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

理学専攻 情報科学コース　寺田愛花　(指導教員：瀬々潤)

1　はじめに

遺伝子の多くは生体内で他の遺伝子と互いに協調することで機能しており，遺伝子機能の解明のために，遺伝子を頂点，協調関係を辺で与えた遺伝子ネットワークが解析されている [1, 2]．遺伝子は常に同じ遺伝子と協調しているわけではなく，協調する相手を変えることで多彩な機能を果たしており，この変化が生物が進化するきっかけの一つであると考えられる．種間で遺伝子ネットワークを比較することで，種が進化で機能獲得してきた過程の解明が期待できる．

ネットワークを比較する手法は大きく二つに分類される．一つ目は，それぞれのネットワークから密な部分ネットワークを求め [1]，これらを比較する手法である．もう一方は，二つのネットワークに共通する部分構造を求める，ネットワークアラインメントと呼ばれる手法である [2]．どちらの手法もネットワークの局所的な構造を比較するため，進化の過程における大域的な協調関係の変化を見つけることは難しい．

本研究ではネットワークを大域的に比較するため，頂点をクラスタに分類し，クラスタ間の協調関係を抽出する．この分類からクラスタ間の関係を表すグラフを構築し，構築したグラフをアラインメントする手法を提案する．

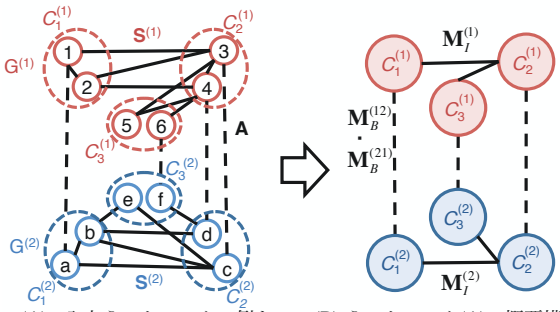
図 1 は，本手法で解析するネットワークと解析結果の例である．図 1(A) は解析するネットワークの例であり，頂点 1 から 6 と a から f で構築された二種の遺伝子ネットワークがある．ネットワークの間をつなぐ点線は，種を越えて保存した遺伝子関係を表している．このネットワークから，本研究では図 1(B) のようなグラフを構築する．これを概要グラフと呼ぶ．図 1(B) の赤と青の頂点はクラスタを表し，辺が密に張られているクラスタをつないでいる．

概要グラフがネットワークから正確に構築され，概要グラフがアラインメントされているとき，次の二つの条件を満たす．(1) ネットワークと概要グラフの情報に差がない．(2) 二つの概要グラフが一致．本研究では，この二つの条件を表す新たな指標と，指標を最小化する手法を提案する．

本手法で，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．構築した概要グラフは，遺伝子ネットワークとの誤差が少なく，クラスタ間の関係でアラインメントされたものであった．また，同じクラスタに分類された遺伝子の多くが同一の機能を有しており，生物学的にも有意な結果を求めることができた．

2　指標の定義と提案手法

ネットワーク $G^{(i)}$ を隣接行列 $\mathbf{S}^{(i)} \in \{0,1\}^{n^{(i)} \times n^{(i)}}$ で定義する．$n^{(i)}$ は，$G^{(i)}$ の頂点数である．$G^{(i)}$ の頂点 $v_p^{(i)}$ と $v_q^{(i)}$ の間に辺が張られている場合は $\mathbf{S}$ の pq 要素が 1，それ以外の場合は 0 である．同様に，$G^{(1)}$ と $G^{(2)}$ の間の辺を隣接行列 $\mathbf{A} \in \{0,1\}^{n^{(1)} \times n^{(2)}}$ と定義し，$\mathbf{A}^{(12)} = \mathbf{A}$，$\mathbf{A}^{(21)} = \mathbf{A}^{\mathrm{T}}$ と表記する．ネットワークは重み無しの無向グラフであり，$\mathbf{S}^{(i)}$ と $\mathbf{A}$ は対



