Vol.2012-BIO-28 No.15
2012/3/29

経路ラベル決定法に基づくネットワーク特徴量の拡張とそれらを用いた辺ラベル付き有向グラフ間類似度の提案

安富祖　仁[†1]　岡崎　威生[†2]　名嘉村　盛和[†2]

社会ネットワーク分析の分野において用いられる、クラスタリング係数や媒介中心性といったネットワーク特徴量を利用することによって、有向グラフ間の類似度を定義することができる。しかし従来のネットワーク特徴量は、遺伝子制御関係ネットワークのような、辺に対してラベルが与えられている有向グラフを扱うことができない。本研究では、従来のネットワーク特徴量を経路ラベル決定法に基づいて辺ラベルをもつ有向グラフへ適用可能なものへと拡張した。また、それらのネットワーク特徴量を用いた辺ラベル付き有向グラフ間類似度の提案を試みた。

Extensions of Network Characteristics based on Path Labeling aims to Similarity for Edge Labeled Directed Graphs

Hitoshi AFUSO ,[†1] Takeo OKAZAKI [†2]
and Morikazu NAKAMURA[†2]

Using network characteristics such as clustering coefficient and betweenness centrality, we can define the network similarity. In biological area, such similarity may reveal the evolutional associations among various species. However, because the traditional network characteristics cannot handle the edge-labeled network such as gene regulartory networks, it is difficult to define the network similarity using those network characteristics. To solve that difficulty, we extended the traditional network characteristics based on path labeling. Using the extended characteristics we also tried to propose network similarity.

†1 琉球大学大学院理工学研究科
†2 琉球大学工学部情報工学科

1. 背　　景

DNA マイクロアレイ[1] やフェノタイプマイクロアレイ[2] といった、多数の遺伝子の発現を観測する技術の発展と、それらから得られる多様な遺伝子発現プロファイルデータを解析する手法の開発[3] によって、様々な生物種の遺伝子制御ネットワークやタンパク質相互作用ネットワークといった生物ネットワークが明らかにされつつある[4]。

複数の生物種から得られた生物ネットワークを比較することで、生物種間の進化的な関連やネットワークが実現している生物的な機能の類似性が明らかになるのではないかと期待されている。生物ネットワークを比較するための手法として、これまでに Natasa *et al*[5] によるネットワークの部分構造に基づく比較手法や、寺田ら[6] や Oleksii *et al*[7] によるネットワークアラインメントに基づく比較手法が提案されてきた。Natasa *et al*[5] の手法では、与えられた無向グラフにおいて graphlet と呼ばれる頂点数が 5 以下の同型でない部分グラフを列挙し、各部分グラフの出現頻度を比較することによって無向グラフの類似度を定義した。しかしこの方法では、頂点数が 5 個以下の部分グラフのみを考慮しているため、構成された類似度が無向グラフの大域的な構造を反映しているとは考えにくい。寺田ら[6] は、局所的な構造を考慮して与えられた無向グラフを、概要グラフと呼ばれる簡約化されたグラフへと変換した後、概要グラフ内の頂点をマッチングすることにより、無向グラフの大域的な構造を反映した類似度を定義した。また、Oleksii *et al*[7] は与えられた無向グラフの頂点間のマッチングを行う際に、遺伝子のシーケンス情報といった頂点間の類似性を利用する類似度を提案し、提案した類似度を用いてヒトと酵母菌のタンパク質相互作用ネットワークを比較した。しかし、これらの手法は無向グラフを対象とした類似度であり、遺伝子発現の因果関係のような、関係に向きが存在するネットワークの類似性を測ることは難しい。また、Oleksii *et al* らの用いた頂点間の類似性は生物学的なものに限られていた。そこで安富祖ら[8] は、社会ネットワーク分析の分野で用いられるクラスタリング係数[9] や中心性[10] といったネットワーク特徴量を利用して、有向グラフの類似性を測る手法を提案した。しかし、安富祖らの使用したネットワーク特徴量は辺にラベルが付置された有向グラフ (辺ラベル付き有向グラフ) を扱うことができないため、遺伝子制御ネットワークのように発現の促進や抑制といった、辺のラベルを考慮する必要がある有向グラフの類似性を測ることが困難となる。

本研究ではクラスタリング係数などのネットワーク特徴量が、特定の種類の経路数を数え上げているという事実に着目し、各辺に付置されたラベルから経路のラベルを決定する方法

