Vol.2011-BIO-24 No.12
2011/3/11

# 大域的な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

寺田愛花[†1]　瀬々潤[†1]

進化による変異を調べる新しい手法として，遺伝子ネットワークの種間比較がある．既存手法は，複合体やパスウエイの保存を調べる局所構造の調査であるが，進化過程では機能モジュールの組み替えの様に，局所構造に変化は無くとも大きな変化が起こる可能性が有る．本研究ではこの様な大域的な変化を同定する手法を開発し，線虫とショウジョウバエの遺伝子ネットワークを比較した結果，成長に関わるネットワークの構造は保存が高く，高次の機能は保存度が薄い事が明らかになった．

# Global Network Alignment using Graph Summarization for Comparison Gene Function

Aika Terada[†1] and Jun Sese[†1]

By comparison of gene networks between species, gain-of-functions in evolutional process might be found. Existing methods to compare networks focus on local network structure, and suitable for finding changes in small modules such as complexes. However, global change such as network alteration among modules may have been caused evolution. To find such global change of network, we introduce a novel method to compare gene networks. We applied our method to networks observed from C. elegans and D. melanogaster, and found that network structure related to growth are highly conserved while networks related to high-order function have little conservation.

†1 お茶の水女子大学　大学院人間文化創成科学研究科
Department of Computer Science, Ochanomizu University

## 1. はじめに

遺伝子の多くは，単一で機能するのではなく，複数の遺伝子が相互作用することで機能する．各遺伝子がどのような機能に関わっているのかを解明するためには，遺伝子単体に着目して解析するだけではなく，遺伝子がどの遺伝子と関連しているのかに着目した解析が重要であり，相互作用により構築される遺伝子ネットワークの解析が盛んに行われている[1)]．主な遺伝子ネットワークの解析手法の一つは，ネットワークの頂点をクラスタリングするグラフクラスタリングである[2),3)]．これらの手法は，同一の種の遺伝子ネットワークから互いに密接に相互作用している遺伝子グループを発見する手法である．

近年，遺伝子ネットワークは複数種で構築されている[4)]．これらのネットワークを種間で比較することで，塩基配列の変異以外に機能獲得のきっかけになった生体内の変異の解明が期待できる．ネットワークアラインメント手法[5),6)] は，遺伝子ネットワークから種間で異なる部分を発見する手法であるが，この手法では，抽出されたネットワーク内の局所的な変化しか発見できない．進化の過程では，このような局所的な変化が伝播し，機能モジュールの間の相互作用が変化するような，ネットワークの大域的な変化が起きていることが考えられる．本研究では，二種の遺伝子ネットワークから遺伝子クラスタの間の変化を発見することで，進化的イベントと遺伝子ネットワークの大域的な変化の関係を発見する手法を提案する．

図 1 (A) は本研究で扱うネットワークとクラスタリングの例である．赤い頂点 1 から 6 と，青い頂点 a から f で構築される二種の遺伝子ネットワークがあり，相互作用している遺伝子同士を実線でつないでいる．また，ネットワークの間の関係を，種を越えて保存したオーソログ関係で与える．図 1 (A) では，ネットワークの間をつなぐ点線がオーソログ関係を示しており，例えば頂点 1 と a はオーソログである．赤と青の四角は頂点のクラスタを表している．本研究では，このクラスタ分類からクラスタ間の関係を表すグラフを構築し，これを概要グラフと呼ぶ．概要グラフをネットワークアラインメントすることで，クラスタの間で変化した相互作用関係を抽出する．図 1(B) は (A) の分類から構築される，概要グラフの例である．四角い頂点はクラスタを表しており，赤は種 1 のクラスタ，青は種 2 のクラスタである．辺が密に張られているクラスタ同士を線でつないでおり，同じ種の遺伝子で構築されたクラスタ間は実線で，異なる種の遺伝子で構築されたクラスタ間は点線でつないでいる．この概要グラフはアラインメントされており，点線でつながれたどの 2 頂点の間も辺の張り方は同じである．本研究では，このようなクラスタ間の関係を抽出し，かつ



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

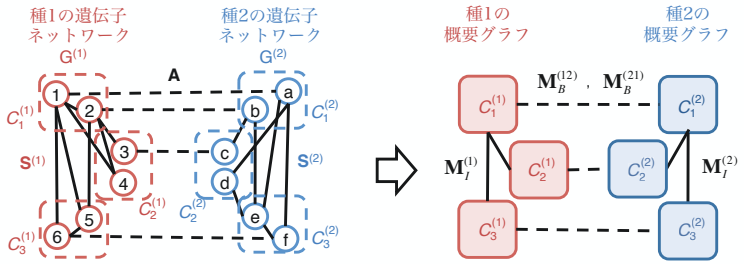


**図 1** 入力するネットワークと解析結果の例
Fig. 1 Example of given networks and aligned summary graphs.

アラインメントされた概要グラフを求めることで，ネットワークを大域的に比較する手法を提案する．

## 2. 関 連 研 究

ネットワーク解析は，近年盛んに行われており，遺伝子ネットワーク以外にも，Web やソーシャルネットワークなどの幅広い分野でネットワークが解析されている．本研究では二つのネットワークを同時に解析しているが，ネットワークの比較手法には各々のネットワークを解析し，その結果を比較する手法も考えられる．

