科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１７　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2012～2013
課題番号：　２４６５１２２７

研究課題名（和文）トランスログ：配列相同性が高くない遺伝子の機能推定手法開発

研究課題名（英文）Translog: How estimate the function of a gene whose seuqnece is unique

研究代表者
瀬々　潤（Sese, Jun）

東京工業大学・情報理工学（系）研究科・准教授

研究者番号：４０３６１５３９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円、（間接経費）　　930,000円

研究成果の概要（和文）：本研究では，遺伝子機能を推定するために，配列情報ではなくネットワークの相同性を基準に，機能推定を行った．ネットワークの比較方法として，大域的ネットワーク比較技術ANGIEを開発した．これは，２つのネットワークが与えられた時，なるべく両者から得られるネットワークの概要構造が一致するように，それぞれのネットワークを簡略化する技術である．また，必ずしもネットワーク構造のみだと精度が高くないので，２つの頂点間に類似度を利用することも可能とした．この技術を使い，異なる種のたんぱく質相互作用ネットワークや発現量の類似性をグラフ，配列相同性を頂点間の類似度としたデータを与え，種間の比較をすることに成功した．

研究成果の概要（英文）：This study developed a method to estimate the functions of genes using similarity of network structure, not using sequence similarity which has been currently mainly used. To this purpose, we developed the method ANGIE which can compare global network structure. Given two different networks, the ANGIE tries to find a brief network structure on each network so that the two brief graphs are similar to each other. Since our initial experimental result showed that the aligned result is far from biological knowledge, we improve the ANGIE so as to give prior similarity knowledge of genes between two species, such as sequence similarity. We applied the ANGIE to yeast and drosophila networks, generated from protein-protein interactions or expression similarity, and the result suggested new functions of genes.

研究分野：　システムゲノム科学

科研費の分科・細目：　ゲノム科学・システムゲノム科学

キーワード：　ネットワーク　比較ゲノム　バイオインフォマティクス

１．研究開始当初の背景

ゲノム網羅的な遺伝子発現量の観測が可能になって 15 年以上が経過した．価格が安価になるだけでなく，対象種の増加も著しく，多様なモデル生物とその近縁種で遺伝子発現量の観測が可能になっている．また，研究開始時の時点で，従来の DNA マイクロアレイを利用した遺伝子発現量の観測は，徐々に新型シーケンサを利用して発現量と配列を求める RNA-seq へと移行しはじめていた．RNA-seq によって最も変わることは，DNA マイクロアレイで利用されていた相補鎖配列に相当するものを事前に決定する必要が無くなることで，これにより，モデル生物だけでなく，非モデル生物に対しても，マイクロアレイの様な新たな機材開発の必要なく，ゲノム網羅的な遺伝子発現の取得が可能となる．

非モデル生物の遺伝子機能予測では，多くの場合は配列を基準に機能予測を行っていた．BLAST に代表される類似配列検索ソフトウエアを用い，非モデル生物から得られた遺伝子配列に最も近い配列を機能既知の遺伝子群から検索．その機能を，得られた遺伝子配列の機能と推定する．機械学習分野ではこの推定手法は最近点分類と呼ばれる手法に分類できる．一方で必ずしも，類似配列が類似の機能を持っているという仮説は正しくない場合もある．例えば，転写因子をコーディングしているタンパク質にシノニマスな変異が起こって，DNA 結合部位に変化があった場合，配列全体に大きな変化は無くとも，機能的な分化は大きくなる可能性がある．また，配列が同一であっても上流因子の変化により機能が異なる場合もあるだろう．このように，遺伝子配列を基にした検索からでは，必ずしも正しい答えが得られるとは限らない状況である．

この問題に対して，本研究では情報科学のグラフ理論的なアプローチを取って解決を試みた．情報科学や数学で扱われるグラフ理論は，頂点と辺からなる構造を扱うものであり，旧来から研究の対象であったが，昨今の Web 技術の進化によって，大規模なグラフ構造とグラフの特徴量の研究が盛んに行われている．Web を用いたグラフとは，各 Web ページを頂点，ページ間のリンクを辺とした構造であり，特定の社内のページや，世界中の Web ページを利用し，有益な構造を導き出す研究である．遺伝子発現量情報同様，2000 年前後から発展し，本研究開発当初には，一定の成熟した分野として確立されていた．また，遺伝子情報に対してもグラフ理論は応用され，タンパク質を頂点とし，実験的に発見された相互作用を辺とするタンパク質相互作用ネットワーク(PPI)，あるいは，遺伝子を頂点とし，遺伝子発現量の変化が類似の遺伝子間に辺を引いた共発現ネットワークなど，グラフ理論の生命情報への応用が進んでいた．

