

一般口演 | バイオインフォマティクス

一般口演20

バイオインフォマティクス

2019年11月24日(日) 09:00 ~ 10:30 F会場 (国際会議場 3階中会議室302)

[4-F-1-03] 1細胞遺伝子発現時系列データの遺伝子相関ネットワークによる分析

○浅野 泰仁¹、小倉 淳²、七野 成之³、上羽 悟史³、松島 綱治³ (1. 東洋大学, 2. 長浜バイオ大学, 3. 東京理科大学)
キーワード : Single-cell transcriptome, Time sequence, Network analysis

慢性炎症の予防には、未病状態(自覚症状を伴わない組織病変)の検出と、その時点での早期対処が重要であると考
えられている。そのためには未病のメカニズム解明が求められるが、未だにその全容は明らかになっていな
い。一方、近年確立された包括的1細胞遺伝子発現解析技術によって、これまで細胞群単位でしか得られなかつた
遺伝子発現データが1細胞単位で取得できるようになった。この1細胞単位の遺伝子発現データの分析のため
に、ノイズ除去を行う MAGICや、擬似的に時系列分析を行う Monocle等の手法も開発されてきたこともあり、包
括的1細胞遺伝子発現解析技術の様々な分野での活用が期待されている。未病のメカニズム解明のためにこの技術
を活用した研究も始まっており、メカニズム解明につながると期待される知見が得られるようになってきてい
る。本研究では、病状が進行するにつれて遺伝子の発現量の変化が観測されることから、遺伝子間の関係も病状
の進行によって変化するのではないかと考え、それを分析するための遺伝子相関ネットワークとその可視化手法
を提案する。具体的には、1細胞単位の遺伝子発現の時系列データを入力とし、各時点の発現データから遺伝子発
現量の相関を求め、相関の絶対値が大きい遺伝子対に辺が存在するようなネットワークを作成する。さら
に、ネットワークの変化を捉えやすくするために、全時点に対して遺伝子の位置を統一した可視化を行う。この
ネットワークの時系列変化を、グラフ理論的アプローチで分析することにより、未病の検出やそのメカニズム解
明に繋がる知識が得られると期待される。その分析の手始めとして、ブレオマイシン投与による肺線維症誘導マ
ウスから作成した1細胞単位の遺伝子発現の時系列データに提案手法を適用し、病状の進行に対して遺伝子相関
ネットワークにどのような変化が見られるかを観察した。

1 細胞遺伝子発現時系列データの遺伝子相関ネットワークによる分析

浅野 泰仁^{*1}、小倉 淳^{*2}、
七野 成之^{*3}、上羽 悟史^{*3}、松島 綱治^{*3}

*1INIAD(東洋大学情報連携学部)、*2 長浜バイオ大学、
*3 東京理科大学

Analysis of Time Sequence Data of Single-cell Transcriptome using Gene Correlation Networks

Yasuhito Asano^{*1}, Atsushi Ogura^{*2}, Shigeyuki Shichino^{*3},
Satoshi Ueha^{*3}, Koji Matsushima^{*3}

*1 INIAD (Faculty of Information Networking for Innovation and Design, Toyo University),
*2 Nagahama Institute of BioScience and Technology, *3 Tokyo University of Science

In order to prevent chronic inflammation, it is considered important to detect a presymptomatic disease (tissue lesion without subjective symptoms) and to take early action at that time. It is desired to elucidate its mechanism for that purpose, although the complete picture has not yet been clarified. On the other hand, techniques for comprehensive single-cell transcriptome have developed in recent years. Several methods also have been proposed for helping analysis of single-cell transcriptome data, including MAGIC for noise reduction and Monocle for pseudo-time sequence analysis. Consequently, the comprehensive single-cell transcriptome techniques are expected to be utilized in various fields. We try to utilize these techniques and methods for elucidating the mechanism of presymptomatic diseases. We consider the relationship between genes might change according to the progress of a disease because the change of transcriptome data is also observed according to that progress. Therefore, we propose gene correlation networks and a method for visualizing the network in order to analyze the change using the single-cell transcriptome data. Concretely, for each matrix of time-sequence data of single-cell transcriptome matrices, we calculate the correlation of a pair of arbitrary genes, and construct a network in which an edge exists between each pair of genes having a large correlation. In addition, we visualize these networks to help analysis of their change so that we unify the position of each gene for all time points. It is expected to obtain knowledge about detection and elucidation of a presymptomatic disease by analyzing the evolution process of these networks in a view point of the graph theory. As a startup of such an analysis, we apply our proposal to the time-sequence data of single-cell transcriptome obtained from mice induced pulmonary fibrosis by bleomycin administration, and observe the evolution process of the obtained gene correlation networks according to the disease progress.