一つのネットワーク解析で頻繁に用いられる手法の一つに，グラフクラスタリング手法[3),7),8)] がある．これらの手法は，クラスタ内の頂点は密に，クラスタ間の辺は疎になるようネットワーク全体をクラスタに分割する．この手法では，クラスタの間の密接な関係を抽出できず，本研究で構築するような，概要グラフを構築することは難しい．

概要グラフを構築する手法には，クラスタ間の密接な関係を抽出し，ネットワークを少ない頂点と辺で近似するグラフの要約手法[9)–11)] がある．これらの手法は一つのネットワークを近似するため，本研究で目的としている二つのネットワークからアラインメントされた概要グラフを構築することは難しい．

本研究と同様，二つのネットワークを同時に解析する手法，ネットワークアラインメント[5),6)] では，ネットワーク内で強く保存している部分ネットワークを抽出することで，ネットワーク内で局所的に保存している部分を発見することができる．この手法ではネットワークを部分的にしか比較することができないが，本研究で導入する概要グラフのアラインメント問題では，ネットワークの全体構造を表す概要グラフを構築し，それらをアラインメントすることで，ネットワークを大域的に比較することができる．

## 3. 概要グラフのアラインメントの定式化

本章では，概要グラフのアラインメントを求めるために新たな指標を導入し，最適化問題として定式化する．まず，ネットワークやクラスタの記法を導入し，次に指標を定義する．

### 3.1 ネットワークとクラスタ，概要グラフの定義

本節では，遺伝子ネットワークとクラスタ分類，概要グラフを定義する．まず，解析するをネットワークを定義する．

**定義 1　ネットワークとネットワーク間の関係**

ネットワークを $G = (G^{(1)}, G^{(2)}, E_B)$ とする．$G^{(1)}$ と $G^{(2)}$ は異なる種のネットワークであり，$E_B$ は $G^{(1)}$ と $G^{(2)}$ の間をつなぐ辺である．$G^{(i)} = (V^{(i)}, E_I^{(i)})$ であり，$V^{(i)}$ は $G^{(i)}$ に含まれる頂点，$E_I^{(i)}$ はそれらをつなぐ辺である．

図 1(A) の赤と青の円は，$V^{(1)}$ と $V^{(2)}$ に含まれる頂点を表しており，それぞれ 1 から 6 と a から f の頂点がある．また，赤い頂点同士をつなぐ実線は $E_I^{(1)}$ に含まれる辺を，青い頂点同士をつなぐ実線は $E_I^{(2)}$ に含まれる辺を表し，赤と青の頂点をつなぐ点線は $E_B$ に含まれる辺である．$G^{(1)}$ と $G^{(2)}$ は，隣接行列 $\mathbf{S}^{(1)}$ と $\mathbf{S}^{(2)}$，$E_B$ を隣接行列 $\mathbf{A}$ で表す．本研究で扱うネットワークは，重み無しの無向グラフであり，$\mathbf{S}^{(1)}$，$\mathbf{S}^{(2)}$，$\mathbf{A}$ は対称行列である．

**定義 2　ネットワークを表す隣接行列**

$G^{(i)}$ を隣接行列 $\mathbf{S}^{(i)} \in \{0,1\}^{n^{(i)} \times n^{(i)}}$ で，$E_B$ を隣接行列 $\mathbf{A} \in \{0,1\}^{n^{(1)} \times n^{(2)}}$ で定義する．$n^{(i)}$ は $G^{(i)}$ にある頂点数である．また，$\mathbf{A}^{(12)} = \mathbf{A}^{(21)} = \mathbf{A}$ と表記する．頂点 $v_p^{(i)}, v_q^{(i)} \in V^{(i)}$ について，辺 $(v_p^{(i)}, v_q^{(i)}) \in E_I^{(i)}$ のとき，$\mathbf{S}^{(i)}$ の $pq$ 要素は 1 であり，それ以外のときは 0 である．$\mathbf{A}$ についても同様であり，$v_p^{(1)} \in V^{(1)}$，$v_q^{(2)} \in V^{(2)}$ について，辺 $(v_p^{(1)}, v_q^{(2)}) \in E_B$ のとき，$\mathbf{A}$ の $pq$ 要素は 1 であり，それ以外のときは 0 である．

例として，図 1(A) の $G^{(1)}$ と $E_B$ の隣接行列 $\mathbf{S}^{(1)}$ と $\mathbf{A}$ を次に示す．

$$\mathbf{S}^{(1)} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

$\mathbf{S}^{(1)}$ の行と列はどちらも頂点 1 から 6 を表しており，$\mathbf{A}$ では行は頂点 1 から 6 を，列は頂点 $a$ から $f$ を表している．$G^{(1)}$ の頂点 3 は，$G^{(2)}$ の頂点 $c$ にのみ点線の辺を張っているため，頂点 3 を表す 3 行目は， 頂点 $c$ を表す 3 列目のみ 1 であり，それ以外が 0 である．

頂点の個数が異なるネットワークは比較が難しいため，本研究では，$G^{(1)}$ と $G^{(2)}$ をそれぞれ $k$ 個のクラスタで近似した概要グラフをアラインメントすることで，ネットワークを大域的に比較する．まず，頂点のクラスタを定義する．

**定義 3　頂点のクラスタ分類**

$V^{(i)}$ をクラスタ $\mathcal{C}^{(i)} = \{C_1^{(i)}, \cdots, C_k^{(i)}\}$ に分類する．$\forall p$ について $C_p^{(i)} \subseteq V^{(i)}$ であり，$C_p^{(i)} \neq \Phi$ である．また， $\forall p, q$ について $C_p^{(i)} \cap C_q^{(i)} = \Phi$ である．

図 1(A) では，$G^{(1)}$ のクラスタを赤い四角で， $G^{(2)}$ のクラスタを青い四角で表している．例えば，$C_2^{(1)} = \{3, 4\}$ である．このクラスタ分類を行列 $\mathbf{C}^{(i)}$ で表す.