２．研究の目的

このような背景のもと，本研究の目標は遺伝子同士の配列類似性が高くなくても，類似の機能を有している遺伝子群を，遺伝子発現量の変化を基にしたグラフ構造を構築し，比較をすることで明らかにすることである．

前述のとおり，DNA シーケンサの安価，高速化により被モデル生物の遺伝子配列や発現量が明らかになっている．一方で，配列が分かっても相同性の高い機能既知遺伝子が無いと，その遺伝子の機能が推定できず，生命活動の理解が広がらない問題点が浮き彫りになっていた．この問題は，次世代シーケンサを用いて，非モデル生物の解析が進む中で，更に大きな問題となると考えられる．本研究では，DNA マイクロアレイや次世代シーケンサを利用して得られた，異なる種間の遺伝子発現量情報を基に，種を超えた発現環境の類似性を発見し，それを基に遺伝子機能を推定する「トランスログ」（トランスクリプトームのホモログ）を提案し，その解析方法を構築した．

３．研究の方法

本研究は２つの段階に分けられる．第一にトランスログ計算方法の開発．次に，その計算方法によって推定された遺伝子機能の確認である．

第一のトランスログの計算のために，グラフの大域的な比較手法の開発を行った．グラフ構造の比較では，一般に局所的なグラフ構造の比較が行われる．たとえば，特定の部分グラフがいずれのネットワークにも含まれているかの調査や，各頂点が隣接している頂点の数（次数）が類似しているかなど，ネットワークの探索，あるいは，局所的な統計量の比較によって，グラフ間を比較するものである．

これらの方法をトランスログの発見に結びつけようとすると必ずしもうまくいかない．第一に，頂点に相当する遺伝子に関して種を越えた類似性を求める事が目標であり，上記のグラフ検索や，次数の類似性の調査も，対応するノードが存在することを仮定している．つまり，配列を用いてホモログ遺伝子が決定している必要がある．これは当初の仮定に反するので，利用することができない．

次に，グラフ構造の類似性を調べる手法として，部分グラフに類似した構造を有している事を基準に，グラフ間の類似性を調査する方法がある．この方法は，システムズバイオロジー研究において，環境間の反応の類似性を見るために利用されることがある手法だが，具体的にどの遺伝子とどの遺伝子が類似しているかを調査するには，また別の計算が必要となるので，トランスログを発見するには比較が大域的過ぎる．

以上の問題点より，前者の手法よりは大域的かつ遺伝子に仮定を少なく，後者の手法よりは，より遺伝子に着目した解析ができるた

めの手法を開発する必要があった．

本研究では，２つのグラフが与えられた時，ノードのクラスタを生成するが，その時，そのクラスタ内外の辺のパターンが類似する様に，クラスタを生成する．また，種間のホモログ関係は，信頼度を定義することが可能であり，硬いホモログ関係は利用するが，それ以外は利用しない様な使い方が可能である．

大域アラインメントの例を図 1 に示す．図 1(A)は与えられたグラフを示しており，Type1 と 2 は，それぞれ別の種から得られた情報である．各頂点は遺伝子，実線による辺は PPI や共発現で得られた同一種内の情報である．破線は異なる種間で既知のホモログ遺伝子である．いくつかの遺伝子にはホモログが無く，情報に欠落がある状況であることがわかる．図 1(B)は本手法で求まるネットワークである．灰色の固まりはクラスタを示し，実線と破線はクラスタ間の関係を示す．例えば，頂点 1 と 4 は同じクラスタに入っており，これらのグループは，Type1 の C2，C3 と接続，また，Type2 の C1 と接続していることが推察される．図 1(A)に戻ると，頂点 1,4 共に Type1 C2 内の頂点に辺を張っている．また，C3 へは頂点 4 のみが，Type2 の C1 には頂点 1 のみが辺を張っている．また，頂点 1 と頂点 4 は互いに接続している．このように，情報としては必ずしも完璧ではないが，全体として見た場合，クラスタ内の頂点は，互いに同じクラスタに辺を張る確率が高い状況になっている．このようなクラスタを見つけることで，頂点 4 は，恐らく頂点 1 や Type2 の C1 に類似した機能を有している可能性があると推定できる．