を提案することにより、各ネットワーク特徴量を辺ラベル付き有向グラフへと適用可能なものへと拡張した。加えて、それらのネットワーク特徴量を用いて、辺ラベル付き有向グラフ間の類似度の提案を試みた。

2. 対象グラフ

本研究では、遺伝子制御ネットワークの単純なモデルを対象として、類似度を提案する。本研究で対象とするグラフ G は頂点集合 V、有向辺集合 E と辺ラベル付置関数 L の組 $G(V, E, L)$ で表現される。頂点集合 V の各要素 v_i は遺伝子に対応し、有向辺集合 E_j の各要素 e は遺伝子間の制御関係に対応している。ここで辺ラベル付置関数 $L : E \to \{1, -1\}$ は次のような関数である。

$$L(e) = \begin{cases} 1 & (\text{辺 e が発現を促進するという関係を表現している場合}) \\ -1 & (\text{辺 e が発現を抑制するという関係を表現している場合}) \end{cases} \quad (1)$$

本研究では、以上のような辺ラベル付き有向グラフに対して類似度を提案することを目的とする。

3. 従来のネットワーク特徴量

本研究で拡張したネットワーク特徴量である、クラスタリング係数、近接中心性、離心中心性、媒介中心性、PageRank について概説する。

クラスタリング係数[9] は、ある頂点に隣接している頂点群が互いに隣接しているかどうかを意味する。クラスタリング係数は元来、無向グラフを対象としたネットワーク特徴量であったが、鈴木[11] によって有向グラフを扱えるものへと拡張された。有向グラフに対するクラスタリング係数は、ある頂点の親集合に属する頂点間の隣接関係に着目するものと、子集合に属する頂点間の隣接関係に着目するものの 2 種類がある。親集合に着目したクラスタリング係数の計算は、ある頂点を終端とし、その頂点の親集合に属する頂点を始端とする長さ 2 の経路数に着目している。

社会ネットワークにおける中心性とは、各頂点すなわちネットワークによって表現される社会的な現象に参加している対象が、そのネットワークにおいて中心的な役割をもっているかどうかを調べるための指標である。どのような性質を中心的とみなすかによって、複数の中心性が定義される。近接中心性では、他の頂点から到達するためのコストが低い頂点が、そのネットワーク内において中心的な役割をもつとみなす。近接中心性は、ある頂点へ他の頂点から到達する最短経路の中で、最長の経路長として定義される。つまり近接中心性では、ある頂点に到達する最短経路の集合に着目している。一方で、他の頂点へ到達するためのコストが低い頂点が、そのネットワーク内において中心的な役割をもつとみなすこともできる。このような視点で定義される特徴量が離心中心性である。離心中心性は、ある頂点から他の頂点へ到達する最短経路の中で最長の経路長と定義される。つまり離心中心性は、ある頂点を始端とする最短経路の集合に着目している。媒介中心性は、ある頂点が他の頂点対を接続している度合いを示している。つまり、ネットワーク内における 2 つの頂点を辺の向きに沿って結ぶときに、より多くの経路に含まれる頂点がそのネットワークの中心部に位置しているとみなす。媒介中心性は、着目する頂点が、他のある 2 つの頂点を結ぶ最短経路に含まれる数によって定義される。つまり離心中心性は、ネットワーク内の任意の 2 頂点間の最短経路の集合に着目している。以上のことから本研究で拡張した 3 つの中心性は、最短経路の集合に着目したものであることがわかる。

PageRank[12] は Web ページの重要度を決定するために Page *et al* によって提案されたネットワーク特徴量である。PageRank では、重要な Web ページからリンクされている Web ページもまた重要であるという、再帰的なアイディアを用いて各 Web ページの重要度を算出する。PageRank はリンクに沿って Web ページ間をランダムに移動するユーザが、十分に長い時間 Web 内を巡回したときに各ページへと到達した回数の割合として捉えられる。つまり PageRank の計算では、各頂点をランダムに決定された経路が通過した回数をカウントしている。

クラスタリング係数、各種中心性、PageRank はそれぞれ、始端と終端を固定した長さ 2 の経路集合、各頂点間の最短経路集合、ランダムに決定された経路集合を考慮してその値を決定していることがわかる。このことから、ある経路に対してそれに含まれる辺のラベルを考慮して経路ラベルを決定することによって、各ネットワーク特徴量を辺ラベル付き有向グラフに適用可能なものへと拡張できると考えられる。