**定義 4　頂点のクラスタ分類を表す行列**

$V^{(i)}$ のクラスタ分類を行列 $\mathbf{C}^{(i)} \in \{0, 1\}^{n^{(i)} \times k}$ で表す．$v_p^{(i)} \in V^{(i)}$， $C_q^{(i)} \in \mathcal{C}^{(i)}$ について，$v_p^{(i)} \in C_q^{(i)}$ であれば $\mathbf{C}^{(i)}$ の $pq$ 要素は 1，それ以外の場合は 0 である．また，全ての頂点はいづれかの一つのクラスタに属するため，制約 $\mathbf{C}^{(i)}\mathbf{1} = \mathbf{1}$ を与える．$\mathbf{1}$ は全ての要素が 1 の列ベクトルである．

次に，クラスタ分類を用いて概要グラフ $L$ を定義する．これはクラスタ間の辺の本数で定義する．

**定義 5　概要グラフ**

グラフ $L_I^{(i)}$ を，頂点を $G^{(i)}$ のクラスタ $C_p^{(i)} \in C^{(i)}$ で，辺を $C_p^{(i)}$ と $C_q^{(i)}$ に張られている辺の本数で構築する．同様に， $L_B$ の頂点を $C^{(1)}$， $C^{(2)}$ で，辺を $C_p^{(1)}$ と $C_q^{(2)}$ の間に張られている辺の本数で構築する．これらを満たす $L_I^{(i)}$ と $L_B$ で構築されるグラフ $L = (L_I^{(1)}, L_I^{(2)}, L_B)$ を概要グラフと呼ぶ.

$L_I^{(i)}$ と $L_B$ は，それぞれ行列 $\mathbf{M}_I^{(i)}$ と $\mathbf{M}_B$ で定義する．

**定義 6　概要グラフを表す行列**

クラスタの間の辺の本数を表す行列を $\mathbf{M}_I, \mathbf{M}_B \in \mathbb{R}^{k \times k}$ とする．$\mathbf{M}_I^{(i)}$ の $pq$ 要素は $C_p^{(i)}$ と $C_q^{(i)}$ の間の辺の本数を表しており，$\mathbf{M}_B^{(ij)}$ の $pq$ 要素は $C_p^{(i)}$ と $C_q^{(j)}$ の間の辺の本数を表している．それぞれ次式で算出できる．

$$\mathbf{M}_I^{(i)} = (\mathbf{C}^{(i)})^{\mathrm{T}} \mathbf{S}^{(i)} \mathbf{C}^{(i)} \tag{1}$$

$$\mathbf{M}_B^{(ij)} = (\mathbf{C}^{(i)})^{\mathrm{T}} \mathbf{A}^{(ij)} \mathbf{C}^{(j)} \tag{2}$$

$\mathbf{M}_I^{(i)}$ と $\mathbf{M}_B^{(ij)}$ は対称行列である．

例として，図 1 (A) の $\mathbf{C}^{(1)}$ と $\mathbf{M}_I^{(1)}$ を示すと，それぞれ $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 3 & 3 \\ 3 & 0 & 0 \\ 3 & 0 & 2 \end{pmatrix}$ である．

### 3.2　概要グラフのアラインメントの良さを示す指標

本節では，クラスタ分類 $\mathbf{C}$ の良さを表す指標を定義することで，概要グラフのネットワークアラインメント問題を定式化する．

本研究では，クラスタ分類から概要グラフを構築し，これらをアラインメントすることで，ネットワークの間で共通なクラスタ間の関係を抽出する．このようなアラインメントを行うため，次の二つの条件を満たすクラスタ分類を求める．

(1)　与えられたネットワークと概要グラフが有する情報に差がない.

(2)　二つの概要グラフが一致している.

条件 1 は構築した概要グラフに対する条件である．ネットワークと概要グラフの間で表している情報に差がないほど，概要グラフはネットワークを正確に表していると言える．条件 2 は概要グラフのアラインメントに対する条件である．二つのネットワークが完全に一致している場合，どの頂点の間も共通の構造を有するようにアラインメントすることが可能であり，ネットワークの間で共通するクラスタ間の強い関係を多く抽出することができる．

まず，この二つの条件それぞれを表す指標を導入し，次に，それらの両方を満たす時に最小となる新たな指標を定義する．

#### 3.2.1　ネットワークと概要グラフの情報の差を表す指標

まず，$\mathbf{C}$ から構築される概要グラフとネットワークの間の情報の差を表す指標を定義する．

ネットワークと概要グラフの差は，頂点が各クラスタに対して張る辺の情報の差で定義する．これは，$G^{(i)}$ の辺については，式 1 の左側から $(\mathbf{C}^{(1)})^{\mathrm{T}}$ の逆行列を掛けた $\mathbf{S}^{(i)}\mathbf{C}^{(i)}$ と $((\mathbf{C}^{(i)})^{\mathrm{T}})^{-1}\mathbf{M}_I^{(i)}$ を近似することに相当する．しかし，$k \times n^{(i)}$ 行列である $(\mathbf{C}^{(i)})^{\mathrm{T}}$ の逆行列は存在しないため，代わりに $\mathbf{C}^{(i)}$ を用いる．また，$\mathbf{S}^{(i)}$ と $\mathbf{M}_I^{(i)}$ には行列の要素が取る範囲に差があるため，行ごとに総和が 1 となるように変換した行列 $\overline{\mathbf{S}}^{(i)}$ と $\overline{\mathbf{M}}_I^{(i)}$ を使用する．図 1 (A) の $G^{(1)}$ では，それぞれ次の行列になる．



