独立成分分析を用いた外生的発現遺伝子同定解析

Identification of exogenously expressed genes by applying independent component analysis

十河 泰弘[*1] 清水 昌平[*1] 鷲尾 隆[*1] 井元清哉[*2]
Yasuhiro Sogawa Shohei Shimizu Takashi Washio Seiya Imoto

[*1]大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka Univ.

[*2]東京大学 医科学研究所
The institute of Medical Science, The University of Tokyo

Many statistical methods have been proposed to estimate causal models in classical situations including fewer variables than observed instances. However, gene expression analysis usually has less observed instances than variables, and thus is an ill-defined problem. Therefore, the existing statistical methods are hardly applied to identification of causal relationships in gene networks . On the other hand, a method named 'eggFinder' developed in our laboratory can find exogenous variables based on non-Gaussianity under situations with orders of magnitude more variables than observed instances. In this paper, we apply this method to three gene expression datasets, analyze the results in terms of genes activated by external causes, and provide their associated insights.

1. 現状課題と研究目的

Simon, Blalock らによる因果推論 [Simon 53, Blalock 61] や，Wright によるパス解析 [Wright 21], TETRAD[Glymour 87][*1] といった従来の因果推論は，多数の事例から少数の変数間の因果関係の推定を行うためのものであるため，コストや倫理面から多数事例の用意が困難で，かつ大量の遺伝子発現変数を含む遺伝子発現データへの適用は難しい．従って，従来手法により遺伝子発現データから遺伝子間発現の因果関係ネットワークを推定することは困難である．しかし，完全な遺伝子間発現因果関係の推定は困難でも，その因果関係について何かしらの情報を得ることができれば，新薬の開発等に役立てることができる．一方で，当研究室で開発された Exogenous generating variable finder(eggFinder) は大量変数間の因果関係における大元の原因となる変数 (外生変数) の集合を，少数事例から同定できる．本研究では，この eggFinder を遺伝子発現データに適用し，少数事例から遺伝子間発現ネットワークにおける因果の根源となる発現遺伝子を同定し，その結果について考察を行う．

2. 因果モデルと eggFinder

因果モデルとは，図 1 のような構造を持つ変数間依存関係ネットワークである．x_i は観測変数，e_i は外的要因変数を表す．外的要因変数は共通した祖先変数を持たず互いに無相関であり，またその変数が従う分布の分散は 0 でなく非ガウス性を有すると仮定する．更に，対象とするモデルが，線形でなおかつ非巡回構造を持ち，Faithfulness[Spirtes 93] が成立するという仮定の下で，eggFinder は適用可能である．Faithfulness とは，同じ祖先変数を持つ子孫変数は必ず相関を持つということである．この因果モデルにおいて，図 1 の x_1 や x_2 のように外的な影響のみを受ける変数を外生変数という．その下流にある変数は，様々な分布に従う確率変数を足し合わせるとその変数の分布はガウス分布に近づく，という中心極限定理からガウス性が強くなる．従って，外生変数の方が一般に非ガウス性が強いと考えられる．eggFinder では，これらの原理に基づいて外生変数の同定を行う．同定を行う際に，非ガウス性の尺度 (ネゲントロピー) として式 (1) を用いる．

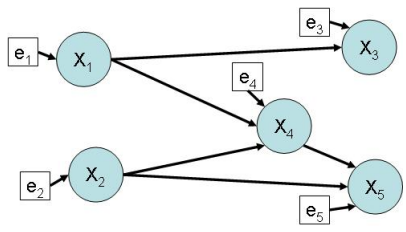

図 1: 因果モデルの例

$$J(\boldsymbol{x_i}) = [E\{G(\boldsymbol{x_i})\} - E\{G(\boldsymbol{z})\}]^2 \quad (1)$$

ただし，本研究では文献 [Hyvärinen 99] に倣い，$G(s) = -exp(-s^2/2)$ とする．$\boldsymbol{z}$ はガウス性変数を表す．この式の前の項は観測変数が従う分布のエントロピーを，後ろの項はガウス分布のエントロピーを表し，その差を取って二乗することで，観測変数が従う分布がガウス分布とどれほど異なっているかを示す．