4. 経路ラベル決定法

各ネットワーク特徴量を辺ラベル付き有向グラフに対して適用可能なものへと拡張するための、経路に含まれる辺のラベルに基づいた経路ラベルの決定法について述べる。本研究では、経路に含まれる辺のラベルに基づいて経路のラベルを決定する方法として、辺ラベルの積による方法、影響伝搬の終端を考慮した方法、辺ラベルの多数決に基づく方法の 3 つを提案した。本研究では、グラフ G 内の長さ n の経路を有向辺 e_k の順列 $(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$

Vol.2012-BIO-28 No.15
2012/3/29

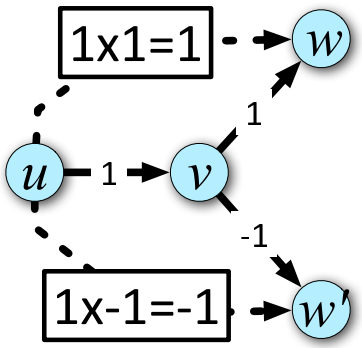

図 1 辺ラベルの積による経路ラベル決定法

として表現する。このとき、辺 e_{k_i} の終端となる頂点と辺 $e_{k_{i+1}}$ の始端となる頂点は一致するものとする。

4.1 辺ラベルの積による経路ラベル決定法

本研究では辺ラベルとして、ある遺伝子が他の遺伝子の発現を促進するならば 1、抑制するならば -1 という二値を使用している。ここで、ある遺伝子 v の発現を促進している遺伝子 u を考える。もしも被制御遺伝子 v が他の遺伝子 w の発現を促進しているならば、遺伝子 v を促進している遺伝子 u は、間接的に遺伝子 w の発現を促進していると考えることができる (図 1 上側の破線)。また逆に、被制御遺伝子 v が他の遺伝子 w' の発現を抑制しているならば、遺伝子 u は間接的に遺伝子 w' の発現を抑制していると考えることもできる (図 1 下側の破線)。つまり、ある遺伝子 v を介した遺伝子 u の遺伝子 w に対する間接的な制御関係は、辺ラベルの積として表現することができる。以上の考察から、ある経路に対してその始端側の辺から終端の辺までラベルの値を逐次掛け合わせた結果を対応する経路のラベルとする方法を提案した。この方法では、ある経路 $(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ の経路ラベル $l(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ は次のように計算される。

$$l(e_{k_1}, e_{k_2}, \cdots, e_{k_n}) = \prod_{i}^{n} L(e_{k_i}) \tag{2}$$

ここで関数 $L(e_{k_i})$ は式 (1) で定義された辺ラベル付置関数である。

4.2 影響伝搬の終端を考慮した経路ラベル決定法

前項で定義した経路ラベル決定法では、経路が長くなった場合に積極的に間接的な影響を与えているとは考えにくい場合にも、促進あるいは抑制の影響を与えているとしてしまう可能性がある。そのような場合の例を図 2 に示す。図 2 において、遺伝子 u から遺伝子 w への間接的な影響を考えた場合、遺伝子 u は遺伝子 w の発現を抑制する遺伝子 v を促進していることから、間接的に遺伝子 w を抑制していると考えられる。しかし遺伝子 v から遺伝子 k への間接的な影響のように、ある遺伝子 k の発現を抑制している遺伝子 w を抑制する遺伝子 v に関しては、間接的に発現を促進していると積極的には言わずに、経路は存在するが影響は与えていないとする立場も考えられる。同様に図 2 中の遺伝子 u から遺伝子 k に対する間接的な影響も、遺伝子間に経路は存在するが影響は伝搬していないと捉えることもできる。以上の考察から、ある経路内において抑制のラベル -1 が出現した以降は、影響が伝搬していないとする経路ラベル決定法を提案した。この方法では、ある経路 $(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ の経路ラベル $l(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ は次のように決定される。

$$l(e_{k_1}, e_{k_2}, \cdots, e_{k_n}) = \begin{cases} 1 & \text{(経路に含まれる全ての辺のラベルが 1 である場合)} \\ -1 & \text{(経路の終端の辺ラベルが } -1 \text{ である場合)} \\ 0 & \text{(経路の終端以外に辺ラベル } -1 \text{ がある場合)} \end{cases} \tag{3}$$

