

複数生物種ネットワークの同時推定 — 半教師つき学習によるアプローチ —

👤 鹿島久嗣 加藤毅（東京大学）
山西芳裕（パリ国立高等鉱業学校）
杉山将（東京工業大学）
津田宏治（産業総合研究所）

👤 本研究は 日本IBM株式会社 東京基礎研究所 のサポートのもとで行われました



複数の生物種のネットワークを、同時に予測する方法を紹介します

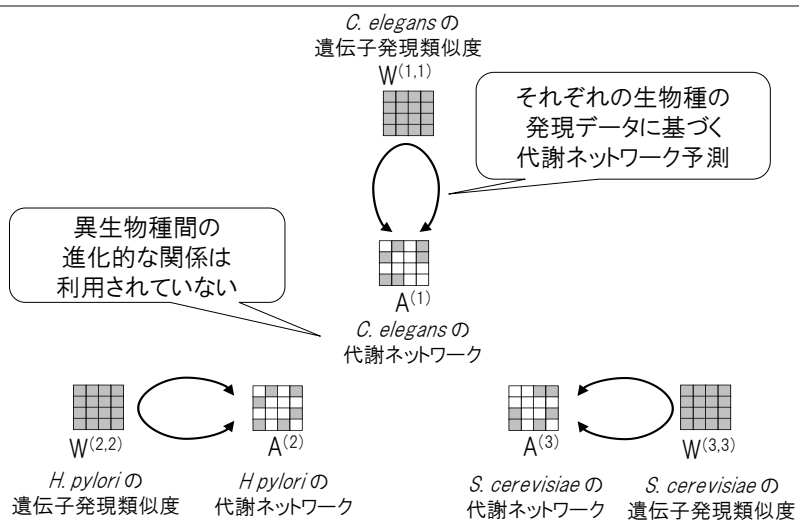
- 複数の生物種のネットワークを、「同時にまとめて」予測することで、予測精度の向上を図る
 - 従来法は、各生物種の生体ネットワークを、それぞれ予測する
- 半教師つき学習によって、予測精度の向上をはかる
 - ネットワークの既知部分をもとに、未知部分を推定
 - 「ラベル伝播法」によって、未知部分の情報も用いる
 - 計算の工夫によって、推定の高速化をはかる
- *S. Cerevisiae*, *H. Pylori*, *C. Elegans* の 3生物種の代謝ネットワークの同時推定に適用し、2つのことを確かめる
 - 個別の生物種を、遺伝子発現データをもとに予測するよりも、異生物種間の配列類似度を用いて、同時に予測する方が精度がよい
 - 同じ情報に基づく、従来法の予測よりも、精度と速度で上回る