Keywords: Single-cell transcriptome, time sequence, network analysis

1. はじめに

炎症は様々な疾患の基礎となる解剖生理学的反応であり、その研究は数多く行われてきた。¹⁾ このうち慢性炎症と関連づけられる病気としては、アレルギー性疾患や自己免疫性疾患が知られているが、近年の高齢化社会にあつて問題となっているがんや糖尿病などの慢性疾患と、慢性炎症が関連していることも示唆されるようになってきた。慢性炎症の予防には、未病状態(自覚症状を伴わない組織病変)の検出と、その時点での早期対処が重要であると考えられている。そのためには未病のメカニズム解明が求められるが、未だにその全容は明らかになっていない。

一方、近年確立された包括的 1 細胞遺伝子発現解析技術によって、これまで細胞群単位でしか得られなかった遺伝子発現データが 1 細胞単位で取得できるようになった。この 1 細胞単位の遺伝子発現データの分析のために、ノイズ除去を行う MAGIC²⁾や、擬似的に時系列分析を行う Monocle³⁾等の手法も開発されてきたこともあり、包括的 1 細胞遺伝子発現解析技術の様々な分野での活用が期待されている。未病のメカニズム解明のためにこの技術を活用した研究も始まっており、メカニズム解明につながると思われる知見が得られるようになってきている。

本研究では、病状が進行するにつれて遺伝子の発現量の変化が観測されることから、遺伝子間の関係も病状の進行によって変化するのではないかと考え、それを分析するための遺伝子相関ネットワークとその可視化手法を提案する。具体的には、1 細胞単位の遺伝子発現の時系列データを入力とし、各時点の発現データから遺伝子発現量の相関を求め、相関の絶対値が大きい遺伝子対に辺が存在するようなネットワークを作成する。さらに、ネットワークの変化を捉えやすくするために、全時点に対して遺伝子の位置を統一した可視化を行う。このネットワークの時系列変化を、グラフ理論的アプローチで分析することにより、未病の検出やそのメカニズム解明に繋がる知識が得られると期待される。

その分析の手始めとして、ブレオマイシン投与による肺線維症誘導マウスから作成した 1 細胞単位の遺伝子発現の時系列データ 2 セットに提案手法を適用し、病状の進行に対して遺伝子相関ネットワークにどのような変化が見られるかを観察した。

2. 関連研究

本節では、単一細胞シーケンスデータの分析に用いられる情報学的手法のうち、本研究で用いている MAGIC²⁾の概

要を説明する。

Dijk らによって提案された MAGIC は、端的に述べれば、細胞ごとの遺伝子発現データのノイズを除去してコントラストを上げる手法である。もともと、単一細胞シーケンズデータには、ドロップアウトなどのノイズがあるため、遺伝子間の相関等の関係も曖昧になってしまうという点が指摘されてきた。そこで、MAGIC では Coifman と Lafon によって提案された拡散マップ⁴⁾を用いて各細胞の特徴量を計算し、「特徴量が似かよった細胞は、その遺伝子発現数も似る傾向がある」という仮説を用いて元の単一細胞シーケンズデータを補完する。なお、入力となる、細胞ごとの遺伝子発現データは細胞数が行数、遺伝子数が列数となる行列で与えられる。行列の各要素は、行に対応する細胞における列に対応する遺伝子の発現数となる。そして、MAGIC の出力も、同じサイズの行列となる。