この経路ラベル決定法において、経路ラベル 0 は遺伝子間に経路は存在するがそれらの間に影響の伝搬は存在しないということを示している。経路ラベル 0 を空ラベルと呼ぶ。

4.3 辺ラベルの多数決に基づく経路ラベル決定法

上記の 2 つの経路ラベル決定法では、与えられた経路のラベルを決定する際に、経路の始端から順次辺ラベルを評価していた。また前述の抑制ラベル -1 を影響伝搬の終端とする方法は、抑制ラベルの出現後は経路が継続したとしても影響が伝搬しないとする方法であるため、多数の空ラベルをもつ経路が出現し、その結果、頂点間に存在する経路の情報が失われてしまう可能性がある。そこで、経路内における辺ラベルの出現順序は考慮せず、各辺ラベルの出現頻度に基づいて経路ラベルを決定する方法を提案した。この方法では、発現促進のラベル 1 をもつ辺を p 本含む長さ n の経路 $(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ の経路ラベル $l(e_{k_1}, e_{k_2}, \cdots, e_{k_n})$ は次のように決定される。

Vol.2012-BIO-28 No.15
2012/3/29

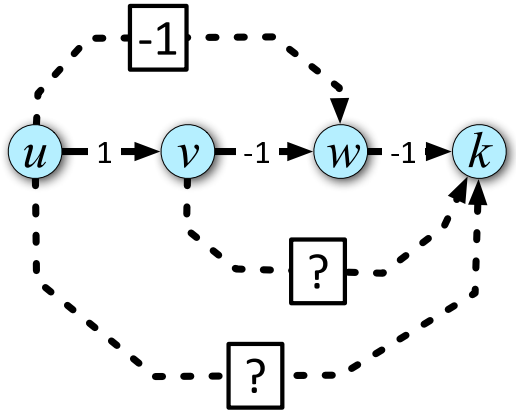

図 2 影響が伝搬していると積極的にいいにくい場合の例

$$l(e_{k_1}, e_{k_2}, \cdots, e_{k_n}) = \begin{cases} 1 & 2p - n > d \\ 0 & |2p - n| < d \\ -1 & n - 2p > d \end{cases} \quad (4)$$

ここで、d は促進ラベル 1 と抑制ラベル -1 の差をコントロールするためのパラメータであり、促進ラベル 1 の数と抑制ラベル -1 の数がある範囲内である場合には、実質的な差がないものとみなし、空ラベル 0 を与えるということを意味している。

以上の 3 つの経路ラベル決定法に基づいて従来のネットワーク特徴量を辺ラベル付き有向グラフが扱えるものへと拡張することができる。

5. 経路ラベル決定法に基づくネットワーク特徴量の拡張

前節で提案した経路ラベル決定法を利用して、従来のネットワーク特徴量の拡張を行う。

5.1 クラスタリング係数の拡張

クラスタリング係数は元来、無向グラフで表現される関係の推移性に着目した特徴量である。しかし、関係の推移性に着目するという観点では、経路ラベルを考慮することができないため、辺ラベル付き有向グラフが扱えるようにクラスタリング係数を拡張することが難しい。そこで、クラスタリング係数を辺ラベル付き有向グラフに対して適用可能なものへと拡張するために、経路ラベルと辺ラベルの一致性に着目する。

従来のクラスタリング係数では、始端と終端との間に直接有向辺が存在するような長さが 2 の経路をカウントしてその値を決定する。同様に拡張したクラスタリング係数では、始端と終端との間の直接的な有向辺の辺ラベルと別の頂点を迂回する長さが 2 の経路の経路ラベルを比較し、それらが一致する回数と一致しない回数をカウントすることによって、その値を決定することとした。拡張したクラスタリング係数は、辺ラベルと経路ラベルの一致性に着目した特徴量であるため、ラベル一致係数と呼ぶこととした。ラベル一致係数は従来の有向グラフ用のクラスタリング係数と同様に、親頂点集合に着目するものと子頂点集合に着目する二種類が定義できる。

