yyamanishi/LinkPropagation/` を参照されたい。
⋆2 厳密には、ペア間の類似度行列は、カーネル関数の条件である半正定性を必ずしも満たすわけではないが、我々の実験においては半正定を確認済みである。
⋆3 P-SVM はノードペア $(i^{(k)}, j^{(k)})$ のリンク強度を、

$$[\mathsf{F}^{(k)}]_{i^{(k)},j^{(k)}} \equiv \sum_{\ell=1}^{n} \sum_{i^{(\ell)},j^{(\ell)}=1}^{m^{\ell}} \alpha^{(\ell)}_{i^{(\ell)},j^{(\ell)}} [\mathsf{W}^{(k,\ell)}]_{i^{(k)},i^{(\ell)}} [\mathsf{W}^{(k,\ell)}]_{j^{(k)},j^{(\ell)}} + [\mathsf{W}^{(k,\ell)}]_{i^{(k)},j^{(\ell)}} [\mathsf{W}^{(k,\ell)}]_{j^{(k)},i^{(\ell)}}$$

によって与える。α はモデルのパラメータである。
⋆4 提案手法では、解の対称性は目標値 A の対称性から自動的に満たされる。

を用いることで、計算効率と記憶領域の問題を回避することにする。我々の実験では、2 乗ヒンジ損失を用いた SVM に漸近的に解が一致することで知られる PUMMA[8] を用いた。また、ベースライン手法としてはカーネル回帰（Nadaraya-Watson 推定[3]）を用いた。

P-SVM の正則化パラメータは $C \equiv 1$ とし、全ての訓練データを 3 周するまで学習を行った。一方、提案手法のハイパーパラメータは $\sigma \equiv 10^{-3}$、$\epsilon \equiv 10^{-5}$ とした。予測精度は 5 回の実験の平均 AUC （Area Under the ROC Curve[6]）を用いた。1 回の実験は、全てのノードペアのうちランダムに 25% か 50%、もしくは 75% を選び訓練データとして用いた。

4.2 結　　果

同時予測 vs. 個別予測　表 1 は、提案手法において個別予測（同一種内発現類似度のみを用いたもの）と同時予測（異種間配列類似度も用いたもの）を比較したものである。訓練データの割合を 25%、50%、75% と変えた場合の平均 AUC を、標準偏差とともに示してある。実験結果を見ると、同時予測を行うことによって殆ど全ての場合に置いて性能が向上（最大で AUC が 0.06 程度改善）していることがわかる。この結果は、異種間の類似度が種をまたいで予測に有用な情報を伝えていることを意味している。

また、C. elegans に対する性能向上は他の種を上回っているが、我々はこの理由は、リンクの密度の違いに起因するのではないかと考えている。3 生物種におけるリンク密度は、それぞれ 0.020（*C. elegans*）、0.012（*H. pylori*）、0.009（*S. cerevisiae*）であり、*C. elegans* のネットワークが最も密である。

提案手法 vs. ペアワイズ SVM（P-SVM）　表 2 は、提案手法と P-SVM、ベースライン手法であるカーネル回帰を比較したものである。表より、提案手法が他手法を一貫して上回っていることが分かる。この差は恐らく、提案手法は半教師つきの手法であるためリンクが未知であるテストデータの情報も用いることができるが、他の手法は通常の教師つき学習法であるためリンクの有無が既知であるようなノードペアの情報しか用いることができないことから来るものであると我々は考えている。

最後に、図 4 は提案手法と P-SVM の実行時間を対数スケールで比較したものである。P-SVM は訓練とテストの 2 つの段階があるが、提案手法はこれをまとめて行うため、総実行時間での比較を行っている。個別予測の実行時間は、各生物種に対する実行時間の和とした。実験は全て Intel® Core™2 Duo CPU T8300（2.40-GHz CPU）および 2.0GB RAM を搭載した Microsoft® Windows XP® マシンの上で、R 言語を用いて行った。なお、P-SVM の訓練は全訓練データを 3 周分行った。

比較より、提案手法は P-SVM よりも遥かに高速であることが分かる。特に、25% の訓練データを用いた時には 100 倍、25% の訓練データを用いた時に至っては 300 倍もの差がある。この差は、オンライン SVM の実装において、全訓練データの繰り返し数（我々の実験では 3 回）を減らせば小さくはなるが、その差は歴然である。また、表 2 の結果からも、繰り返し数を減らすことは予測性能の低下を引き起こすものと思われる。