以下で、MAGIC が実際に行っている処理の概要を述べる。

1. 前処理

まず、発現数を各細胞ごとに正規化する。その後、PCA(主成分分析)を適用して、次元数を削減し、各細胞を 20 から 100 次元のベクトルで表現する。
2. 細胞×細胞の affinity 行列の作成
 1. で得られた各細胞のベクトルを用いた各細胞間の距離を求め、行数・列数ともに細胞数となる、細胞間距離行列を作成する。さらにこの行列の要素となる距離に対して、adaptive Gaussian kernel を用いた非線形写像を適用する。この結果をその行及び列に対応する細胞間の affinity と呼ぶ。この操作は、簡単に言えば、線形の距離を変換してコントラストを上げたことに対応する。
3. 遷移確率行列への変換
 2. で得られた affinity 行列を対称化・正規化することで遷移確率行列の形にする。遷移確率行列とは、マルコフ連鎖で用いられる対称行列であり、各行各列の和がそれぞれ 1 になる。さらに、この行列を $t (> 0)$ 乗する。これは、元の遷移確率行列の各要素が、1 ステップのランダムウォークである点からある点へと遷移する確率を表していたのに対し、 t ステップでの遷移確率に対応する行列となる。これは、やはりコントラストを上げるのであるが、遷移確率行列をグラフの隣接行列と考えたときに、遷移確率が高い辺が密に存在する部分の近傍部分グラフの値が高まり、そうでない部分の値が低くなるような方法である。
4. 元の行列への掛け合わせ
 3. で得られた行列を、左から元の(MAGIC の入力となる)行列に掛ける。これによって得られる行列のサイズは、元の行列と同一になる。さらに、この掛け合わせによって元の行列と各要素の値のスケールが変わってしまうので、再スケーリングすることによって元の行列とスケールを合わせ、得られた行列を出力する。

本研究では、MAGIC の提案者の一人である Krishnaswamy の研究室が公開している実装⁵⁾の Python 版を用いている。

3. 遺伝子相関ネットワーク構築手法

本節では、まず、取得した細胞ごとの遺伝子発現データに

対して、提案する遺伝子相関ネットワークを構築する手法を説明する。なお、データを取得した日時の系列を (t_1, t_2, \dots, t_d) と表し(d はデータの個数、ある日時 t_i における細胞ごとの遺伝子発現データを $M(t_i)$ で表す。このデータにおける細胞数を $cell(t_i)$ 、遺伝子数を $gene(t_i)$ とすると、 $M(t_i)$ は行数 $cell(t_i)$ 、列数 $gene(t_i)$ の行列となる。

次に、時系列遺伝子発現データ $M(t_1), M(t_2), \dots, M(t_d)$ の各 $M(t_i)$ に対して得られた遺伝子相関ネットワークを、他の $M(t_j)$ に対して得られた遺伝子相関ネットワークと可視化して比較するための手法を提案する。

3.1 遺伝子相関ネットワーク

遺伝子相関ネットワークは重み付き無向グラフで表され、頂点集合は遺伝子の集合であり、辺集合は、遺伝子対の集合となる。直感的には、遺伝子発現データにおいて相関の高い遺伝子対間に辺が存在する。辺集合を構築する具体的な手法は、以下で説明する。

3.2 前処理

時系列相関ネットワークの構築にあたって、 $M(t_i)$ に現れる遺伝子のうち、発現数の少ないものを除く。これは MAGIC でも行われていることである。提案手法では、上記の MAGIC の実装に含まれる、発現数上位の x パーセント(x はパラメータ)のみを残す関数を用いて、発現数上位 1000 程度の遺伝子を残している。この残った遺伝子(その個数を $gene'(t_i)$ とする)に対応する部分行列を、 $M'(t_i)$ と表す。