以上の手法を定量的に計算できるように，EM アルゴリズム的な手法を用いたアルゴリズム ANGIE (Aligning Networks Globally with Interconnect Edges) を開発した．ANGIE では，3 種類のグラフを入力とする．２つは種内の遺伝子の類似情報を表すグラフ(図 1 内の実線．Type1 と Type2)．もう一つは，種間の隣接情報（図 1 内の破線で表されるグラフ）である．

C をクラスタ，F を各ノードの特徴ベクトル，M を与えられたグラフに関する情報として，現在のクラスタ候補と特徴ベクトルの関係を，以下の J で表した．

$$J(\vec{C}) = Dist\left(\vec{F}, \vec{C}\left(\bar{\vec{M}}_{\mathrm{I}} + \alpha\begin{pmatrix} 0 & \bar{\vec{M}}_{\mathrm{B}} \\ \bar{\vec{M}}_{\mathrm{B}}^{\mathrm{T}} & 0 \end{pmatrix} + \beta\begin{pmatrix} \bar{\vec{M}}_{\mathrm{I}}^{(2)} & \bar{\vec{M}}_{\mathrm{B}}^{\mathrm{T}} \\ \bar{\vec{M}}_{\mathrm{B}} & \bar{\vec{M}}_{\mathrm{I}}^{(1)} \end{pmatrix}\right)\right),$$

ここで α と β は重みであり，α が大きいとホモログ間に張られたエッジの重要性が高く，β が大きいとクラスタ間のエッジを信用し，ノイズを許さなくなるパラメータである．

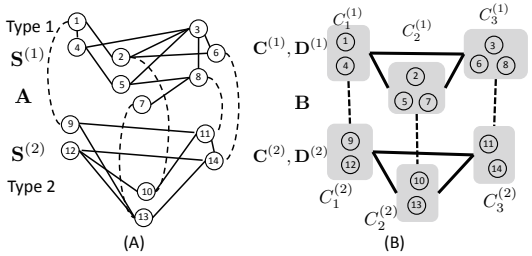

図 1. 提案する大域グラフアラインメントの例．(A)与えられたグラフ．(B)求められる大域アラインメント

ANGIE では，この J(C)を最小化するために，始めにランダムにクラスタに属する頂点 C を決め，そこから J(C)が最小化するように値を更新．収束するまで（あるいは，一定回数終了するまで)，C を更新し続けるアルゴリズムを取っている．

４．研究成果

本提案アルゴリズムにおいてパラメータとなっている α と β の値に対する結果の影響度を計測した．

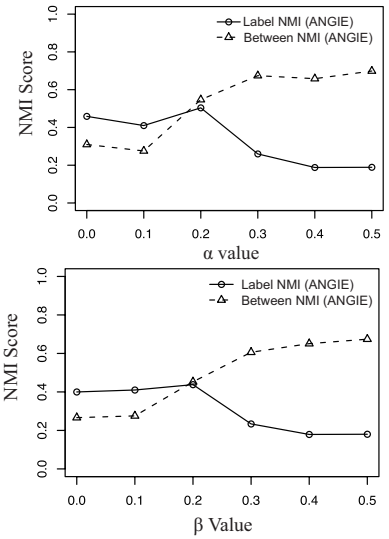

図 2. α と β の値の変化による性能変化に関する考察．

二つの擬似データを，まずクラスタ間のネットワークを作成．それを基に頂点間のネットワークを生成し，それらの辺をランダムに入れ替えることで作成した．評価軸として，クラスタ間に張られた（予め用意した際の）クラスタラベルの相同性と，辺の相同性を調べた．その結果を図 2 に示す．図 2 の結果より，まず，一致度を表す NMI(Normalized Mutual Information)スコアは 0.5 前後であり，ホモログ間の辺の数が少ないにも関わらず，ある程度互いに一致する結果が得られていることが分かる．次に，少なくとも擬似データにおいては，α（ホモログ間の信頼度）や β（クラスタ間の辺の信頼度）を増加させると辺の一致度は上がるが，クラスタの性能は低下す