3. eggFinder のアルゴリズム

以下に eggFinder 全体のアルゴリズムを示す．

1. 最初に対象とする因果モデル内の全観測変数の集合を V_x とする．そして，同定の結果得られる外生変数の集合を $E = \phi$，E 内の変数のいずれとも相関を持たない観測変数の集合を $U_E^{(1)} = V_x$，$m = 1$ とする．
2. 同定された何れかの外生変数と他の全ての観測変数が相関を持つまで，すなわち，$U_E^{(m)} = \phi$ になるまで以下の手順を繰り返す．
   (a) 式 (1) を用いて，U_E 内の観測変数の中から $x_m = argmax_{x \in U_E^{(m)}} J(x)$ となる x_m を探索する．

連絡先: 十河 泰弘
大阪大学 産業科学研究所 知能推論研究分野
〒 567-0047 大阪府茨木市美穂ケ丘 8-1
e-mail:sogawa@ar.sanken.osaka-u.ac.jp

*1 http://www.phil.cmu.edu/projects/tetrad/index.html

(b) $U_E^{(m)}$ 内の変数のうち，上記 (a) で探索した最も非ガウス性の強い観測変数 x_m を外生変数として E に加える．すなわち，$E = E \cup \{x_m\}$ とする．

(c) E 内の各変数と $U_E^{(m)}$ 内の各変数のそれぞれで検定手法 1 による相関性の検定を行い，$m = m + 1$ として，E 内の何れの変数とも相関を持たない観測変数の集合を新たに $U_E^{(m)}$ とする．

検定手法 1 (ピアソンの積率相関係数の検定) 変数 X, Y の値のペア N 個に対して，(X, Y) 間の相関係数 γ_{XY} は下式で表される．但し，x_i, y_i はそれぞれ i 番目の事例における変数 X, Y の値であり，$\overline{x}, \overline{y}$ はそれぞれ x, y の平均値を表す．

$$\gamma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

相関係数 γ_{XY} に対してその相関係数が 0 であるかどうかの検定を行う．

1. 前提
   - 帰無仮説 H_0：「相関係数 $= 0$」
   - 対立仮説 H_1：「相関係数 $\neq 0$」
   - 有意水準 α で両側検定を行う．
2. 次式で検定統計量 t_0 を計算する．

$$t_0 = \frac{|\gamma_{XY}|\sqrt{n-2}}{\sqrt{1-\gamma_{XY}^2}} \tag{2}$$

3. t_0 は，自由度が $n-2$ の t 分布に従う．
4. 有意確率を P =Pr$\{|t| \geq t_0\}$ とする．
5. 帰無仮説の採否を決める．
   - $P > \alpha$ のとき，帰無仮説を採択する．
   - $P \leq \alpha$ のとき，帰無仮説を棄却する．

手順 2(c) において相関性の検定を行うが，その際，False Discovery Rate(FDR) が 5%となるように統制を行う．上記のアルゴリズムのように検定を複数回繰り返すと，帰無仮説が実際には真であるのに棄却してしまうという誤りが起こる確率が有意水準より大きくなってしまう．これを多重比較の問題といい，このようなことが起こらないように,「検定の結果，有意とみなされた結果の数のうち，実は帰無仮説が正しい結果の割合をある一定の値以下に統制しよう」という考え方が FDR である．当研究では以下の手法を用いて，FDR が 5%になるよう統制を行う．それぞれの変数間での検定の結果得られた l 個の有意確率を値の小さなものから $P_{(1)}, \cdots, P_{(l)}$ とし，FDR を q とすると，$P_{(k)} \leq \frac{k}{l}q(k = 1, 2, \cdots, l)$ を満たす最大の $P_{(k)}$ を有意水準とする．eggFinder では新たに外生変数が追加されるたびに上記の手法を用いて FDR の統制を行い，次の外生変数の同定を行う．

4. 外性的発現遺伝子同定解析

4.1 解析対象データと解析手順の概要

本研究では，National Center for Biotechnology Information [*2]に公開されている，3 つのデータについて解析を行った．以下に示す 3 つのデータは Affymetrix 社の GeneChip を用いてマイクロアレイ分析を行うことによって得られた遺伝子の発現度を示す数値データである．マイクロアレイ分析とは，基準とする標準細胞状態での遺伝子発現状態に対して，興味がある別の細胞状態での遺伝子発現状態の相対的な活性の違いを調べるものである [Washio 07].