(A)　入力ネットワークの例と頂点のクラスタリング　　(B) ネットワーク(A)の概要構造と概要構造のアラインメント

図 1: 入力するネットワークと解析結果の例

象行列である．$G^{(i)}$ を k 個のクラスタ $C_1^{(i)}, \ldots, C_k^{(i)}$ に分類し，これを行列 $\mathbf{C}^{(i)} \in \{0,1\}^{n_i \times k}$ で表す．$G^{(i)}$ の頂点 $v_p^{(i)}$ が $C_q^{(i)}$ に属しているとき，$\mathbf{C}^{(i)}$ の pq 要素は 1 であり，それ以外は 0 である．

また，$\mathbf{M}_I^{(i)} = \mathbf{C}^{(i)\mathrm{T}}\mathbf{S}^{(i)}\mathbf{C}^{(i)}$，$\mathbf{M}_B = \mathbf{C}^{(i)\mathrm{T}}\mathbf{A}^{(ij)}\mathbf{C}^{(j)}$ とする．$\mathbf{M}_I^{(i)}, \mathbf{M}_B^{(ij)} \in \mathbb{R}^{k \times k}$ である．$\mathbf{M}_I^{(i)}$ の pq 要素は $C_p^{(i)}$ と $C_q^{(i)}$ の間の辺の本数を表し，$\mathbf{M}_B$ の pq 要素は，$C_p^{(1)}$ と $C_q^{(2)}$ の間の辺の本数である．これらの各行の総和が 1 となるようにノーマライズした行列を概要グラフとし，$\overline{\mathbf{M}}_I^{(i)}$ と $\overline{\mathbf{M}}_B^{(ij)}$ で定義する．それぞれの pq 要素は，$\overline{\mathbf{M}}_I^{(i)} = \mathbf{M}_{I_{pq}}^{(i)} / \sum_{r=1}^{k} \mathbf{M}_{I_{pr}}^{(i)}$，$\overline{\mathbf{M}}_B^{(ij)} = \mathbf{M}_{B_{pq}}^{(ij)} / \sum_{r=1}^{k} \mathbf{M}_{B_{pr}}^{(ij)}$ である．また，$\tilde{\mathbf{M}}^{(ij)}$ を対角要素が $\overline{\mathbf{M}}^{(ij)}$ と等しく，それ以外の要素が 0 の行列とする．

2.1　指標の定義

行列 $\mathbf{G}$，$\mathbf{C}$，$\mathbf{M}$ をそれぞれ，$\begin{pmatrix} \mathbf{S}^{(1)} & w_1\mathbf{A} \\ w_1\mathbf{A}^{\mathrm{T}} & \mathbf{S}^{(2)} \end{pmatrix}$，$\begin{pmatrix} \mathbf{C}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{(2)} \end{pmatrix}$，$\begin{pmatrix} \overline{\mathbf{M}}_I^{(1)} & w_1\overline{\mathbf{M}}_B^{(12)} \\ w_1\overline{\mathbf{M}}_B^{(21)} & \overline{\mathbf{M}}_I^{(2)} \end{pmatrix}$ とし，クラスタ分類を表す行列 $\mathbf{C}$ の良さを表す指標を定義する．w_1 は任意のパラメータであり，ネットワークをつなぐ辺に対する重要度である．概要グラフがネットワークを誤差なく正確に表しているとき，行列の積 $\mathbf{GC}$ と $\mathbf{CM}$ は一致することから，ネットワークと概要グラフの差を表す指標を次式で定義する．

$$\Delta_S(\mathbf{C}) = Dist(\mathbf{GC}, \mathbf{CM}) \tag{1}$$

$\overline{\mathbf{M}}^{(1)} = \overline{\mathbf{M}}^{(2)}$ と $\overline{\mathbf{M}}_B^{(ij)} = \tilde{\mathbf{M}}_B^{(ij)}$ を満たすとき，二つの概要グラフは構造が全く等しく，クラスタ同士も対応の取れた完全なアラインメントと言える．このことから，概要グラフのアラインメントの良さを表す指標を次式で定義する．