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

$$\overline{\mathbf{S}}^{(1)}\mathbf{C}^{(1)} = \begin{pmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 2/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 \end{pmatrix}, \mathbf{C}^{(1)}\overline{\mathbf{M}}^{(1)} = \begin{pmatrix} 1/4 & 3/8 & 3/8 \\ 1/4 & 3/8 & 3/8 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 3/5 & 0 & 2/5 \\ 3/5 & 0 & 2/5 \end{pmatrix}$$

$\mathbf{M}_I^{(i)}$ が $\mathbf{S}^{(i)}$ の情報を欠落することなく概要を表していれば，$\overline{\mathbf{S}}^{(i)}\mathbf{C}^{(i)}$ と $\mathbf{C}^{(i)}\overline{\mathbf{M}}_I^{(i)}$ は完全に一致した行列になる．これは，$\mathbf{A}^{(ij)}$ と $\mathbf{M}_B^{(ij)}$ でも同様である．

この関係を用いて， ネットワークと概要グラフの情報の差を行列の間の距離で定義する．まず，二つの行列 $\mathbf{X}$ と $\mathbf{Y}$ の距離，$Dist(\mathbf{X},\mathbf{Y})$ を定義し，次に，ネットワークと概要グラフの情報の差を表す指標を定義する．

**定義 7　　行列の距離の定義**

行列 $\mathbf{X}$ の $p$ 行を $\mathbf{x}_p$ で表す．また，ベクトル $\mathbf{x}_i$， $\mathbf{y}_j$ のコサイン距離を $CosDist(\mathbf{x}_i,\mathbf{y}_j)$ で表す．$n$ 行 $m$ 列の行列 $\mathbf{X}$， $\mathbf{Y}$ の間の距離は，$Dist(\mathbf{X},\mathbf{Y}) = 1/n\sum_{i=1}^{n} CosDist(\mathbf{x}_i,\mathbf{y}_i)$ とする．

**定義 8　　ネットワークと概要グラフの情報の差を表す指標**

$G$ と概要グラフの情報の差を，次式で定義する．

$$\Delta_S(\mathbf{C}) = Dist(\mathbf{GC},\mathbf{CM}) \tag{3}$$

ただし， $\mathbf{G}$，$\mathbf{C}$， $\mathbf{M}$ はそれぞれ，$\begin{pmatrix} \overline{\mathbf{S}}^{(1)} & w_1\overline{\mathbf{A}}^{(12)} \\ w_1\overline{\mathbf{A}}^{(21)} & \overline{\mathbf{S}}^{(2)} \end{pmatrix}$，$\begin{pmatrix} \mathbf{C}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{(2)} \end{pmatrix}$，$\begin{pmatrix} \overline{\mathbf{M}}_I^{(1)} & w_1\overline{\mathbf{M}}_B^{(12)} \\ w_1\overline{\mathbf{M}}_B^{(21)} & \overline{\mathbf{M}}_I^{(2)} \end{pmatrix}$ であり，$\mathbf{0}$ は全ての要素が 0 の零行列である．$\overline{\mathbf{S}}^{(i)}$， $\overline{\mathbf{A}}^{(ij)}$， $\overline{\mathbf{M}}_I^{(i)}$， $\overline{\mathbf{M}}_B^{(ij)}$ は $\mathbf{S}^{(i)}$， $\mathbf{A}^{(ij)}$， $\mathbf{M}_I^{(i)}$， $\mathbf{M}_B^{(ij)}$ を各々の行の総和が 1 となるようにした行列である．行列 $\mathbf{X}$ を変換した行列 $\overline{\mathbf{X}}$ の $pq$ 要素は次式で算出する．

$$\overline{\mathbf{X}}_{pq} = \mathbf{X}_{pq} / \sum_{r=1} \mathbf{X}_{pr} \tag{4}$$

$w_1$ はネットワークをつなぐ辺に含まれるノイズを考慮する任意のパラメータである．

### 3.2.2 概要グラフのアラインメントの良さを表す指標の定義

本節では， $\mathbf{C}$ から構築される概要グラフのアラインメントの良さを表す指標を定義する．

二つのネットワークが完全に一致しているとき，どの頂点に対しても辺の張り方が完全に一致する頂点がもう一方のネットワークに存在し，全ての頂点を含むネットワークアラインメントできる．この関係から，二つのネットワークの概要グラフが完全に一致しているとき，概要グラフが最適なアラインメントをされていることを示す指標を定義する．また，$C_p^{(1)}$ と $C_p^{(2)}$ が辺の張り方が共通のクラスタとして対応づくため，概要グラフの間の関係を表す行列 $\overline{\mathbf{M}}_B^{(ij)}$ は対角行列になる．この関係から，概要グラフのアラインメントの良さを次のように定義する．