3.3 MAGIC の適用と相関行列の計算

$M'(t_i)$ に 2 節で概要を述べた MAGIC を適用する。これによって得られた行列の各列は、各遺伝子の特徴ベクトルとみなすことができる。これによって、各遺伝子対の得られた相関係数を計算することが可能である。結果として、行数及び列数が $gene'(t_i)$ となり、各要素が対応する遺伝子対の相関係数となる行列 $C(t_i)$ が得られる。

3.4 辺集合の構築

ここでは、上で得られた相関行列 $C(t_i)$ から、遺伝子相関ネットワーク $G(t_i)=(V(t_i), E(t_i))$ を構築する。 $V(t_i)$ は、 $C(t_i)$ に現れる全遺伝子の集合となる。辺集合 $E(t_i)$ は、直観的にはある閾値を超えた相関係数を持つ遺伝子対からなる。言い換えると、 $u \in V(t_i), v \in V(t_i)$ の任意の遺伝子対 (u, v) に対して、 $C(t_i)$ の対応する要素すなわち u 行 v 列の値が、与えられた値の区間(または区間の集合)に含まれるならば、 $(u, v) \in E(t_i)$ となる。また、この辺の重み $w(u, v)$ はその値となる。例えば、区間 $[0.8, 1.0]$ を与えた場合は、辺集合は相関係数 0.8 以上の遺伝子対の集合となる。また、区間の集合 $\{[-1.0, 0.8], [0.8, 1.0]\}$ を与えた場合は、辺集合は相関係数の絶対値が 0.8 以上の遺伝子対の集合となる。

第 4 節では、各時系列遺伝子発現データ $M(t_i)$ について遺伝子相関ネットワークを構築した結果、得られた点集合・辺集合のサイズについてまとめている。

3.5 遺伝子相関ネットワークの可視化

時系列遺伝子発現データ $M(t_1), M(t_2), \dots, M(t_d)$ に対して得られた遺伝子相関ネットワークの系列 $G(t_1), G(t_2), \dots, G(t_d)$

を可視化して分析するためには、すべての遺伝子相関ネットワークにおいて、同じ遺伝子が同じ場所に配置された方が良い。仮に同じ遺伝子の場所が、異なる2時点 t_i, t_j でのネットワーク $G(t_i), G(t_j)$ のそれぞれの可視化において異なってしまうては、例えばその遺伝子に接続する辺がどう変化したのか極めて分かりづらいものになってしまうからである。

遺伝子の場所を固定する方法は様々なものが考えられるが、本研究では単純さと、場所にある程度の意味付けが可能なることを重視して、以下の可視化手法を採用することにした。

1. 遺伝子に対応する、遺伝子相関ネットワークの点の Y 座標を、その遺伝子が存在する染色体の番号に対応させる。具体的には、1 番染色体に属する遺伝子が可視化結果の一番上に配置され、上から順に等間隔で 2 番染色体、3 番染色体、...、22 番染色体、X 染色体、Y 染色体となるように配置される。図 1 は可視化結果であるが、各点は緑色の円で表されている。一番上の点列が 1 番染色体に対応する。図 1 を上から下に見ていったときに、点列が最初は垂直方向に等間隔に並んでいるが、最下部の周辺では垂直方向に大きな空間が空いているのがわかる。これは、20-22 番染色体に属する遺伝子がこのネットワークに含まれていなかったことを意味する。
2. 次に、点の X 座標を、その染色体内の遺伝子の位置に対応させる。遺伝子の染色体内の開始点が先頭に近い方が図中の左に配置され、末尾に近い方が右に配置される。図 1 を上から下に見ていったときに、点列が最初は水平方向に長く、しだいに短くなっているのは、1 番染色体が最も長く、22 番染色体に至るまで、基本的にはだんだん短くなっていることに対応する。
3. 遺伝子対 (u, v) が辺集合に含まれるときは、 u, v に対応する図中の点間を結ぶ線分でこれを表す。今回は重み $w(u, v)$ が非負の時は青色、負の時は赤色で表現した。図 1 は区間の集合 $\{-1.0, -0.8\}, [0.8, 1.0\}$ に対して得られた遺伝子相関ネットワークの可視化であるが、赤い辺(青い辺と重なった部分は黒く見えている)より青い辺が圧倒的に多い、すなわち負の相関よりも正の相関を持つ遺伝子対が圧倒的に多いことがわかる。辺の重みの絶対値の大小によって線分の太さを変えることも可能である。