参　考　文　献

1) Basilico, J. and Hofmann, T.: Unifying collaborative and content-based filtering, *Proceedings of the 21st International Conference on Machine Learning (ICML)* (2004).
2) Ben-Hur, A. and Noble, W.: Kernel methods for predicting protein-protein interactions, *Bioinformatics*, Vol.21, No.Suppl. 1, pp.i38–i46 (2005).
3) Bishop, C.: *Pattern Recognition and Machine Learning*, Springer (2006).
4) Chen, C., Weirauch, M., Powell, C., Zambon, A. and Stuart, J.: A search engine to identify pathway genes from expression data on multiple organisms, *BMC Systems Biology*, Vol.1, p.20 (2007).
5) Golub, G. and Loan, C.V.: *Matrix computations (3rd ed.)*, Johns Hopkins University Press (1996).
6) Gribskov, M. and Robinson, N.: Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, *Computers and Chemistry*, Vol.20, pp.25–33 (1996).
7) Hwang, T., Sicotte, H., Tian, Z., Wu, B., Kocher, J.-P., Wigle, D., Kumar, V. and Kuang, R.: Robust and efficient identification of biomarkers by classifying features on graphs, *Bioinformatics*, Vol.24, No.18, pp.2023–2029 (2008).
8) Ishibashi, K., Hatano, K. and Takeda, M.: Online Learning of Approximate Maximum p-Norm Margin Classifiers with Biases, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)* (2008).
9) Joachims, T.: *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, Kluwer Academic Publishers (2003).
10) Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y.: KEGG for linking genomes to life and the environment, *Nucleic Acids Research*, Vol.36, No.Database issue, pp.D480–484 (2008).
11) Kato, T., Tsuda, K. and Asai, K.: Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, Vol.21, No.10, pp.2488–2495 (2005).
12) Kharchenko, P., Vitkup, D. and Church, G.: Filling gaps in a metabolic network using expression information, *Bioinformatics*, Vol.20, pp.449–453 (2004).

表 1 個別予測と同時予測の比較結果。予測精度は AUC で計測したもの。 同時予測の性能は、個別予測のそれを上回っていることが分かる。

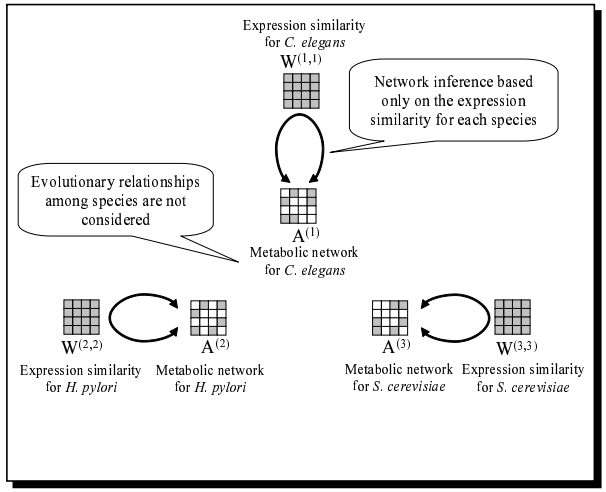
	C. elegans		H. pylori		S. cerevisiae		全体	
訓練データの割合	提案手法 (個別予測)	提案手法 (同時予測)	提案手法 (個別予測)	提案手法 (同時予測)	提案手法 (個別予測)	提案手法 (同時予測)	提案手法 (個別予測)	提案手法 (同時予測)
25 %	0.702±0.004	0.747±0.005	0.600±0.007	0.616±0.007	0.851±0.005)	0.865±0.004	0.749±0.002	0.780±0.002
50 %	0.712±0.005	0.776±0.008	0.617±0.009	0.635±0.008	0.901±0.005	0.909±0.005	0.786±0.005	0.820±0.005
75 %	0.727±0.008	0.791±0.008	0.629±0.016	0.653±0.021	0.921±0.008	0.928±0.009	0.806±0.006	0.840±0.005