ることが判明した．これは，擬似データを作成した際に，ノイズとして辺の入れ替えを行っていることに対応しており，予想通りαやβが大きいと，ノイズに弱くなる傾向を示した．

次に，実データでの結果を調査するために，線虫とショウジョウバエのPPIを基にしたネットワークを利用し，ANGIEを用いて比較を行った．その結果が図3である．ショウジョウバエは897遺伝子，1,855相互作用，線虫は478遺伝子,732相互作用のデータである．ホモログ関係として,KEGGに含まれている線虫-ショウジョウバエ間のホモログ関係を利用した．これらのホモログには57の機能グループが関連づいていた．また，クラスタ数は40とした．アラインメントを得る際に，αとβを決める必要があるが，ここではそれぞれ0.1刻みに動かし，NMIの平均値が最大となるようなαとβを選択した．

図3の結果から言えることは，まずホモログ関係の存在するクラスタ（図３中央）と，ホモログ関係の存在しないクラスタ（図３両脇）が明確に存在していることである．その原因を調べるため，機能を調べてみると，ホモログ関係のあるクラスタには,MAPKパスウエイなど基本代謝や種を越えて保存が知られている機能に関連したものが多く，両脇のクラスタには種固有の機能が関連したものが多かった．これらの結果から，ANGIEによるクラスタリング結果を精査することで，機能未知の遺伝子に対して機能の示唆を与える事ができる可能性が得られた．

同様の実験としてPPIネットワークではなく遺伝子の共発現情報を利用して作成したネットワークを基にANGIEを実行し，結果を求めた．これは，ショウジョウバエにおけるモデル生物のDrosophila melanogasterの近縁6種のマイクロアレイから得られた遺伝子

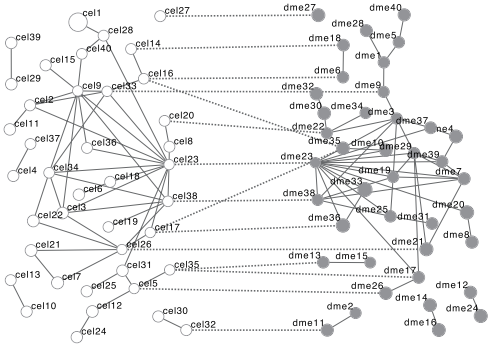

図 3. 線虫とショウジョウバエから得られたネットワークの大域的アラインメント結果．白丸が線虫，黒丸がショウジョウバエである．実線はクラスタ間につながりが深いこと，破線はホモログ関係が多いことを示している．

発現情報を基に，進化過程において類似の発現変動を示す遺伝子間に辺を貼り，ANGIEによって比較を行った．

その結果，種間のクラスタ間のグラフが非常に似た形状のものが得られ，ショウジョウバエ種間の変化は必ずしも大きくない傾向が示された．その一方で，クラスタ間の構造が異なる場所を調べると，成長やセルサイクルに関連する因子が多かった．

これは，実験ではどのショウジョウバエにおいても発生から同じ時間で計測しているが，実際には発生の速さに違いがあり，結果として異なる発生のタイミングで調査された事を示していると考えられる．

以上より，大域的なグラフアラインメント手法を提案しているANGIEは，生物学的な意味を捉えられるネットワーク解析手法である．これらのクラスタの詳細を調べることで,当初目的であったトランスログの発見へと結びつけることが可能である．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

1. Izawa A, Sese J. 2013. RECOT: a tool for the coordinate transformation of next-generation sequencing reads for comparative genomics and transcriptomics. Source Code Biol Med 8: 6.（査読有り）

2. Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. 2013. Statistical significance of combinatorial regulations. PNAS 110: 12996-13001.（査読有り）

〔学会発表〕（計１件）

1. Terada A, Sese J. 2012. Global Alignment of Protein-Protein Interaction Networks for Analyzing Evolutionary Changes of Network Frameworks. 4th International Conference on Bioinformatics and Computational Biology (BICoB-2012), pp. 196-201. Las Vegas, Nevada, USA. March 12 - 14, 2012.（査読有り）

〔図書〕（計０件）
〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
なし

６．研究組織
(1)研究代表者
瀬々 潤（Sese, Jun）
東京工業大学 大学院情報理工学研究科

准教授
　研究者番号：40361539
(2)研究分担者
なし
(3)連携研究者
なし