モチベーション

3

THE UNIVERSITY OF TOKYO

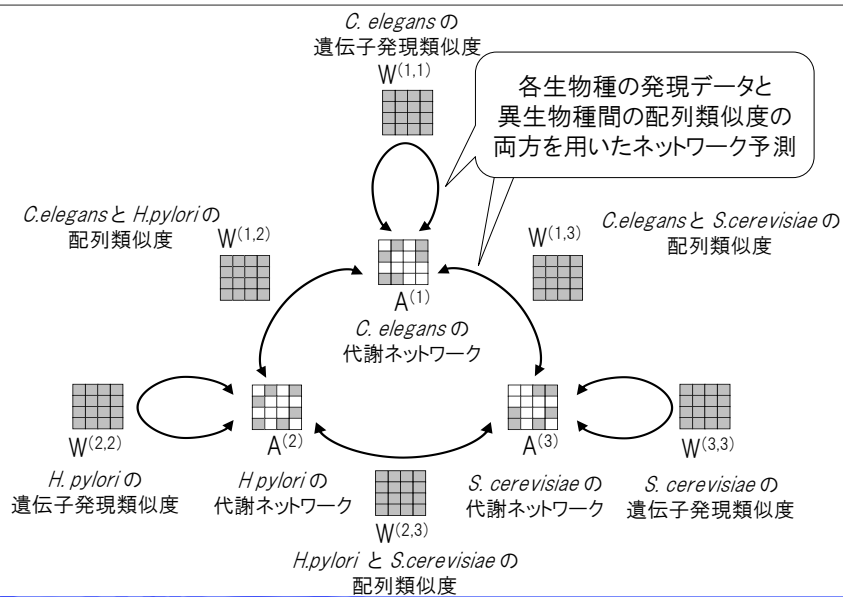
現在ある生体ネットワーク予測は、種ごとに行われるのがふつうです



4

THE UNIVERSITY OF TOKYO

一方、我々は、これらを「同時に」予測することを試みます



5

THE UNIVERSITY OF TOKYO

対象

6

THE UNIVERSITY OF TOKYO

用いるデータは *C.elegans*, *H.pylori*, *S.cerevisiae* の代謝ネットワークと遺伝子発現データとアミノ酸配列データです

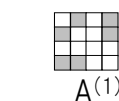
- 代謝ネットワーク: KEGG PATHWAYデータベースから抽出した、3つの生物種 *C. elegans*, *H. pylori*, *S. cerevisiae* の代謝ネットワーク
 - それぞれ、532、291、722個の酵素タンパク質をもつ
- 補助情報として、発現データと配列データが与えられている
 - KEGG GENES データベースから抽出された、各酵素のアミノ酸配列
 - Gene Expression Omnibus [Stuart et al.,2003] から作られた、マイクロアレイデータ
 - ・ それぞれの生物種で各酵素に対し 1209、293、753 の計測値

7

THE UNIVERSITY OF TOKYO

3生物種の代謝ネットワークを、3つのグラフの隣接行列として表現します
これらは部分的にわかっており、未知の部分を推定するのが目的です

- 各生物種の代謝ネットワークを、それぞれひとつのグラフとして表現する
 - グラフの各頂点が、ひとつの酵素を表す
 - ・ それぞれ、532、291、722個の頂点をもつグラフになる
 - 2つの酵素(=頂点)が、代謝経路の中で、連続する反応を修飾するときにリンクをもつ
- グラフ構造は、隣接行列として表現できる
 - 3つの隣接行列として表現できる
- ネットワークには、分かっている部分と、分からない部分がある
 - 隣接行列には、分かっている部分と、分からない部分がある



$A^{(1)}$
C. elegans の
代謝ネットワーク



$A^{(2)}$
H. pylori の
代謝ネットワーク



$A^{(3)}$
S. cerevisiae の
代謝ネットワーク

分かっているところ

分かっていないところ

8

THE UNIVERSITY OF TOKYO

遺伝子発現データと配列データは、酵素間の類似度として用います

- 与えられた発現データと配列データは、全て、酵素の類似度に変換して用いる
- 2つの酵素が、同一生物種のものの場合、2つの酵素の類似度を、発現の類似度とする
 - 発現類似度は、ガウスカーネルによる類似度

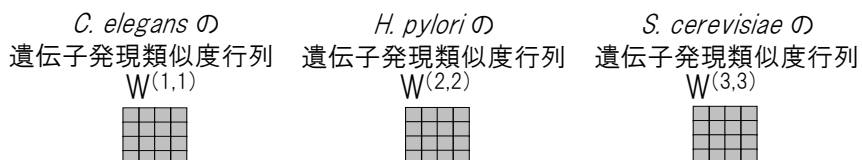
$$k(\mathbf{x}, \mathbf{x}') = \exp \left(- \|\mathbf{x} - \mathbf{x}'\| / \gamma \right)^2$$

・ ただし $\gamma \equiv 2$ とする

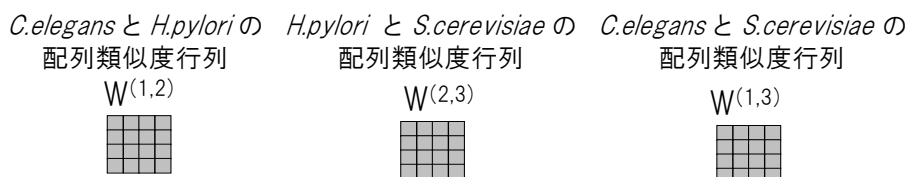
- 2つの酵素が、異生物種のものの場合、2つの酵素の類似度を、配列の類似度とします
 - 配列類似度は $[0,1]$ に正規化した Smith-Waterman スコア