(A) 白血病細胞に薬物投与を行ったデータ
白血病の治療薬として用いられるメトトレキサート (以下 MTX) やメルカプトプリン (以下 MP)，またその組み合わせを人の白血病細胞に投与した場合の 12600 個の遺伝子の発現度についてまとめたデータである．

(B) 乳がん細胞株にホルモン刺激を行ったデータ
乳がん細胞株 (MCF7) に成長ホルモンである上皮成長因子 (以下 EGF) やヘレグリン (以下 HRG) を様々な濃度で投与して，時系列に 22277 個の遺伝子の発現度を計測し，それらについてまとめたデータである．

(C) 悪性度に対する乳がん細胞のデータ
上記の 2 つのデータと異なり，刺激等を加えることなく，悪性度の指標である G1，G2，G3 といった乳がんの悪性度ごとに乳がん細胞の遺伝子の発現度についてまとめたデータである．

これら 3 つのデータについて，以下に説明する検定手法 2 の Welch の t 検定を用い，刺激や悪性度の違いに対して共通した発現を示す遺伝子をデータから除く処理や，刺激や悪性度の異なる標本を統合する処理を行った．

検定手法 2 (Welch の t 検定) Welch の t 検定では，帰無仮説が正しいと仮定した場合に，統計量が t 分布に従うことを利用して，標本の分散が等しくない 2 組の標本 $x_1, \cdots, x_m$ および $y_1, \cdots, y_n$(標本サイズは m および n とする) について平均に有意差があるかどうかの検定を行う．それぞれの標本平均を $\bar{X}, \bar{Y}$ とし，不偏分散を U_x, U_y とすると，検定は以下の手順によって行われる．

1. 前提
   - 帰無仮説 H_0：「2 組の標本について平均に有意差がない」
   - 対立仮説 H_1：「2 組の標本について平均に有意差がある」
   - 有意水準 α で両側検定を行う．
2. 次式で検定統計量 t_0 を計算する．

$$t_0 = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{U_x}{m} + \frac{U_y}{n}}} \tag{3}$$

3. t_0 は，以下に示す自由度 ν の t 分布に従う．

$$\nu = \frac{(\sqrt{\frac{U_x}{m} + \frac{U_y}{n}})^2}{\frac{U_x^2}{m^2(m-1)} + \frac{U_y^2}{n^2(n-1)}} \tag{4}$$

*2 http://www.ncbi.nlm.nih.gov/

表 1: 薬物刺激と標本数

グループ No.	投与した薬物	標本数
(1)	高濃度 MTX のみ	22
(2)	高濃度 MTX と MP	10
(3)	低濃度 MTX と MP	14
(4)	MP のみ	12

表 2：MTX の濃度差による外生的発現遺伝子候補

遺伝子名
31411_at
36525_at
31332_at
36824_at
36059_at

表 3：MP の有無による外生的発現遺伝子候補

遺伝子名
636_at
35492_at
32757_at
34088_at
1027_at

4. 有意確率を $P = \Pr\{|t| \geq t_0\}$ とする．

5. 帰無仮説の採否を決める．

   - $P > \alpha$ のとき，帰無仮説を採択する．
   - $P \leq \alpha$ のとき，帰無仮説を棄却する．

4.2 薬物刺激を加えた白血病細胞の遺伝子発現データの解析

まず最初に，ここでは (A) 薬物刺激を加えた白血病細胞の遺伝子発現データの解析手順について述べる．このデータでは，白血病細胞への薬物の投与パターンによって 4 種類の標本があり，パターンごとに標本数が異なる．各薬物投与パターンの標本数を表 1 に示す．ここでは，12600 個の遺伝子を対象として，発現度の測定を行っている．前処理において，投与薬物の異なる標本間で発現度が有意に違う遺伝子のみを同定解析の対象として選択するために，表 1 の 4 種類の標本のうちで，MP 投与の有無 {(1)⇔ (2)}，MTX の投与濃度の違い {(2)⇔ (3)} の条件ペアについてデータ間に検定手法 2 を適用し，それぞれ投与薬物条件によって最も発現に違いのある 1000 個の遺伝子を選択した．そして，各条件ペアで選択した遺伝子集合について，それぞれ {(1)+(2)}，{(2)+(3)} のデータへ統合した．最後に，これら 2 つのデータに eggFinder を適用して外生変数候補を同定し，前者の条件ペアについて MP の投与の有無によって直接発現する遺伝子群，後者の条件ペアについて MTX の投与濃度の違いによって直接発現する遺伝子群を導出した．