表 2 提案手法、ペアワイズ SVM（P-SVM）、カーネル回帰（KR）の比較。予測精度は AUC で計測したもの。提案手法は他手法を上回っていることがわかる。

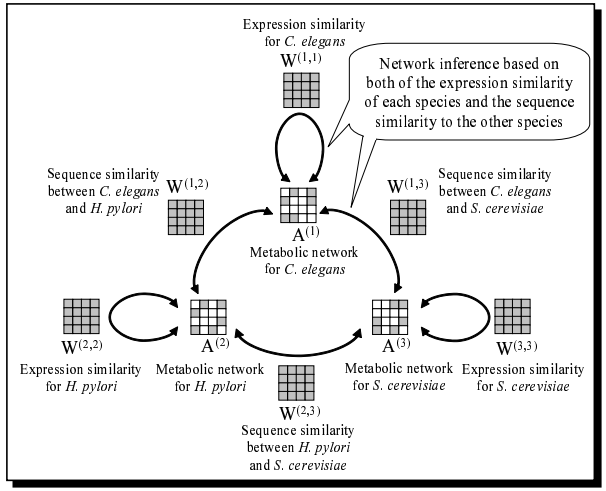
	C. elegans			H. pylori		
訓練データの割合	KR (同時予測)	P-SVM (同時予測)	提案手法 (同時予測)	KR (同時予測)	P-SVM (同時予測)	提案手法 (同時予測)
25 %	0.593±0.002	0.722±0.007	0.747±0.005	0.565±0.009	0.604±0.002	0.616±0.007
50 %	0.599±0.006	0.752±0.008	0.776±0.008	0.565±0.005	0.628±0.012	0.635±0.008
75 %	0.605±0.012	0.774±0.013	0.791±0.008	0.575±0.009	0.648±0.018	0.653±0.021
	S. cerevisiae			全体		
訓練データの割合	KR (同時予測)	P-SVM (同時予測)	提案手法 (同時予測)	KR (同時予測)	P-SVM (同時予測)	提案手法 (同時予測)
25 %	0.822±0.009	0.832±0.007	0.865±0.004	0.727±0.002	0.746±0.005	0.780±0.002
50 %	0.883±0.002	0.884±0.005	0.909±0.005	0.755±0.003	0.789±0.006	0.820±0.005
75 %	0.914±0.006	0.914±0.004	0.928±0.009	0.765±0.004	0.813±0.004	0.840±0.005

13) Laub, A.: *Matrix Analysis for Scientists and Engineers*, Society for Industrial and Applied Mathematics (2005).
14) Matthews, L., Vaglio, P., Reboul, J., Ge, H., Davis, B., Garrels, J., Vincent, S. and Vidal, M.: Identification of potential interaction networks using sequence based searches for conserved protein-protein interactions or "interlogs", *Genome Research*, Vol.11:2, pp.2120–2126 (2001).
15) Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. and Kanehisa, M.: KAAS: an automatic genome annotation and pathway reconstruction server, *Nucleic Acids Research*, Vol.35, pp.W182–W185 (2007).
16) Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. and Morris, Q.: GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function, *Genome Biology*, Vol.9, No.Suppl. 1, p.S4 (2008).
17) Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. and Maltsev, N.: The use of gene clusters to infer functional coupling, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.96, pp.2896–2901 (1999).
18) Oyama, S. and Manning, C.: Using Feature Conjunctions across Examples for Learning Pairwise Classifiers, *Proceedings of the 15th European Conference on Machine Learning (ECML)*, pp.322–333 (2004).
19) Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. and Yeates, T.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles., *Proceedings of the National Academy of Sciences of the United States of America*, Vol.96, pp.4285–4288 (1999).
20) Smith, T. and Waterman, M.: Identification of common molecular subsequences, *Journal of Molecular Biology*, Vol.147, pp.195–197 (1981).
21) Stuart, J., Segal, E., Koller, D. and Kim, S.: A gene-coexpression network for global discovery of conserved genetic modules, *Science*, Vol.302, No.5643, pp.249–255 (2003).
22) Tamada, Y., Bannai, H., Imoto, S., Katayama, T., Kanehisa, M. and Miyano, S.: Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models, *Journal of Bioinformatics and Computational Biology*, Vol.3, No.6, pp.1295–1313 (2005).
23) Tsuda, K., Shin, H. and Schölkopf, B.: Fast protein classification with multiple