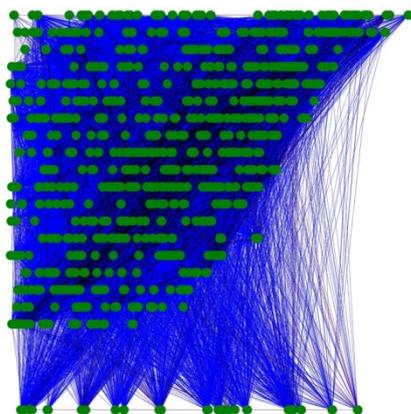


図 1 遺伝子相関ネットワークの可視化

第 4 節では、時系列遺伝子発現データに対して得られた遺伝子相関ネットワークをそれぞれ可視化して時系列順に比

較した結果を説明している。

4. 実験結果と考察

本節では、時系列遺伝子発現データに対して、提案した手法で遺伝子相関ネットワークを作成した結果と、それらを可視化して時系列順に比較した結果を説明し、考察を行っている。なお、データセットとしては、東京理科大学松島研究室で作成された、プレオマイシン投与による肺線維症誘導マウスの単一細胞シークエンスデータを 2 セット用いた。以降、それぞれデータセット 1、データセット 2 と呼ぶことにする。これらのデータセットはそれぞれ、 t_1 :0 日目、 t_2 :3 日目、 t_3 :7 日目、 t_4 :10 日目でデータ取得を行ったものである。各遺伝子発現データ $M(t_1), M(t_2), M(t_3), M(t_4)$ はそれぞれ約 2500 から約 7000 の細胞と約 25000 から約 30000 の遺伝子からなる。

4.1 遺伝子相関ネットワークの作成結果

表 1 から 4 は、与えた相関係数の値の区間 $[0.8, 1.0]$, $[0.95, 1.0]$, $[-1.0, -0.6]$, $[-1.0, -0.8]$ のそれぞれについて得られた遺伝子相関ネットワークの辺数をまとめたものである。点数は、どのデータであっても 3 節で説明したように発現数上位の約 1000 点を選んでいるためほとんど差がないので、ここでは省略した。上記の区間を選んだ理由は、正の相関係数を持つ辺の数が相対的に非常に多く、負の相関係数を持つ辺が相対的に少なかったため、正の区間は狭めにしたということと、絶対値 0.6 未満の辺に強い相関関係があるとは言いにくいことである。なお、それぞれの表には、遺伝子相関ネットワークの作成の際に MAGIC を適用した場合と、適用しなかった場合双方の結果を記してある。

表 1 区間 $[0.8, 1.0]$ に対する遺伝子相関ネットワーク辺数

	t_1	t_2	t_3	t_4
データセット 1 (MAGIC 有)	26,171	50,577	37,624	35,393
データセット 1 (MAGIC 無)	68	66	60	66
データセット 2 (MAGIC 有)	19,779	21,743	27,869	24,706
データセット 2 (MAGIC 無)	86	77	70	68

表 2 区間 $[0.95, 1.0]$ に対する遺伝子相関ネットワーク辺数

	t_1	t_2	t_3	t_4
データセット 1 (MAGIC 有)	7,359	8,705	5,489	7,477
データセット 1 (MAGIC 無)	19	26	3	14
データセット 2 (MAGIC 有)	6,172	6,254	4,111	4,744
データセット 2 (MAGIC 無)	9	8	3	9

表 3 区間[-1.0, -0.6]に対する遺伝子相関ネットワーク辺数

	t_1	t_2	t_3	t_4
データセット 1 (MAGIC 有)	5,060	35,183	22,330	9,674
データセット 1 (MAGIC 無)	0	0	0	0
データセット 2 (MAGIC 有)	3,500	10,897	12,798	11,550
データセット 2 (MAGIC 無)	0	0	0	0