解析の結果，データ (A) からは表 2 や表 3 のような外生的発現遺伝子候補が同定された．また，検定手法 2 で選択したそれぞれの投与薬物条件における 1000 個の遺伝子集合間で共通する遺伝子の数は 134 個であった．得られた外生的発現遺伝子候補の違いや t 検定で選択した遺伝子集合間で共通する遺伝子の数が少ないことから，MP の有無と MTX の濃度差によって影響を受ける遺伝子は異なる，即ち MP と MTX という 2 つの薬品は異なる遺伝子に作用することがわかる．

4.3 ホルモン刺激を加えた乳がん細胞の遺伝子発現データの解析

次に，ここでは乳がん細胞に対して 0.1nmol，0.5nmol，1.0nmol，10.0nmol の 4 つの異なった濃度で EGF や HRG によるホルモン刺激を行い，5 分，10 分，15 分，30 分，45 分，60 分，75 分の 7 つの経過時間でその遺伝子発現を調べたデータ (B) について解析手順を述べる．このデータでは，それぞれの濃度と時間につき 1 標本のみが用意されており，22277 個の遺伝子を対象として，発現度の測定を行っている．簡略化のため，それぞれの標本を「濃度 0.1nmol の HRG を投与し，経過時間 5 分を調べた標本」ならば「H-0.1-5」という表記にする．このデータの解析においても，検定手法 2 を用いて経過時間の差によって発現度に最も有意差のある 1000 個の遺伝子を解析対象とした．表 4 に検定手法 2 を適用した標本の組み合わせについて示す．これらの遺伝子群に対し，以下の 2 通りの解析を行った．

表 4: データ (B) での検定手法 2 を適用した標本の組み合わせ

遺伝子群	検定手法 2 を用いた標本
T_1^1	{H-0.1-5,H-0.1-10,H-0.5-5,H-0.5-10} ⇳ {H-0.1-60,H-0.1-90,H-0.5-60,H-0.5-90}
T_2^1	{E-0.1-5,E-0.1-10,E-0.5-5,E-0.5-10} ⇳ {E-0.1-60,E-0.1-90,E-0.5-60,E-0.5-90}

解析 1 HRG 刺激を行った標本に対し，7 個の経過時間ごとに，0.1nmol，0.5nmol，1.0nmol，10.0nmol の濃度で計測した 4 個の発現データをまとめて 1 個の標本として扱った．そして，各濃度の標本の平均をそれぞれの標本内の各データから引いて，標本の平均を 0 にし，その 7 個の標本を合わせたものを T_1^1 の遺伝子群を対象として eggFinder に適用した．

解析 2 EGF 刺激を行った標本に対し，7 個の経過時間ごとに，0.1nmol，0.5nmol，1.0nmol，10.0nmol の濃度で計測した 4 個の発現データをまとめて 1 個の標本として扱った，そして，各濃度の標本の平均をそれぞれの標本内の各データから引いて，標本の平均を 0 にした後に，その 7 つの標本を合わせたものを T_2^1 の遺伝子群を対象として eggFinder に適用した．

解析の結果，前節と同様に様々な外生的発現遺伝子候補を得ることができた．T_1^1 と T_2^1 の間で共通する遺伝子の数は 125 個と多く存在することから，投与したホルモンに関係なく，経過時間の差によって影響を受ける遺伝子が多く存在すると考えられる．しかし，それぞれの解析の結果得られた外生的発現遺伝子候補間で共通する遺伝子はない．これらのことから，経過時間の差によって外生的に発現する遺伝子はホルモンによって異なるが，それらの遺伝子によって影響を受ける下流の遺伝子は共通したものが多いと考えられる．

4.4 乳がん細胞の遺伝子発現データの解析

最後に，薬物刺激等を加えずに，異なる悪性度の乳がん細胞の遺伝子発現を調べたデータ (C) の解析手順について述べる．このデータでは，乳がんの悪性度の低いものから G1，G2，G3 という 3 種類の標本があり，悪性度ごとに標本数が異なる．各悪性度ごとの標本数について表 5 に示す． ここでは 22683 個の遺伝子を対象として，発現度の測定を行っている．このデータの解析も，検定手法 2 を用いて各悪性度間で発現度に最も有意差のある 1000 個の遺伝子を解析対象とした．検定手法 2 を用いた標本の組み合わせについて表 6 に示す．これらの遺伝子群に対して，表 7 に示す遺伝子群と標本の組み合わせで