ある辺ラベル付き有向グラフ G 内の頂点 v のラベル一致係数 $LabelConsistency_G(v)$ を、辺ラベルと経路ラベルが一致する回数 $match$ と一致しない回数 $mismatch$ の 2 つの値を用いて次のような式で定義した。

$$LabelConsistency_G(g) = \begin{pmatrix} \frac{match}{|P_v|(|P_v|-1)} \\ \frac{mismatch}{|P_v|(|P_v|-1)} \end{pmatrix} \quad (5)$$

ここで P_v は頂点 v の親集合、または子集合を表す。ラベル一致係数では、長さが 2 の経路に着目するため、辺ラベルの多数決による経路ラベルの決定法を用いた拡張は行えない。そこでラベル一致係数の計算に必要な $match$ 値と $mismatch$ 値は積による経路ラベル決定法と影響伝搬を考慮した経路ラベル決定法のいずれかで計算することとした。

5.2 各中心性の拡張

近接中心性、離心中心性、媒介中心性の 3 つの中心性は、それぞれ各頂点間の最短経路に着目して計算される特徴量である。そこで、各中心性の計算に使用する最短経路の数え上げを、経路ラベルが促進ラベル 1 になる場合と抑制ラベル -1 になる場合の 2 通りの場合にわけてカウントすることで、辺ラベル付き有向グラフに対しても適用可能な中心性へと拡張できる。従来の 3 つの中心性それぞれに対して、促進ラベル 1 をもつ経路に着目する場合と抑制ラベル -1 をもつ経路に着目する場合の 2 つの値が定義される。例えば、辺ラベル付き有向グラフ G の頂点 v に対する拡張された近接中心性 $ExtendedCloseness_G(v)$ は次のように定義される。

$$ExtendedCloseness_G(v) = \begin{pmatrix} \max_{u \in G} d_{\mathbf{positive}}(u, v) \\ \max_{u \in G} d_{\mathbf{negative}}(u, v) \end{pmatrix} \quad (6)$$

ここで $d_{\mathbf{positive}}(u, v)$ と $d_{\mathbf{negative}}(u, v)$ はそれぞれ、頂点 u と v の間の促進ラベル 1 をも

つ最短経路の長さと抑制ラベル −1 をもつ最短経路の長さを表す。拡張された離心中心性、媒介中心性も同様に、それぞれ促進ラベル 1 をもつ最短経路に着目した場合と抑制ラベル −1 をもつ最短経路に着目した場合の 2 つの値をもつ。

5.3 PageRank の拡張

PageRank は Random surfer model に従うユーザが Web ページ間をランダムに巡回することを十分長い時間行ったときの、各 Web ページにユーザが訪れた回数によって決定される。ここで、従来の PageRank のユーザを遺伝子間の影響、Web ページ間のランダムな移動経路を頂点間のランダムウォークと捉え直すことによって、従来の PageRank を辺ラベル付き有向グラフを扱うことができるものへと拡張できる。具体的には、ランダムに影響の初期位置を決定してそこから辺に沿って十分長いステップ数で頂点間の巡回を行いながら、各ステップでの経路ラベルを決定し、各頂点に促進ラベル 1 の経路が通過した回数と抑制ラベル −1 の経路が通過した回数をカウントすることにより、拡張した PageRank を定義する。

Page *et al* らは辺を全く無視したジャンプを考慮することによって、有向グラフの強連結性を保証し、それによってユーザの各 Web ページへの巡回回数の割合が初期位置に依存しない分布に収束するということを示した。また初期位置に依存しない分布が遷移確率行列の第一固有ベクトルに一致することを示し、PageRank の計算を固有値問題へと帰着させることで各 Web ページの PageRank の計算を効率よく行う方法を提案した。しかし、辺ラベル付き有向グラフにおいては影響の伝搬元、すなわち影響の初期位置に依存して各頂点への経路ラベルが決定されるため、固有値問題へと計算を帰着することは難しい。そこで、PageRank がランダムに移動するユーザが十分長い時間、巡回を行った後の各 Web ページへの巡回割合と捉えられることを利用し、モンテカルロ法[13] によって拡張した PageRank を計算することとした。モンテカルロ法に基づいた拡張された PageRank の計算ステップを以下に示す。