遺伝子発現データと配列データの類似度は、類似度行列として表現します

- 計算した発現データと配列データは、全て、酵素の類似度行列に変換して用いる
- 3生物種の場合、6個の類似度行列ができる
 - 各生物種内の遺伝子発現類似度行列(正方行列)



- 異生物種間の配列類似度行列(矩形行列)



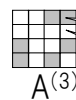
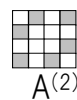
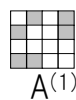
アプローチ

11

THE UNIVERSITY OF TOKYO

補助情報(類似度)を利用した、複数隣接行列の穴埋め問題として定式化します

- 3つの隣接行列の穴埋めを行いたい



分かっているところ

分かっていないところ

- 補助情報として、類似度行列が与えられる

— 類似度行列は全体として対称

・ 対角ブロック

— $W^{(1,2)T} = W^{(2,1)}$

— $W^{(2,2)T} = W^{(2,2)}$

— $W^{(3,3)T} = W^{(3,3)}$

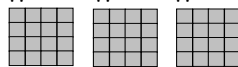
・ 非対角ブロック

— $W^{(1,2)T} = W^{(2,1)}$

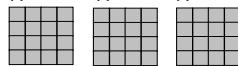
— $W^{(1,3)T} = W^{(3,1)}$

— $W^{(2,3)T} = W^{(3,2)}$

$W^{(1,1)}$ $W^{(1,2)}$ $W^{(1,3)}$



$W^{(2,1)}$ $W^{(2,2)}$ $W^{(2,3)}$



$W^{(3,1)}$ $W^{(3,2)}$ $W^{(3,3)}$



非対角ブロックは
異生物種間の配列類似度

対角ブロックは
各生物種内の発現類似度

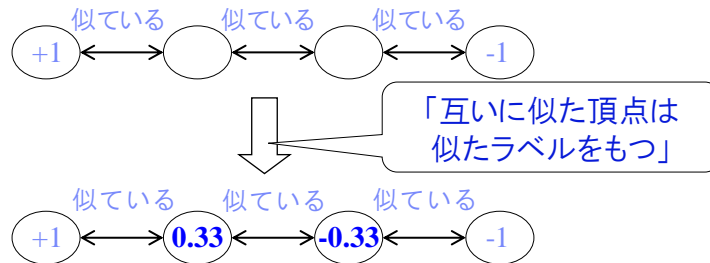
12

THE UNIVERSITY OF TOKYO

どのような推論方式を使うか？

「似た頂点は、似たラベルをもつ」という、ラベル伝播法の考え方を我们用います

- 半教師つき学習の代表的手法「ラベル伝播」を用いる
- 元々は、頂点の分類(リンクの予測ではなく)のための手法
- 「似た頂点は、似たラベルをもつ」の原則を用いる

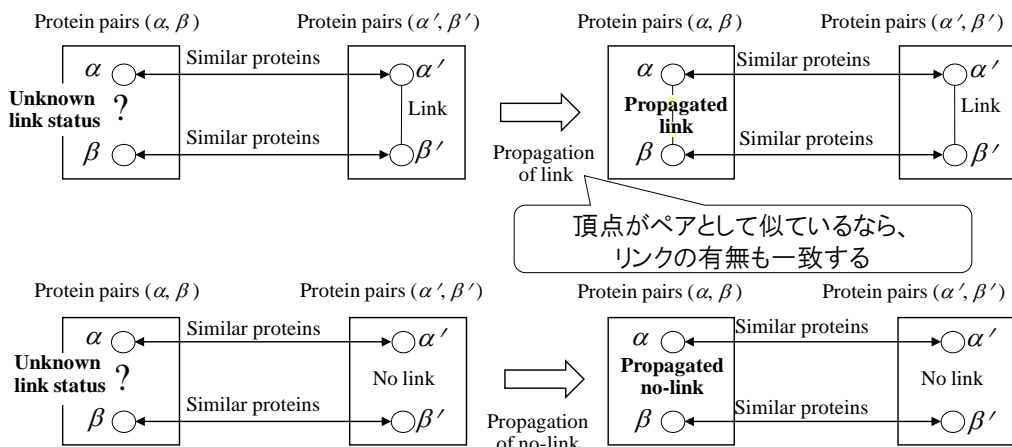


13

THE UNIVERSITY OF TOKYO

ラベル伝播法の考え方を、頂点ペアに対して適用します

- 「似た頂点ペアは、似たリンク(の有無)をもつ」の原則を用いる
- この原則をネットワーク全体で整合性が取れるように適用する



14

THE UNIVERSITY OF TOKYO

結果

15

THE UNIVERSITY OF TOKYO

予測精度の比較実験によって、同時予測の有効性と、従来手法に対する優位性を示します

■ 2つの比較を行う:

- 個別の生物種を遺伝子発現データをもとに予測する

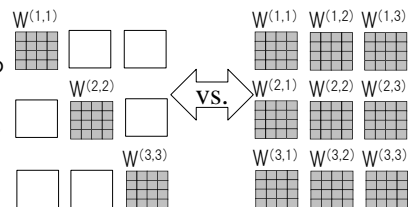
vs.

- 異生物種間の配列類似度を用いて同時に予測する

- 従来法(ペアワイズSVM、カーネル回帰)

vs.

提案法



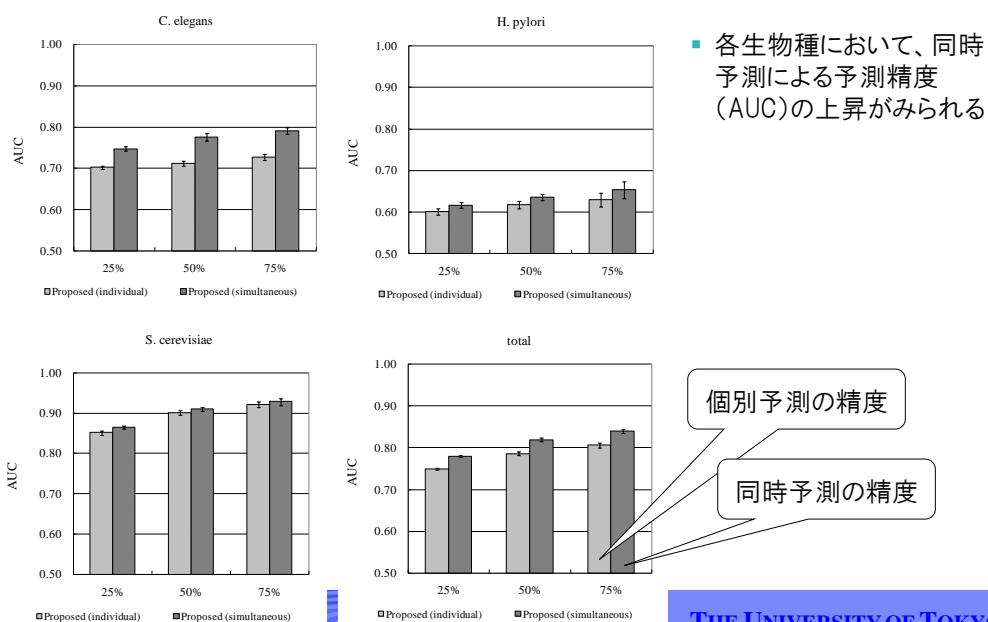
■ 予測精度によって評価する

- KEGGから持ってきたネットワークの構造は正しいものとする
- 隣接行列の要素のうち、25%/50%/75%をランダムに隠し、これを予測
- 予測精度は、以下の2つの指標で測る
 - ・ AUC
 - ・ sensitivity=specificityとなる点