表 5: 悪性度と標本数

悪性度	標本数
G1	68
G2	166
G3	55

表 6: データ (C) での検定手法 2 による標本の組み合わせ

遺伝子群	検定手法 2 による標本の組み合わせ
T_1^2	G1⇔G2
T_2^2	G1⇔G3
T_3^2	G2⇔G3

表 7: 対象遺伝子群と解析標本

解析 No.	対象遺伝子	解析標本
(1)	T_1^2	G1
(2)	T_2^2	G1
(3)	T_1^2	G2
(4)	T_3^2	G2
(5)	T_2^2	G3
(6)	T_3^2	G3

解析を行った．解析の結果，前節と同様に様々な外性的発現遺伝子候補を得ることができた．ここで，検定手法 2 で選択された遺伝子群間において共通する遺伝子数を表 8 に，それぞれの解析により得られた外生的発現遺伝子候補間で共通する外性的発現遺伝子候補の数を表 9 に示す． 表 8 を見ると，T_2^2 と T_3^2 の間では 616 個の遺伝子が共通であることがわかり，各遺伝子の発現はかなり似た変化をしていることがわかる．これに対して，T_1^2 と T_3^2 の間で共通している遺伝子は 149 個と少なく，異なる変化をしていることがわかる．このことから，G3 の乳がん細胞の遺伝子発現パターンは G1 や G2 に比べて独自であると考えられる．さらに，表 9 を見ると，(5) と (6) の間では 12 個の外生的発現遺伝子候補が共通であることがわかり，他と比べると多く，安定して求まっており，表 8 に示されたのと同様に G3 が特有の外生的発現遺伝子を持つことがわかる．これに対して，G1，G2 は特有の外生的発現遺伝子を持たないことがわかる．総合的には G3 の独自性が際立っていることが伺え，このことから悪性度が高い乳がん細胞は G1，G2 とは異なった遺伝子発現をしていると考えられる．

5. 課題

今後の課題として，次の 2 つが挙げられる．1 つ目は，より多くの遺伝子発現データについて解析を行い，結果をデータベース化することである．本研究では，3 種類の遺伝子発現データを扱ったが，eggFinder による解析を適用可能なデータは大量に存在する．今後はそれらの解析結果をデータベース化することで，専門家が容易にアクセス可能な仕組みを作る必要がある．

2 つ目は，社会学などの他分野のデータを eggFinder に適用することである．これにより，より広範な分野に本研究の成果を還元できる．

6. 結論

従来の統計的因果推論を用いて，小規模事例しか持たず，かつ大規模次元データである遺伝子発現データから遺伝子発現に関する因果推論を行うことは困難であった．そこで，本研究では少数事例からなる大規模次元データの外生変数の同定をすることができる eggFinder を用いて，様々な遺伝子発現データの解析を行った．この解析によって従来は，少数事例からでは導出困難と思われた遺伝子発現に関する知見を得ることができた．

表 8: 各遺伝子群間で共通する遺伝子の数

	T_1^2	T_2^2	T_3^2
T_1^2	-	351	149
T_2^2	351	-	616
T_3^2	149	616	-

表 9: 各解析の外生的発現遺伝子候補間で共通する候補数

	(1)	(2)	(3)	(4)	(5)	(6)
(1)	-	2	1	0	0	0
(2)	2	-	0	1	1	1
(3)	1	0	-	0	1	0
(4)	0	1	0	-	0	0
(5)	0	1	1	0	-	12
(6)	0	1	0	0	12	-

謝辞

筆者等が本研究を行うにあたり，助言を頂いた大阪大学 産業科学研究所 高次推論方式研究分野 猪口 明博 助教に感謝する．

参考文献

[Glymour 87] Glymour, Clark., Richard Scheines, Peter Spirtes, Kevin Kelly. Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. Academic Press, 1987.

[Wright 21] Wright, S.Correlation and causation. Journal of Agricultural 20 pp.557-585,1921.

[Spirtes 93] Spirtes, P., C. Glymour, R. Scheines. Cusation, Prediction, and Search. Springer Verlag, 1993

[Washio 07] 鷲尾隆，樋口知之，井元清哉，玉田嘉紀，佐藤健，元田浩.「グラフマイニングとその統計的モデリングへの応用 」，統計数理, vol.54, no.2 ,2007

[Hyvärinen 99] Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans.on Neural Networks 10 pp.626-634, 1999.

[Simon 53] Simon, H. Causal Ordering and Identifiability. In Models of Discovery, pages 53-80. D. Reidel, Derdrecht, Holland, 1953.

[Blalock 61] Blalock, Hubert M. Causal Inferences in Nonexperimental Research. The Univ. of North Carolina Press, Chapel Hill, North Carolina, 1961.