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

本研究では，行列 $\mathbf{X}$，$\mathbf{Y}$ の距離 $Dist(\mathbf{X}, \mathbf{Y})$ をベク

トル $\mathbf{x}_p$， $\mathbf{y}_p$ のコサイン距離を用いて定義する．$\mathbf{x}_p$，$\mathbf{y}_p$ の q 要素はそれぞれ，$\mathbf{X}$， $\mathbf{Y}$ の pq 要素であり，$Dist(\mathbf{X},\mathbf{Y})$ は全ての行のコサイン距離の平均とする．

Δ_S と Δ_A で $\mathbf{C}$ の良さを表す指標を次式で定義する．

$$J(\mathbf{C}) = \Delta_S(\mathbf{C}) + w_2\Delta_A(\mathbf{C}) \tag{3}$$

w_2 は任意のパラメータであり，概要グラフのアラインメントに対する重要度を表す．$J(\mathbf{C})$ は，0 に近いほど概要グラフのアラインメントが良いことを示す．

$J(\mathbf{C})$ 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，$\mathbf{C}^{(i)\mathrm{T}}\mathbf{1} = \{n^{(i)}/k\}^{1\times k}$ とし，構成するクラスタの大きさを均等にする．$\mathbf{1}$ は全ての要素が 1 の列ベクトルである．

2.2 提案手法

$\mathbf{C}$ は離散的な行列であり，$J(\mathbf{C})$ が最小となる $\mathbf{C}$ を求めることは難しい．本研究では，k-means に基づいた手法，ALIgnment with Cluster using Edge connectivity (ALICE) を提案する．ALICE は，クラスタの重心を含む行列 $\mathbf{M}$ の計算，頂点をクラスタに分類という二つのステップを交互に行う．

行列 $\mathbf{M}$ の計算では，まず，概要グラフ $\overline{\mathbf{M}}_I^{(i)}$ と $\overline{\mathbf{M}}_B^{(ij)}$ を計算する．次に，二つの概要をアラインメントするため，$\overline{\mathbf{M}}_I^{(i)}$ と $\overline{\mathbf{M}}_B^{(ij)}$ の代わりに $\mathbf{M'}_I^{(i)} = \overline{\mathbf{M}}_I^{(i)} + w_2\overline{\mathbf{M}}_I^{(j)}$ と $\mathbf{M'}_B^{(ij)} = \overline{\mathbf{M}}_B^{(ij)} - w_2(\overline{\mathbf{M}}_B^{(ij)} - \tilde{\mathbf{M}}_B^{(ij)})$ で $\mathbf{M}$ を構築する．$\mathbf{M}$ の p 行は，クラスタ C_p の接続性を表しており，この接続性をクラスタの重心をベクトル $\mathbf{m}_p$ で表す．$\mathbf{m}_p$ の q 要素は， $\mathbf{M}$ の pq 要素で与える．

頂点の分類では，頂点とクラスタの重心の距離を計算し，距離が最小となるクラスタに頂点を分類する．$\mathbf{GC}$ の p 行は頂点 v_p の接続性を表しており，これをベクトル $\mathbf{v}_p$ で表す．$\mathbf{v}_p$ の q 要素は， $\mathbf{GC}$ の pq 要素で与える．v_p と C_q の重心の距離は，ベクトルのコサイン距離 $CosDist(\mathbf{v}_p^{(i)}, \mathbf{m}_q^{(i)})$ を含む次式で算出する．

$$\begin{aligned} Dist(\mathbf{v}_p^{(i)}, \mathbf{m}_q^{(i)}) = CosDist(\mathbf{v}_p^{(i)}, \mathbf{m}_q^{(i)}) \\ + \frac{k}{n^{(i)}}(n_q^{(i)} - n_{orig}^{(i)}) \end{aligned} \tag{4}$$

第 3 項は，$\mathbf{C}^{(i)\mathrm{T}}\mathbf{1}$ となるよう, クラスタに属する頂点数に対するペナルティである．$n_q^{(i)}$ は $C_q^{(i)}$ に属する頂点数であり，$C_{orig}^{(i)}$ は更新前に頂点 $v_p^{(i)}$ が属しているクラスタである．