16

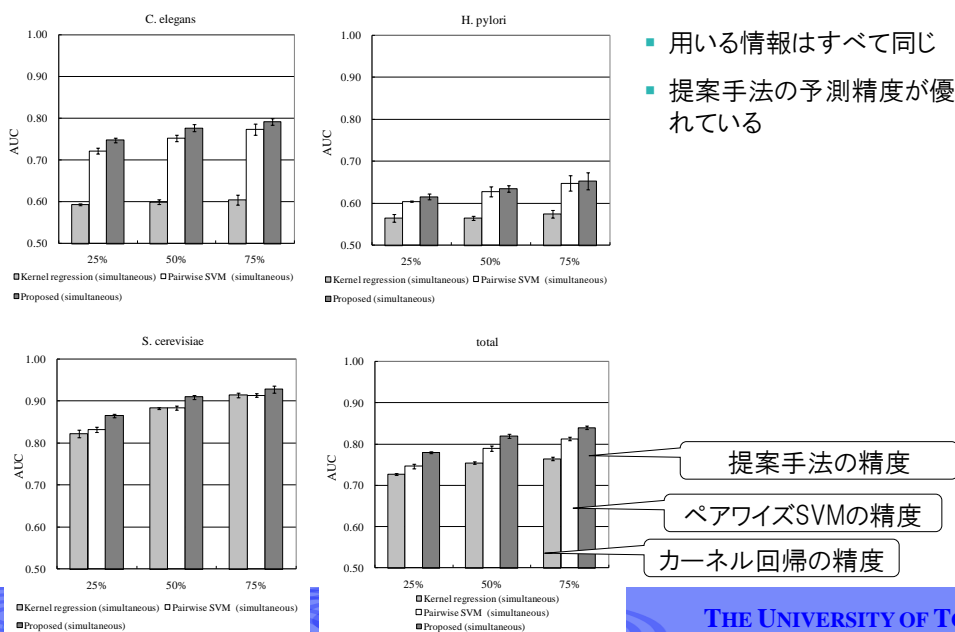
THE UNIVERSITY OF TOKYO

個別の生物種を、遺伝子発現データをもとに予測するよりも、異生物種間の配列類似度を用いて、同時に予測する方が精度がよいことが確認されました



THE UNIVERSITY OF TOKYO

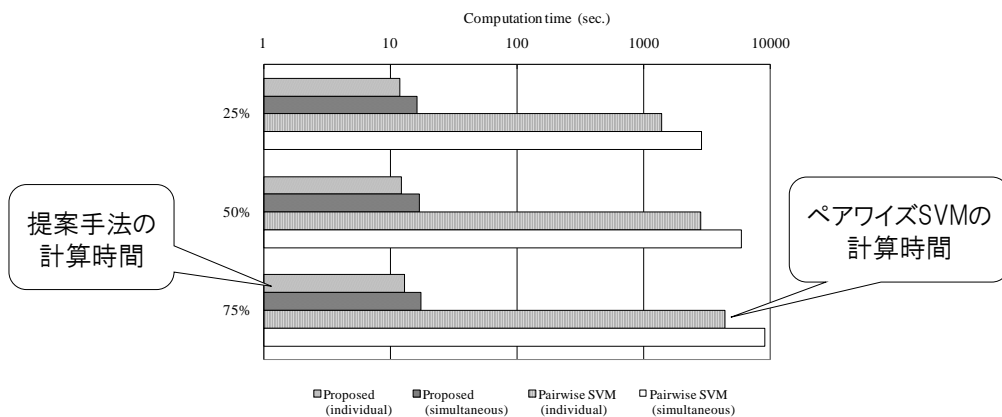
同じ情報に基づく、従来法の予測よりも、精度で上回ることが確認されました



THE UNIVERSITY OF TOKYO

同じ情報に基づく、従来法の予測よりも、速度で上回ることが確認されました

- ペアワイズSVMと比較して
 - ペアワイズSVMはオンライン学習アルゴリズム (passive-aggressive) で実装
 - ・ 全てのデータを3周処理する



19

THE UNIVERSITY OF TOKYO

解法の詳細

20

THE UNIVERSITY OF TOKYO

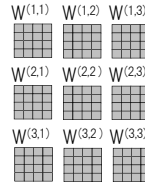
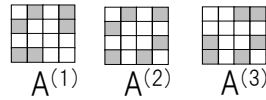
隣接行列と類似度行列から、予測隣接行列を推定します

入力

隣接行列: $A = \{ A^{(1)}, A^{(2)}, A^{(3)} \}$

- ・リンクのある部分は +1
- ・リンクのない部分は -1
- ・わからない部分は 0

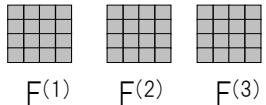
類似度行列: $W^{(1,1)}, \dots, W^{(3,3)}$



出力

予測隣接行列: $F = \{ F^{(1)}, F^{(2)}, F^{(3)} \}$

- ・リンクの確からしさに応じて $[-1, +1]$ の値を割り当てる
 - リンクの有無が分かっている部分については +1 か -1 に近い値が入っているはず