(1) 影響伝搬の初期位置となる頂点をランダムに選択する。
(2) 現在の頂点の子集合から次に影響を伝搬させる頂点を選択する。
(3) 初期位置となる頂点から選択された頂点への経路ラベルを決定する。
(4) 指定されたステップ数になるまで (2) から (3) を繰り返す。
(5) 指定された試行数だけ (1) から (3) を繰り返す。
(6) 各頂点に対して促進ラベル 1 の経路が通過した回数と促進ラベル 1 をもつ経路総数との割合を求める。
(7) 同様に各頂点に対して抑制ラベル −1 がの経路が通過した回数と抑制ラベル −1 をもつ経路総数との割合を求める。

以上のステップにより、辺ラベル付き有向グラフへと拡張した PageRank を計算することができる。拡張 PageRank の値は各頂点について 2 つずつの値をもつ。

6. ネットワーク特徴量を利用したグラフの数ベクトル表現

前節までに、社会ネットワーク分析などの分野において用いられるネットワーク特徴量である、クラスタリング係数、近接中心性、離心中心性、媒介中心性、PageRank を拡張し、辺ラベル付き有向グラフへと適用可能なネットワーク特徴量を提案した。新たなネットワーク特徴量は親頂点集合に着目したラベル一致係数、子頂点集合に着目したラベル一致係数、拡張近接中心性、拡張離心中心性、拡張媒介中心性、拡張 PageRank の合計 6 つで、それぞれの特徴量が 2 次元のベクトルとして表現されている。これらのネットワーク特徴量を用いることにより、辺ラベル付き有向グラフ G 内の各頂点 v を 12 次元のベクトルとして表現できる。本研究では、頂点数が N である辺ラベル付き有向グラフ G を、N 本の 12 次元ベクトルを縦に並べた $12 \times N$ 次元ベクトルで表現し、そのベクトル間の類似度を定義することによって対応する辺ラベル付き有向グラフ間の類似度の定義を試みた。

$$v \rightarrow \begin{pmatrix} ParentsLabelConsistency_G^+(v) \\ ParentsLabelConsistency_G^-(v) \\ \vdots \\ ExtendedPageRank^+(v) \\ ExtendedPageRank^-(v) \end{pmatrix} \tag{7}$$

$$G \rightarrow \begin{pmatrix} ParentsLabelConsistency_G^+(v_1) \\ ParentsLabelConsistency_G^-(v_1) \\ \vdots \\ ExtendedPageRank^+(v_N) \\ ExtendedPageRank^-(v_N) \end{pmatrix} \tag{8}$$

ここで $ParentsLabelConsistency_G^+(v)$ と $ParentsLabelConsistency_G{}^-(v)$ はそれぞれ、頂点 v の親頂点集合に着目したラベル一致係数の第 1 次元、ラベル一致係数の第 2 次元の値を表す。$ExtendedPageRank^+(v)$ と $ExtendedPageRank^-(v)$ も同様である。辺ラベ

Vol.2012-BIO-28 No.15
2012/3/29

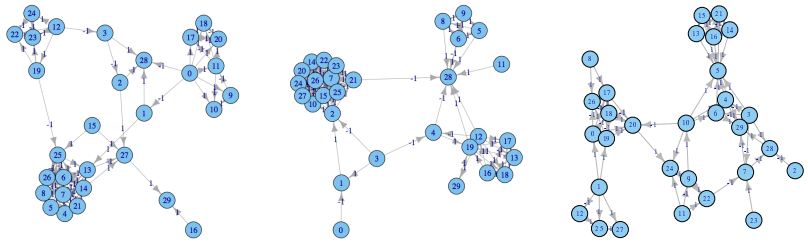

図 3 実験に使用した辺ラベル付き有向グラフの例

ル付き有向グラフの数ベクトル表現 (8) を利用して類似度の定義を試みる。

7. グラフに対するベクトル表現の妥当性検証実験

辺ラベル付き有向グラフの数ベクトル表現 (8) を利用してグラフ間の類似度を定義するために、数ベクトル表現によって辺ラベル付き有向グラフの類似性が測れるかどうかの検証を行った。検証を行うために辺ラベル付き有向グラフを複数用意し、それらのグラフを数ベクトル表現して、その空間内での点の分布を調べた。空間内での点の分布に偏りがあれば、数ベクトル表現によって用意した辺ラベル付き有向グラフにおける何らかの構造類似性を捉えることができていると考えられるため、提案した数ベクトル表現によって辺ラベル付き有向グラフの類似度を定義できると考えられる。