networks, *Bioinformatics*, Vol.21 Suppl. 2 (2005).
24) Vishwanathan, S. V.N., Borgwardt, K. and Schraudolph, N.: Fast computation of graph kernels, *Advances in Neural Information Processing Systems 19* (2007).
25) Walhout, A., Sordella, R., Lu, X., Hartley, J., Temple, G., Brasch, M., Thierry-Mieg, N. and Vidal, M.: Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development, *Science*, Vol.287, pp.116–122 (2000).
26) Weston, J., Elisseeff, A., Zhou, D., Leslie, C. and Noble, W.: Protein ranking: from local to global structure in the protein similarity network, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.101, No.17, pp. 6559–6563 (2004).
27) Yamanishi, Y., Vert, J.-P. and Kanehisa, M.: Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information, *Bioinformatics*, Vol.21, pp.i468–i477 (2005).
28) Yamanishi, Y., Vert, J. and Kanehisa, M.: Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, Vol.20 Suppl 1, pp.i363–370 (2004).
29) Zhou, D., Bousquet, O., Weston, J. and Schölkopf, B.: Learning with local and global consistency, *Advances in Neural Information Processing Systems 16*, pp.321–328 (2004).
30) Zhu, X., Ghahramani, Z. and Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions, *Proceedings of the 20th International Conference on Machine Learning (ICML)* (2003).



(a) 複数生物種ネットワークの個別予測



(b) 複数生物種ネットワークの同時予測

図 1 3 つの生物種（*C. elegans*、*H. pylori*、 *S. cerevisiae*）のネットワークの (a) 個別予測と (b) 同時予測。個別予測が遺伝子発現量の類似度などの種内の情報を用いるのに対し、同時予測では種間をまたいだ配列類似度なども用いる。

Vol.2010-BIO-21 No.19
2010/6/19

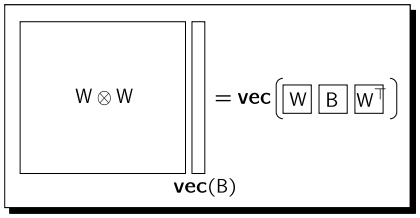
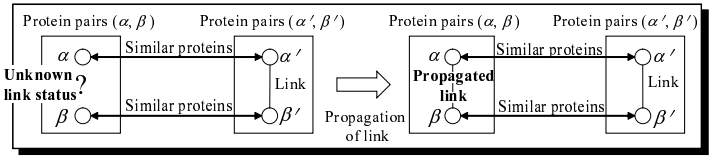
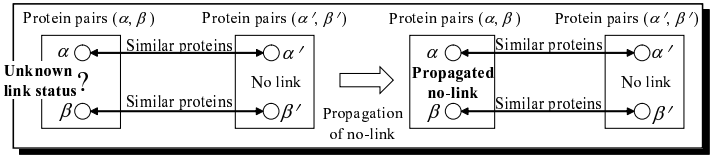


図 2 “vec トリック”[13),24)] を用いることによって 2 つの行列のクロネッカー積とベクトル化された行列との掛け算（左辺）を、行列の掛け算（右辺）に置き換えることができる。これによって、計算量のオーダーが一段下がり、記憶量量のオーダーは元々の大きさの平方根にまで削減される。



(a) リンクがある場合のリンク伝播



(b) リンクが無い場合のリンク伝播

図 3 タンパク質ペア (α, β) と (α', β') に対するリンク伝播の働き。図 (a) 2 つのタンパク質ペアが互いに似ていれば、片方のペア間のリンクが、リンク未知のもう片方のペアに伝播する。図 (b) 同様に「リンクが無い」という状態も伝播する。リンク伝播法はこの原則を全タンパク質ペアに対して同時に適用する。

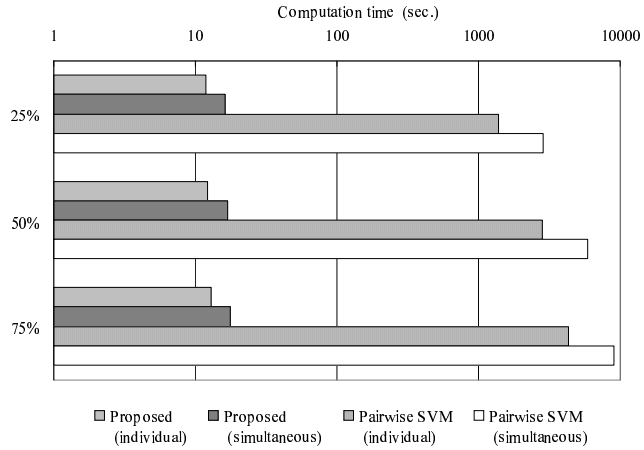


図 4 実行時間の比較。提案手法はペアワイズ SVM（P-SVM）よりも一貫して速いことが分かる。