21

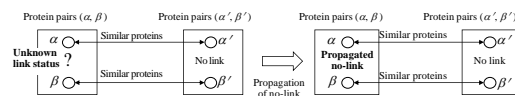
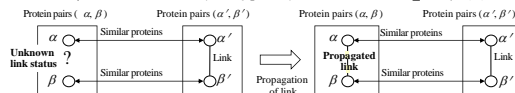
THE UNIVERSITY OF TOKYO

最適化(最小化)問題として定式化します

- 予測隣接行列 F の満たすべき2つの制約を表した、2つの目的関数の線形和として定義する

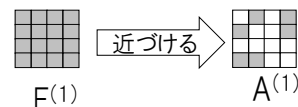
$$J(F) \equiv \sigma J_1(F) + J_2(F)$$

- J_1 : 「似た頂点ペアは、似たリンク(の有無)をもつべし」を表す目的関数と



- J_2 : 「ネットワークの分かっている部分は、予測もそれに近づけるべし」を表す目的関数

- $\sigma > 0$ は2つの項のバランスを取る定数



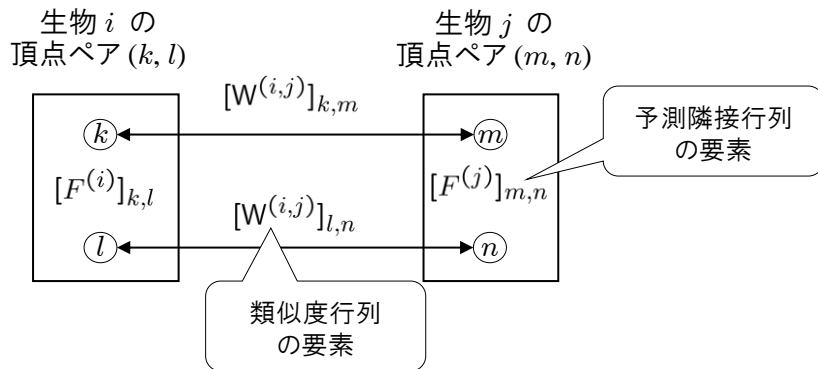
22

THE UNIVERSITY OF TOKYO

「似た頂点ペアは、似たリンク(の有無)をもつべし」ための目的関数

- 生物 i の頂点ペア (k, l) と生物 j の頂点ペア (m, n) は、 k と m 、 l と n がそれぞれ似ている(類似度行列の要素が大きい)ときに、近い予測隣接行列の要素をもつ

$$J_1(F) \equiv \sum [W^{(i,j)}]_{k,m} [W^{(i,j)}]_{l,n} ([F^{(i)}]_{k,l} - [F^{(j)}]_{m,n})^2$$