**定義 9　　概要グラフのアラインメントの良さを表す指標**

$\mathbf{M}_I^{(1)}$ と $\mathbf{M}_I^{(2)}$ のアラインメントの良さを，次式で定義する．

$$\Delta_A(\mathbf{C}) = Dist(\overline{\mathbf{M}}_I^{(1)},\overline{\mathbf{M}}_I^{(2)}) + \sum_{i,j=1,2} Dist(\overline{\mathbf{M}}_B^{(ij)},\tilde{\mathbf{M}}_B^{(ij)}) \tag{5}$$

$\tilde{\mathbf{M}}_B^{(ij)}$ は対角成分が $\overline{\mathbf{M}}_B^{(ij)}$ と等しく，それ以外は 0 の行列である．

### 3.2.3 概要グラフの誤差とアラインメントの良さを表す指標の定義

$\mathbf{C}$ から構築した概要グラフは，入力したネットワークと情報に差が少ないほど $\Delta_S(\mathbf{C})$ が小さく，概要グラフ同士がアラインメントできているほど $\Delta_A(\mathbf{C})$ が小さくなる．この二つを加算することで， $\mathbf{C}$ の良さを測る指標を定義する．

**定義 10　　構築した概要グラフの情報の誤差とそのアラインメントの良さを表す指標**

$\mathbf{C}$ について，概要グラフとネットワークの情報の差と，概要グラフのアラインメントの良さを示す指標を次式で定義する．

$$\Delta(\mathbf{C}) = \Delta_S(\mathbf{C}) + w_2\Delta_A(\mathbf{C}) \tag{6}$$

$w_2$ はネットワークに含まれるノイズや，二つのネットワークの間の変異を考慮する任意のパラメータである．

$\Delta(\mathbf{C})$ は，クラスタの個数が少ないほど値が小さくなる傾向があるため，本研究ではクラスタに含まれる頂点数が同程度になるように分類する．また， $\forall C_p^{(i)} \in \mathcal{C}^{(i)}$ について $C_p^{(i)} \neq \Phi$ であるため，$\mathbf{C}$ は $\mathbf{C1} = \mathbf{1}$ を満たす行列である．

## 4. 提 案 手 法

本章では， 3 章で定義した $\Delta(\mathbf{C})$ を最適化する手法，ALIgnment with Cluster using Edge connectivity (ALICE) を提案する．

Algorithm 1 に，ALICE の概略を示す．$\mathbf{C}$ は $\mathbf{C1} = \mathbf{1}$ を満たすため，各行に 1 が一カ所のみ立っている行列である．ALICE は，$k$-means のように，$\mathbf{M}$ と $\mathbf{C}$ の計算を $\Delta(\mathbf{C})$ が収束するまで繰り返す．



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

**Algorithm 1** : ALICE($\mathbf{G}$, $k$, $w_1$, $w_2$)

**Require:** ネットワークを表す隣接行列 $\mathbf{G}$，クラスタ数 $k$， パラメータ $w_1$ と $w_2$.
1: 各頂点をランダムに分類して $\mathbf{C}$ を構成する.
2: **repeat**
3: // 行列 $\mathbf{M}$ の更新
4: $\overline{\mathbf{M}}^{(i)}$ と $\overline{\mathbf{M}}_B^{(ij)}$ を計算する.
5: $\overline{\mathbf{M}}'^{(i)}_I \leftarrow \overline{\mathbf{M}}_I^{(i)} + w_2\overline{\mathbf{M}}_I^{(j)}$
6: $\overline{\mathbf{M}}'^{(ij)}_B \leftarrow \overline{\mathbf{M}}_B^{(ij)} - w_2(\overline{\mathbf{M}}_B^{(ij)} - \tilde{\mathbf{M}}_B^{(ij)})$
7: $\overline{\mathbf{M}}_I^{(i)} \leftarrow \overline{\mathbf{M}}'^{(i)}_I$, $\overline{\mathbf{M}}_B^{(ij)} \leftarrow \overline{\mathbf{M}}'^{(ij)}_B$ とする.
8: // 行列 $\mathbf{C}$ の更新
9: **for** $1 \le p \le n^{(i)}$ **do**
10: **for** $1 \le q \le K$ **do**
11: $G^{(i)}$ の頂点 $v_p^{(i)}$ と $C_q^{(i)}$ の距離を算出し，距離が最小のクラスタに $v_p^{(i)}$ を分類する.
12: **end for**
13: **end for**
14: **until** $\Delta(\mathbf{C})$ が収束するまで繰り返す.
15: return $\mathbf{C}$.