検証に用いる辺ラベル付き有向グラフ群を、酵母菌の遺伝子制御ネットワークを収録した YeastNet2[4] から、Modularity[14] というネットワーク特徴量に基づいてサブネットワークを 50 個ランダムにサンプリングすることによって用意した。YeastNet2 に収録されている有向グラフには辺ラベルが与えられていないため、ランダムにサンプリングされた有向グラフの各辺に対してランダムにラベルを与えた。実験に使用する辺ラベル付き有向グラフの例を図 3 に示す。 ネットワーク特徴量を計算するために使用する経路ラベル決定法には、影響伝搬の終端を考慮した経路ラベル決定法を用いた。検証実験を行う際に、グラフを表現した数ベクトルの空間が通常のベクトル空間のように整った空間であるとは限らないため、通常の内積や相関を用いた類似度では正しくグラフ間の類似性が測ることができない可能性がある。そこで本研究では、整った空間でない場合でも類似度を測ることのできるカーネル主成分分析[15] を用いて特徴空間での点の分布を調べた。カーネル主成分分析においては予備実験の結果から、カーネル関数としてガウシアンカーネルをパラメータ $\sigma = 0.4$ を指定して使用することとした。

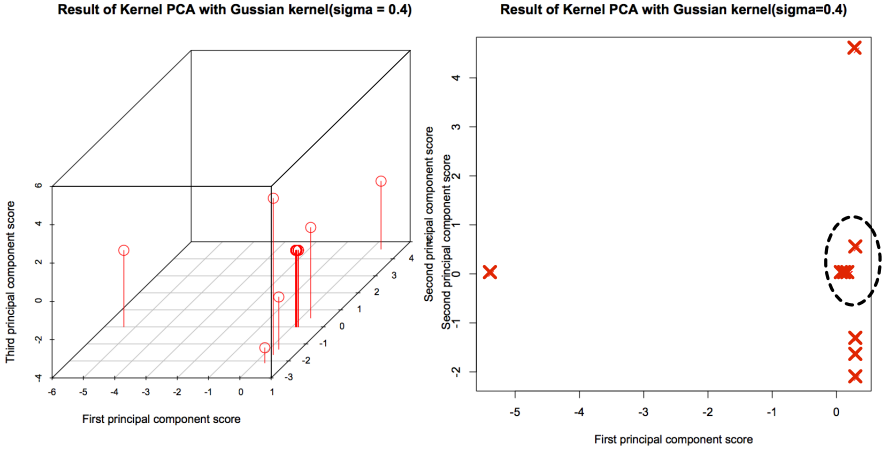

図 4 カーネル主成分分析を行った結果

点の分布を調べた結果を図 4 に示す。 図 4 左では、第一主成分得点が x 軸、第二主成分得点が y 軸、第三主成分が z 軸に示されている。また図 4 右では第一主成分が x 軸、第二主成分が y 軸に示されている。図 4 をみると、空間内において点が局在している部分が複数存在することがわかる。このことから、数ベクトル表現 (8) を用いてネットワーク類似性を定義することができる可能性が示された。

また 50 個の辺ラベル付き有向グラフのうち、45 個がほぼ近いごく近い位置に局在していることがわかる (図 4 右の破線部分)。これは、YeastNet2 という単一の生物種のネットワークから Modularity という単一の指標を用いて辺ラベル付き有向グラフを生成したことが原因である可能性がある。本実験において、類似していないと判定された 4 つの辺ラベル付き有向グラフと破線部分に局在している辺ラベル付き有向グラフの一部を図 5 に示す。 図 5 の上段と下段を比較すると、類似していないとみなされたグラフ群は、極度に緊密な部分が存在していることがわかる (図 5 下段破線部分)。しかしながら、類似しているとみなされたグラフ群にも一部緊密な部分が存在している場合があった。拡張した数ベクトル表