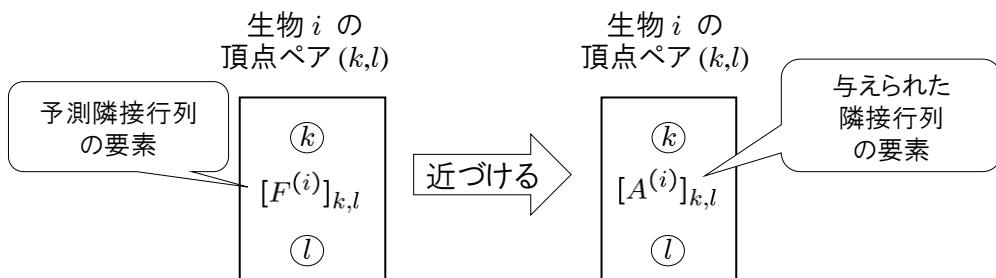
23

THE UNIVERSITY OF TOKYO

「ネットワークの分かっている部分は、予測をそれに近づけるべし」ための目的関数

- 隣接行列の分かっている部分は、予測隣接行列の対応する値を近づける

$$J_2(F) \equiv \sum ([F^{(i)}]_{k,l} - [A^{(i)}]_{k,l})^2$$



24

THE UNIVERSITY OF TOKYO

目的関数は行列を用いてシンプルに書き換えることができます

- 全体の目的関数 $J(F) = \sigma J_1(F) + J_2(F)$

- 1項目は、グラフラプリアン

$$L^{(i,j)} \equiv D^{(i,j)} \otimes D^{(i,j)} - W^{(i,j)} \otimes W^{(i,j)}$$

を用いて

$$\begin{aligned} J_1(F) &\equiv \sum_{(i,j,k,l,m)} [W^{(i,j)}]_{k,m} [W^{(i,j)}]_{l,n} ([F^{(i)}]_{k,l} - [F^{(j)}]_{m,n})^2 \\ &= \sum_{(i,j)} \text{vec}(F^{(i)})^\top L^{(i,j)} \text{vec}(F^{(j)}) \end{aligned}$$

- 2項目は、

$$\begin{aligned} J_2(F) &\equiv \sum_{(i,k,l)} ([F^{(i)}]_{k,l} - [A^{(i)}]_{k,l})^2 \\ &= \sum_i \|F^{(i)} - A^{(i)}\|^2 \end{aligned}$$

$$\text{vec} \left(\begin{bmatrix} \downarrow \uparrow \\ \downarrow \uparrow \end{bmatrix} \right) = \begin{pmatrix} \downarrow \\ \uparrow \\ \downarrow \\ \uparrow \end{pmatrix}$$

25

THE UNIVERSITY OF TOKYO

予測隣接行列は連立方程式を解くことで得られます

- 最終的な目的関数

$$J(F) = \sigma \sum_{(i,j)} \text{vec}(F^{(i)})^\top L^{(i,j)} \text{vec}(F^{(j)}) + \sum_i \|F^{(i)} - A^{(i)}\|^2$$

- 目的関数を最小化する解は、以下の(大きな)連立方程式を解くことで得られる

$$\sum_j (\sigma L^{(i,j)} + \delta(i=j)I) \text{vec}(F^{(j)}) = \text{vec}(A)^{(i)}$$

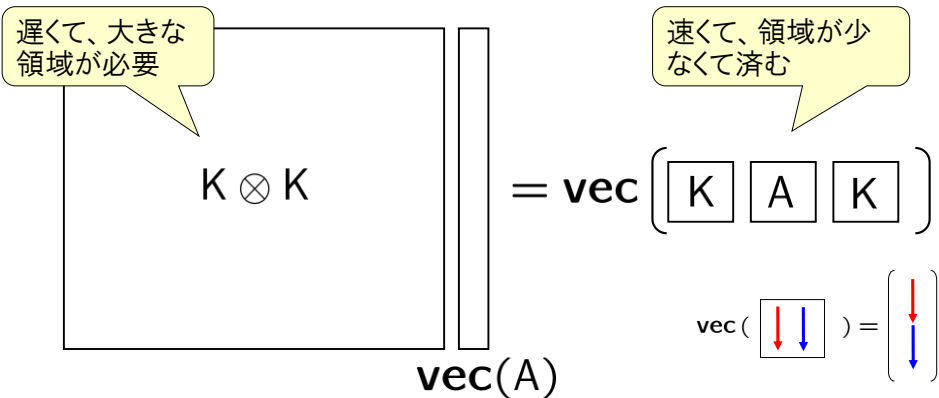
- 連立方程式は、たとえば、共役勾配法などによって解く

26

THE UNIVERSITY OF TOKYO

計算のボトルネックがありますが、これは大幅に効率化・省スペース化できます

- 行列のクロネッカー積と、ベクトル化した行列の積の掛け算が計算のボトルネックになる
- 「行列のクロネッカー積」と「ベクトル化された行列」の積についての公式が有用
- 左辺より、右辺がずっと効率的に計算できる


$$K \otimes K \quad \text{vec}(A) = \text{vec} \left(\begin{bmatrix} K & A & K \end{bmatrix} \right)$$
$$\text{vec} \left(\begin{bmatrix} \downarrow & \downarrow \\ \downarrow & \downarrow \end{bmatrix} \right) = \begin{bmatrix} \downarrow \\ \downarrow \end{bmatrix}$$