Algorithm 1 の 11 行目では，頂点とクラスタの距離を算出し，その距離が最小となるクラスタに $v_p$ を分類する．頂点とクラスタの距離は，頂点 $v_p$ が各クラスタに対して張る辺の情報を表すベクトル $\mathbf{v}_p$ と，クラスタの重心ベクトル $\mathbf{m}_q$ の距離で定義する．$\mathbf{v}_p$ と $\mathbf{m}_p$ の $q$ 要素は，それぞれ，$\mathbf{GC}$ の $pq$ 要素と $\mathbf{CM}$ の $pq$ 要素で与える．$\mathbf{v}_p$ と $\mathbf{m}_q$ の距離は，ベクトルのコサイン距離 $CosDist(\mathbf{v}_p, \mathbf{m}_q)$ を含む次式で定義する．

$$Dist(\mathbf{v}_p^{(i)}, \mathbf{m}_q^{(i)}) = CosDist(\mathbf{v}_p^{(i)}, \mathbf{m}_q^{(i)}) + \frac{k}{n^{(i)}}(n_q^{(i)} - n_{orig}^{(i)})$$

第 2 項は，属する頂点数が均等になるよう，クラスタに含まれる頂点数に対するペナルティである．$n_{orig}^{(i)}$ は，更新前に頂点 $v_p^{(i)}$ が属しているクラスタ $C_{orig}^{(i)}$ が含む頂点数であり，$n_q^{(i)}$ は，更新後のクラスタ $C_q^{(i)}$ が含む頂点数である．

## 5. 疑似データの解析

本章では，疑似ネットワークを解析し，二つのパラメータ $w_1$ と $w_2$ の効果と，ALICE のノイズ耐性を示す．疑似ネットワークは実データに即して疎なものを作成した．クラスタ数は 20 個あり，多くの頂点に対して辺を張るハブのクラスタ $C_{20}^{(i)}$ と，$C_{20}^{(i)}$ に辺を張るクラスタを 19 個作成している．$p = 1, \cdots 19$ について，$C_p^{(i)}$ と $C_{p+1 \bmod 19}^{(i)}$ に属する頂点

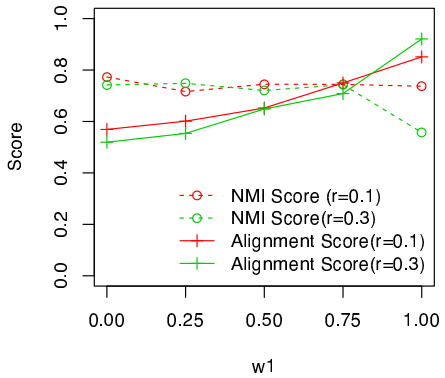


**図 2** $w_1$ による NMI スコアと $J_A(\mathbf{C})$ の変化
Fig. 2 NMI score and $J_A(\mathbf{C})$ according to $w_1$.

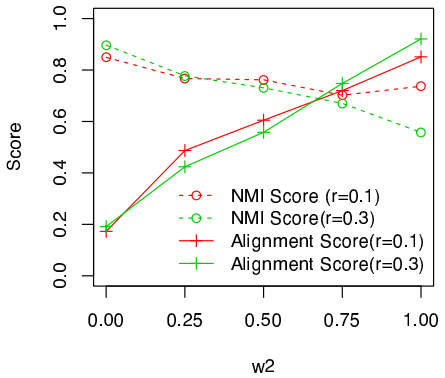


**図 3** $w_2$ による NMI スコアと $J_A(\mathbf{C})$ の変化
Fig. 3 NMI score and $J_A(\mathbf{C})$ according to $w_2$.

の間に辺を張った．ネットワークの間の辺は，ランダムに選んだ 10 個のクラスタに関して $C_p^{(1)}$ と $C_p^{(2)}$ の間に作成している．また，作成した $G^{(1)}$， $G^{(2)}$， $E_B$ に対し，$r$% の辺をランダムに追加することでノイズを付与した．

解析結果は二つの指標で計測する．一つ目は，疑似データで作成した正しいクラスタ分類と解析結果のクラスタ分類の類似度である．これは，正規化平均相互情報量 (NMI スコア)[12] で計測する．本研究では， $G^{(1)}$ と $G^{(2)}$ それぞれでこのスコアを算出し，その平均を NMI スコアとしている．NMI スコアでは，概要グラフのアラインメントの良さを計測できないため，これを式 6 を用いて $J_A(\mathbf{C}) = 1 - \Delta_A(\mathbf{C})$ で計測する．どちらの指標も，適切な分類であるほど 1.0 に近く，最小値は 0 である．

二つのパラメータ $w_1$ と $w_2$ の効果を調べるため，$r$ を 0.1 と 0.3 でノイズを加えたデータを解析した．図 2 は $w_1$ の効果を調べるため，$w_2 = 1.0$ で $w_1$ を変えて解析した結果であり，図 3 は $w_1 = 1.0$ とし，$w_2$ を変えて解析した結果である．図 2 の x 軸は $w_1$ を，図 3 の x 軸は $w_2$ を示している．y 軸は図 2，図 3 ともに NMI スコアと $J_A(\mathbf{C})$ の値を示しており，$J_A(\mathbf{C})$ は Alignment Score と記している．

図 2 と 図 3 では，類似した傾向がある．$J_A(\mathbf{C})$ の値は $w_1$， $w_2$ が大きくなるほど高いスコアになっており，ノイズの割合による違いはない．NMI スコアは，ノイズが低ければ $w_1$， $w_2$ に依らずスコアに大きな変化はない．しかし，ノイズを多く含む $r = 0.3$ では，$w_1$ と $w_2$ のどちらも大きいとスコアが減少している．これらのことから，ノイズが少ない