表 4 区間[-1.0, -0.8]に対する遺伝子相関ネットワーク辺数

	t_1	t_2	t_3	t_4
データセット 1 (MAGIC 有)	8	1,352	1,063	111
データセット 1 (MAGIC 無)	0	0	0	0
データセット 2 (MAGIC 有)	10	80	214	191
データセット 2 (MAGIC 無)	0	0	0	0

表 1 から 4 より、MAGIC を適用することによって辺の数が大幅に増えていることが分かる。これは、MAGIC を適用する前に見つからなかった相関関係が、数多く見つかるようになったことを意味する。特に、負の相関に関しては、MAGIC 適用前は全く見つかっていなかったが、MAGIC を適用することによって見つかるようになったことが注目される。

正の相関を表す辺の数は、時系列を追って見たところ特徴的な変化を見つけることが難しかったが、負の相関を表す辺の数の時系列変化については、特徴的なものが見つかったため、次の可視化の結果と共に考察することにする。

4.2 遺伝子相関ネットワークの可視化結果

ここでは、4.1 節で作成した遺伝子相関ネットワークを可視化した結果について説明する。上で説明したように、今回は負の相関を表す辺の数の時系列変化が特徴的であったと考えられるので、表 4(MAGIC 有の部分)に対応する、すなわち区間[-1.0, -0.8]に対応する遺伝子相関ネットワークの可視化結果を提示する。



図 2 遺伝子相関ネットワーク(データセット 1, t_1)

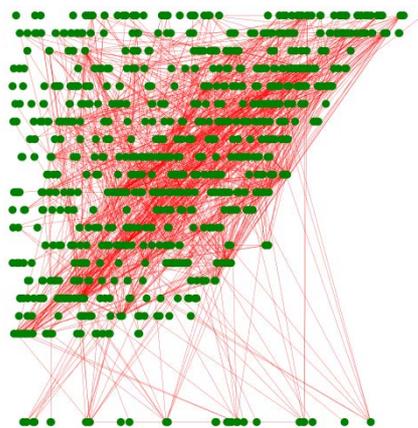


図 3 遺伝子相関ネットワーク(データセット 1, t_2)

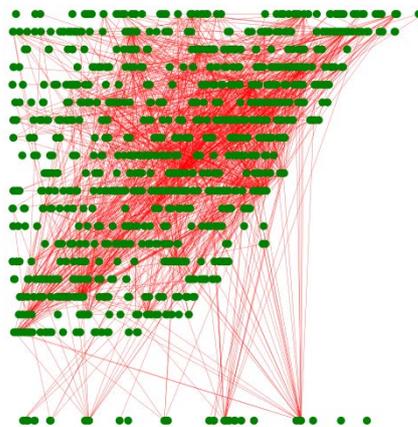
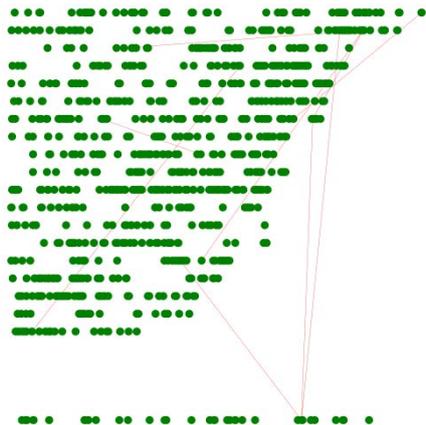
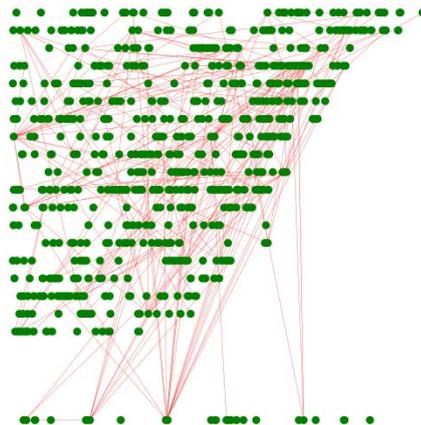
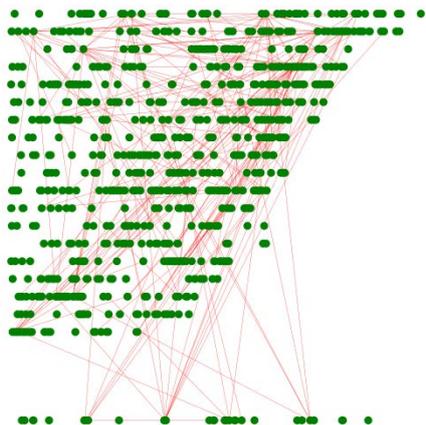
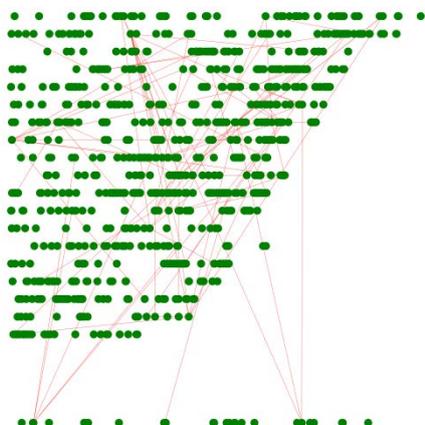