27

THE UNIVERSITY OF TOKYO

結論

28

THE UNIVERSITY OF TOKYO

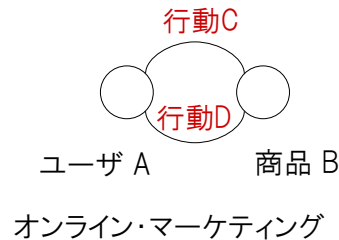
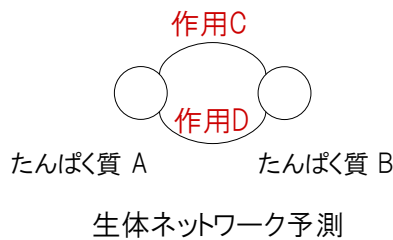
複数の生物種のネットワークを、同時に予測する方法を紹介しました

- 複数の生物種のネットワークを、「同時にまとめて」予測することで、予測精度を向上しました
 - 従来法は、各生物種の生体ネットワークを、それぞれ予測する
- 半教師つき学習によって、予測精度を向上しました
 - ネットワークの既知部分をもとに、未知部分を推定
 - 「ラベル伝播法」によって、未知部分の情報も用いる
 - 計算の工夫によって、推定の高速化をはかる
- *S. Cerevisiae*, *H. Pylori*, *C. Elegans* の 3 生物種の代謝ネットワークの同時推定に適用し、2つのことを確かめました
 - 個別の生物種を、遺伝子発現データをもとに予測するよりも、異生物種間の配列類似度を用いて、同時に予測する方が精度がよい
 - 同じ情報に基づく、従来法の予測よりも、精度と速度で上回る

展望

紹介した手法は、さらに「複数タイプのリンク」を扱えるように拡張できます

- 「リンクの種類」を考えることで、適用対象が広がる
 - 生体ネットワーク分析: 作用の種類や、作用が起こる環境、など
 - オンライン・マーケティング: 購買、評価、商品情報の閲覧、など
 - 時間もリンクの種類として考えることができる
- 異なるタイプのリンク間の相関を利用することで、予測精度があがる可能性がある



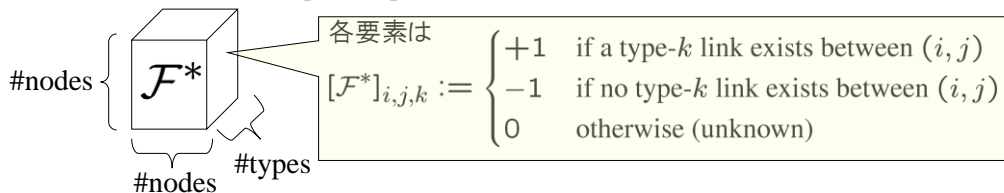
※ Hisashi Kashima, Tsuyoshi Kato, Yoshihiro Yamanishi, Masashi Sugiyama and Koji Tsuda:
 Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction
 In Proc. SIAM Data Mining Conference, 2009.

31

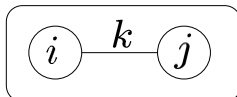
THE UNIVERSITY OF TOKYO

複数タイプリンク予測問題は、3階のテンソルの補完問題として捉えられます

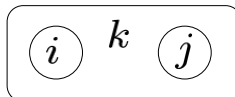
- ネットワーク構造の既知部分が、3階のテンソル \mathcal{F}^* で与えられる
 - 既知部分は +1 か -1 で、未知部分は 0
- 目標: 未知部分 (0) を $[-1, +1]$ の値で(リンクの確信度に応じて)埋める



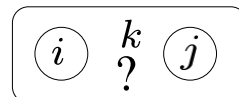
$$[\mathcal{F}^*]_{i,j,k} = +1$$



$$[\mathcal{F}^*]_{i,j,k} = -1$$



$$[\mathcal{F}^*]_{i,j,k} = 0$$



※ Hisashi Kashima, Tsuyoshi Kato, Yoshihiro Yamanishi, Masashi Sugiyama and Koji Tsuda:
 Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction
 In Proc. SIAM Data Mining Conference, 2009.

32

THE UNIVERSITY OF TOKYO

ありがとうございました