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

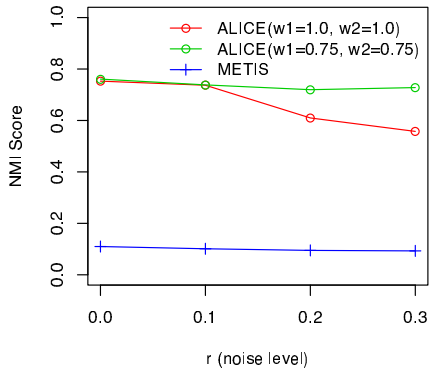


**図 4** ノイズの増加に対する NMI スコアの変化
Fig. 4 Change of NMI score with respect to adding noise.

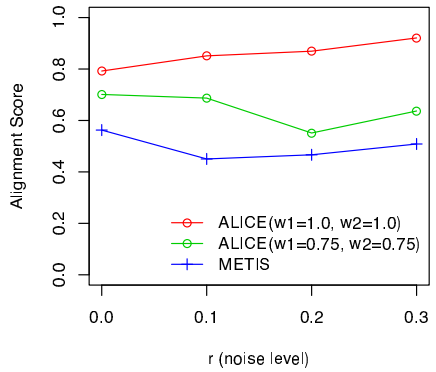


**図 5** ノイズの増加に対する $J_A(\mathbf{C})$ の変化
Fig. 5 $J_A(\mathbf{C})$ value with respect to adding noise.

データは $w_1$ と $w_2$ を高めにし，ノイズが多いデータは $w_1$ と $w_2$ を低めにすることで，ノイズを含むデータに対しても正解セットとの差が小さく，概要グラフもアラインメントされている適切な結果を得られることが分かる．

次に，ALICE のノイズ耐性を示すため，ノイズの割合を変化させたデータで解析を行った．異なるノイズの割合のデータを解析していることを考慮し，ALICE のパラメータは $w_1 = w_2 = 1.0$ と $w_1 = w_2 = 0.75$ の二通りで解析した．また，本手法と比較するため，グラフクラスタリング手法 METIS[2] でも解析した．図 4 は解析結果の NMI スコアを，図 5 は $J(\mathbf{C})$ を示しており，どちらの図も $x$ 軸はノイズの割合 $r$ である．この二つの図から，NMI スコア，$J(\mathbf{C})$ のいづれも，全てのノイズの割合のデータに対し，ALICE の方が METIS よりもスコアが高いことが分かる．また，ALICE のパラメータの違いを比較すると，$r = 0.1$ 以下では NMI スコアには差が見られないが，$J(\mathbf{C})$ は $w_1 = 1.0$，$w_2 = 1.0$ で解析した結果の方がスコアが高い．$r = 0.2$ 以上では $J(\mathbf{C})$ のスコアは　$w_1 = 1.0$，$w_2 = 1.0$ の方が高いものの，NMI スコアは $w_1 = 0.75$, $w_2 = 0.75$ の方が高い．この結果から，本手法が METIS と比較してノイズ耐性が強く，また，ノイズの割合に応じて $w_1$ と $w_2$ を調節することで，正解セットとのずれが少ない結果を求められることが分かる．

## 6. 実データの解析

第 5 章で，本手法のノイズ耐性を示した．本章では，線虫とショウジョウバエの遺伝子ネットワークをクラスタリングし，アラインメントされた概要グラフを構築する．

解析した遺伝子ネットワークは iRefIndex[4] を，ネットワークの間をつなぐオーソログ関係は HomoloGene[13] を利用した．ネットワークの頂点数と辺の本数は線虫が 4,098 個と 9,289 本，ショウジョウバエが 5,813 個と 21,147 本であり，ネットワークの間の辺は 966 本である．

このネットワークを，ALICE と METIS[2] を用いて 100 個のクラスタに分類する．ALICE のパラメータ $w_1$ と $w_2$ はどちらも 1.0 で解析した．

解析結果を $J_A(\mathbf{C})$ で比較する．その結果，ALICE は 0.601，METIS は 0.506 であり，ALICE の方がアラインメントされた概要グラフを構築できたことが分かる．図 6 は，ALICE で求めた概要グラフを可視化した結果である．このグラフでは，赤の頂点が線虫の遺伝子のクラスタを，青の頂点がショウジョウバエの遺伝子のクラスタを表している．頂点の大きさで属する遺伝子の個数を示し，遺伝子数が多いクラスタ程大きい頂点で表している．頂点に表示している文字はクラスタの ID であり，cel は線虫の，dme はショウジョウバエのクラスタである．同一種では，辺の密度が高いクラスタを実線でつないでおり，各々の種で 120 本の辺を張っている．異なる種の間は，クラスタ間の関係でアラインメントされたクラスタ同士を，オーソログ関係の多いクラスタから順に点線で 30 本つないでいる．赤い辺は二つの種の間で共通の辺であり，クラスタ間でアラインメントされた関係である．

図 6 には，赤い辺でつながれたアラインメントされている大きな部分グラフがある（図で矢印で示している）．この部分グラフに含まれているクラスタについて，既知の機能 Gene Ontology[14] と比較した．その結果，developmental process や growth の子タームに該当する，成長に関連する機能に関わっている遺伝子が多く属していた．これらのクラスタの相互作用が共通していることは，いづれの種でも発生過程に大きな違いが無い，既知な知見とも合致している．また，この共通している部分グラフに含まれているクラスタ cel59 と dme59，cel2 と dme2，cel57 と dme57 はクラスタ間の関係は類似しているものの，点線ではつながれていないオーソログ関係の遺伝子が少ないクラスタである．これらのクラスタには，現在ではオーソログとは知られていない遺伝子が属していることが推測できる．