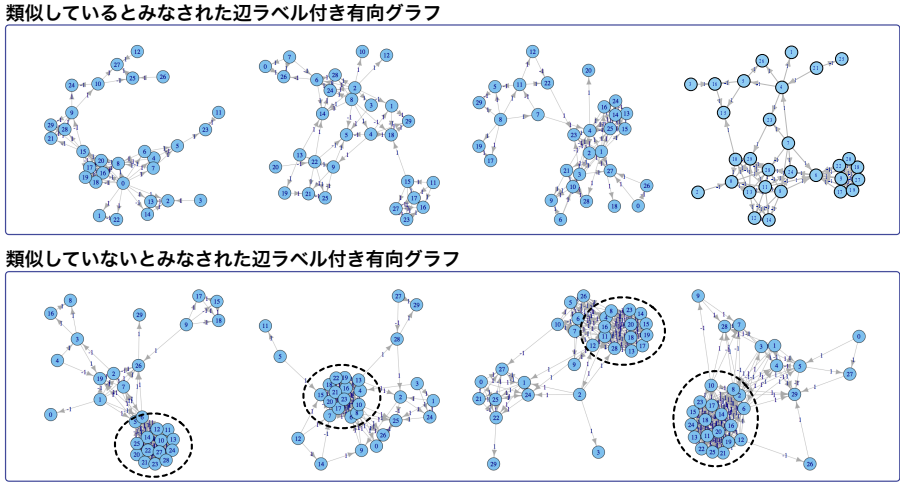

図 5 カーネル主成分分析の結果、類似しているとみなされたグラフ群 (上段) と類似していないとみなされたグラフ群 (下段)

現による辺ラベル付き有向グラフ間の類似度では、どのような構造が同一なものとみなされているかを調べるには、さらなる検証が必要となることがわかった。

8. ま と め

本研究では、従来のネットワーク特徴量を辺ラベル付き有向グラフに適用可能なものへと拡張し、それらを用いたグラフ間類似度の提案を試みた。検証実験においては、提案したネットワーク特徴量を用いることにより、グラフ間の類似性を測ることができる可能性が示されたが、どのような構造を同一なものとみなしているかに関しては、より詳細な検証が必要であることがわかった。

参 考 文 献

1) Stanford Microarray Database, http://smd.stanford.edu
2) Biolog Corp, $PhenotypeMicroArrays^{TM}$, http://www.biolog.com/pmTechDesOver.shtml
3) 安富祖 仁, 岡崎 威生, 名嘉村 盛和, “共分散選択と PageRank に基づく評価関数による遺伝子ネットワーク推定”, SIG-BIO, 58, pp.5-8, 2008
4) Insuk L, Zhihua L, Edward. M, “An Improved, Bias-Reduced Probabilistic Gene Network of Baker’s Yeast, *Saccharomyces cerevisiae*”, PLoS One 3:2(10), 2007
5) Natasa Przulj, “Biological network comparison using graphlet degree distribution”, ECCB Vol.23, pp.177-183, 2006
6) 寺田 愛花, 瀬々 潤, “大域的ネットワークアラインメントを用いた遺伝子機能の比較”, SIG-BIO, 24(12), pp.1-7, 2011
7) Oleksii K, Natasa P, “Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human”, ECCB Vol.00, pp.1-7, 2010
8) 安富祖 仁, 岡崎 威生, 名嘉村 盛和, “生物ネットワークアラインメントのためのノード削除応答に基づいたノード間類似度”, SIG-BIO technical report, 26, 2011
9) Watts D.J, Strogatz S.H, “Collective dynamics of small world networks”, Nature, Vol.393, pp.440-442, 1998
10) David Eppstein, Joseph Wang, “Fast Approximation of Centrality”, Journal of Graph Algorithm and Applications, Vol.8(1), pp.39-45, 2004
11) 鈴木 智也, “情報伝達に基づいた有向重み付き複雑ネットワーク解析”, 情報処理学会論文誌, 数理モデル化と応用 Vol.2, pp.70-78, 2009
12) Sergey B, Lawrence P, “The anatomy of a large-scale hyper textual Web search engine”, Computer Networks and ISDN System, Vol.30, Issues1-7, pp.107-117, 1998
13) Mooney, Christopher Z, “Monte Carlo simulation”, Quantitative applications in the social sciences, Vol.116, pp.103-111, 1997
14) Ulrik B, Daniel D, Marco G, Robert G, “On Finding Graph Clustering with Maximam Modularity”, Graph-Theoretic Concepts in Computer Science, Lecture notes in computer science, Vol.4769, pp.121-132, 2007
15) 赤穂昭太郎, “カーネル多変量解析-非線形データ解析の新しい展開-”, シリーズ確率と情報の科学, 岩波書店, 2008