図 4 遺伝子相関ネットワーク(データセット 1, t_3)



図 5 遺伝子相関ネットワーク(データセット 1, t_4)

図6 遺伝子相関ネットワーク(データセット2, t_1)図9 遺伝子相関ネットワーク(データセット4, t_4)図7 遺伝子相関ネットワーク(データセット2, t_2)図8 遺伝子相関ネットワーク(データセット3, t_3)

以上の図から、負の相関の辺は健常時(t_1)には極めてわずかであることがわかる。その中では、2番染色体の遺伝子を含む辺が比較的多いことが見て取れる。その後(t_2 , t_3)は、病状の進行と共に負の相関の辺は急激に増加する。特に、データセット1ではその傾向が顕著である。そしてさらにその後炎症が定着したと考えられる時点(t_4)では、負の相関は減少している。データセット1では特にその傾向は顕著であり、データセット2では減少はわずかである。2個のデータセットだけでは確実なことは無論言えないが、この負の相関の一過性の増加は、病気の進行で細胞機能の転換が起きている可能性もある。

5. むすび

本論文では、1細胞遺伝子発現データから構築可能な遺伝子相関ネットワークと、その時系列変化を分析するための可視化手法を提案した。今後の課題は、さらに数多くのデータセットについて提案手法を適用し、今回同様の観測結果が得られるか検証すること、PageRankのようにネットワーク上の重要な点(遺伝子)を発見する手法を適用して病気に関係することが知られている遺伝子の働きの関係を調査すること等が挙げられる。それによって、未病と関係する遺伝子等が突き止められれば、未病予防にも有用な知見が得られると期待される。

参考文献

- 1) 松島綱治, 上羽悟史, 七野成之, 中島 拓弥. もっとよくわかる! 炎症と疾患～あらゆる疾患の基盤病態から治療薬までを理解する (実験医学別冊 もっとよくわかる! シリーズ). 羊土社, 2019.
- 2) Dijk D. et al. Recovering gene interactions from single-cell data using data diffusion. Cell, 2018; 174(3): 716-729.
- 3) Qiu X, Hill A, Packer J, Lin D, Ma Y, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nature Methods, 2017; 14: 309-315.
- 4) Coifman R, Lafon S. Diffusion maps. Applied and Computational Harmonic Analysis, 2006; 21(4): 5-30.
- 5) Krishnaswamy Lab, GitHub. [https://github.com/KrishnaswamyLab (cited 2019-Aug-30)].