© 2011 Information Processing Society of Japan

Vol.2011-BIO-24 No.12
2011/3/11

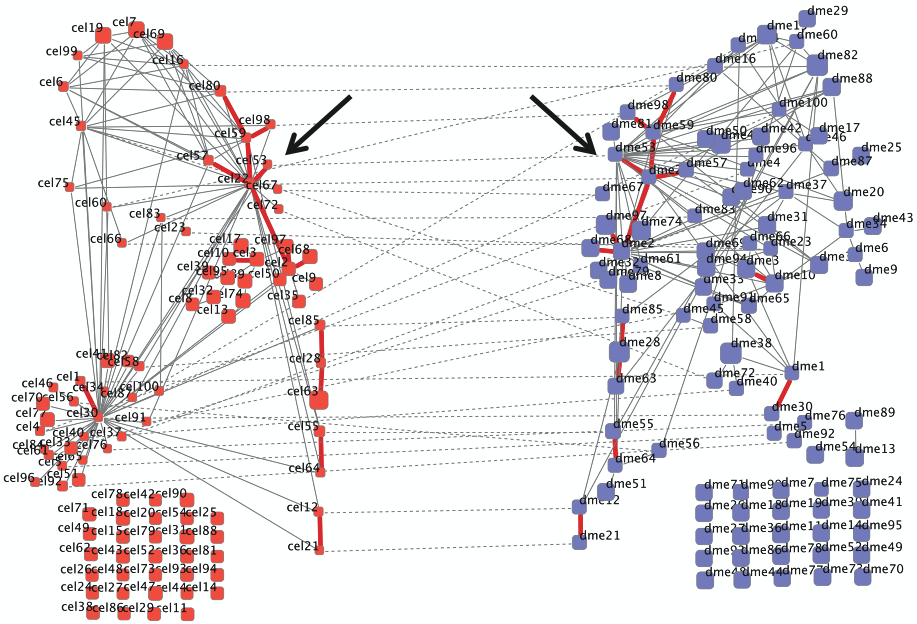

**図 6** 線虫 (cel) とショウジョウバエ (dme) を比較した結果を表す概要グラフ
Fig. 6 Summary graphs aligned between C. elegans and D. melanogaster.

## 7. おわりに

本論文では，二つのネットワークから大きく共通する構造を抽出することでネットワークを比較する，概要グラフアラインメント問題を導入した．この問題を適応することで，遺伝子ネットワークを種間で大域的に比較する事ができる．疑似ネットワークの解析から，本手法がノイズの多いデータにおいてもクラスタ同士をアラインメントできることを示した．また，線虫とショウジョウバエの遺伝子ネットワークを解析した結果，成長に関わる遺伝子のクラスタの間に，共通する相互作用の構造があることが分かった．

本研究では二つのネットワークの比較を行ったが，今後は三つ以上のネットワークを同時に解析するように改良することを考えている．進化の過程に沿って 3 種以上のネットワークを同時に比較することで，どのようにネットワークが変化し，各々の種が機能獲得してきたのか発見することに繋がるだろう．また，種間比較以外にも時系列に沿ったネットワーク比較に応用でき，例えば，細胞が分裂する各段階とネットワークの変化を対応づけることができるだろう．

## 参 考 文 献

1) Barabasi, A.-L. and Oltvai, Z.N.: Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.*, Vol.5, pp.101–113 (2004).
2) Karypis, G. and Kumar, V.: A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs, *SIAM J. SCI. COMPUT.*, Vol.20, pp.359–392 (1999).
3) Ding, C. H.Q., He, X., Zha, H., Gu, M. and Simon, H.D.: A min-max cut algorithm for graph partitioning and data clustering, *ICDM '01*, pp.107–114 (2001).
4) Razik, S., Magklaras, G. and Donaldson, I.M.: iRefIndex: A consolidated protein interaction database with provenance, *BMC Bioinformatics*, Vol.9, p.405 (2008).
5) Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S. and *et al.*: Conserved patterns of protein interaction in multiple species, *Proc. Natl. Acad. Sci.*, Vol.102, pp.1974–1979 (2005).
6) Singh, R., Xu, J. and Berger, B.: Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proc. Natl. Acad. Sci.*, Vol.105, pp.12763–12768 (2008).
7) Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning, *KDD '01*, pp.269–278 (2001).
8) Girvan, M. and Newman, M. E.J.: Community structure in social and biological networks, *Proc. Natl. Acad. Sci.*, Vol.99, pp.7821–7826 (2002).
9) Long, B., Xu, X., Zhang, Z. and Yu, P.S.: Community Learning by Graph Approximation, *ICDM '07*, pp.232–241 (2007).
10) Tian, Y., Hankins, R.A. and Patel, J.M.: Efficient Aggregation for Graph Summarization, *SIGMOD '08*, pp.567–580 (2008).
11) Navlakha, S., Rastogi, R. and Shrivastava, N.: Graph summarization with bounded error, *SIGMOD '08*, pp.419–432 (2008).
12) Strehl, A. and Ghosh, J.: Relationship-Based Clustering and Visualization for High-Dimensional Data Mining, *INFORMS. J. Comput.*, Vol.15, pp.208–230 (2003).
13) Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K. and *et al.*: Database resources of the National Center for Biotechnology Information, *Nucleic. Acids. Res.*, Vol.34, pp.173–180 (2006).
14) Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H. and *et al.*: Gene Ontology: tool for the unification of biiology, *Nat. Genet.*, Vol.25, pp.25–29 (2000).



© 2011 Information Processing Society of Japan