3 解析結果

線虫とショウジョウバエの遺伝子ネットワークを本手法で解析した．遺伝子ネットワークは，iRefIndex [3] で公開されているデータを利用している．線虫とショウジョウバエのネットワークの頂点数はそれぞれ，4,098 個と 5,813 個，辺の本数は 10,231 本と 24,818 本である．グラフをつなぐ辺は，HomoloGene [4] のデータを利用し，その本数は 966 本である．このネットワークを，本手法 ALICE とグラフクラスタリングで頻繁に用いられる手法 METIS [5] により 100 個のクラスタに分類し，二つの手法の結果を比較する．ALICE のパラメータ $w1$ と $w2$ はどちらも 1.0 で解析した．

3.1 概要グラフの誤差とそのアラインメントの比較

解析結果を，概要グラフとネットワークの誤差と，概要グラフのアラインメントの良さから比較する．指標 $\Delta_S Dist(\mathbf{GC}, \mathbf{CM})$ は概要グラフの誤差を表す指標であるが，今回解析したネットワークは $\mathbf{S}^{(i)}$ に対し，$\mathbf{A}$ が非常に疎である．そのため，$\mathbf{S}^{(i)}$ と $\overline{\mathbf{M}}_I^{(i)}$，$\mathbf{A}^{(ij)}$ と $\overline{\mathbf{M}}_B^{(ij)}$ の誤差をそれぞれ $Dist(\mathbf{S}^{(i)}\mathbf{C}^{(i)}, \mathbf{C}^{(i)}\overline{\mathbf{M}}_I^{(i)})$，$Dist(\mathbf{A}^{(ij)}\mathbf{C}^{(j)}, \mathbf{C}^{(i)}\overline{\mathbf{M}}_B{}^{(ij)})$ で算出し，これらの平均 Δ'_S で概要グラフの誤差を評価する．概要グラフのアラインメントの良さは，$\Delta_A(\mathbf{C})$ で評価する．

Δ'_S の値はそれぞれ，ALICE 0.458，METIS 0.530 であり，Δ_A の値は ALICE 0.399，METIS 0.494 であった．Δ'_S と Δ_A 共に ALICE の方が小さいことから，求めた概要グラフとそれらのアラインメント共に，ALICE の方が良い結果である．

3.2 生物学的な知見との比較

クラスタの結果を遺伝子の既知の機能と比較する．遺伝子の機能には，Gene Ontrogy [6] を利用した．

それぞれの種のネットワークに含まれている全遺伝子が有する機能の確率に対し，同一のクラスタに属している頂点の機能の確率について二項検定を行った．それぞれのクラスタについて最小の確率を算出し，全クラスタの平均をもとめた結果，線虫もショウジョウバエも 0.03 未満の値であった．このことから，構成したクラスタに含まれる遺伝子が，生物学的な知見にも類似した機能を果たしていることが分かった．

これらの機能を比較することで，二種の間で共通の機能や，進化の過程でそれぞれの種が獲得してきた機能の発見が見込まれる．

4 今後の課題

二つのネットワークを大域的に比較する手法 ALICE を提案した．今後は三つ以上のネットワークを同時に解析できるよう改良したいと考えている．また，ALICE は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多彩なデータに応用可能である．このようなデータを時系列に沿って比較することで，時間の経過によってネットワークが変化してきた様子を抽出できる．

参考文献

[1] Victor Spirin and Leonid A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci*, Vol. 100, pp. 12123–12128, 2003.

[2] Mohsen Bayati, Margot Gerritsen, and *et al.* Algorithms for large, sparse network alignment problems. In *Proc. of ICDM*, pp. 705–710, 2009.

[3] Sabry Razik, George Magklaras, and *et al.* irefindex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, Vol. 9, p. 405, 2008.

[4] David L. Wheeler, Tanya Barrett, and *et al.* Gene ontology: tool for the unification of biiology. *Nucleic Acids Res*, Vol. 34, pp. 173–180, 2006.

[5] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J.SCI.COMPUT.*, Vol. 20, pp. 359–392, 1999.

[6] Michael Ashburner, Catherine A. Ball, and *et al.* Gene ontology: tool for the unification of biiology. *Nat. Genet.*, Vol. 25, pp. 25–29, 2000.