Vol.2019-BIO-59 No.3
2019/9/8

テンソル分解に基づく教師なし学習による変数選択はMicroRNAトランスフェクションにより仲介されるmRNAの配列非特異的オフターゲット調節の普遍的性質を同定することができる

田口善弘[1,a)]

概要：microRNA は標的の mRNA を分解したり、翻訳を阻害することで遺伝子の発現を抑えることが主な機能であると信じられてきた。しかし、microRNA の標的を同定するために人工的に導入された microRNA の効果は、期待に反して多数の mRNA の発現の上昇を招く。本研究では近年著者が提案してきた「主成分分析を用いた教師なし学習による変数選択」および「テンソル分解を用いた教師なし学習による変数選択」を用いて、人工的に導入された microRNA の効果を研究した。その結果、人工的に導入された microRNA によって発現変化する mRNA の種類には（導入する microRNA の種類や、導入する培養細胞の種類によらない）普遍性があることが解った。この効果は microRNA と結合するタンパクが導入した microRNA によって奪われてしまい、既存の microRNA による標的 mRNA の抑制効果が阻害されてしまったことによると解釈できることが解った。

1. はじめに

microRNA(miRNA) は短鎖 (22 塩基長程度) の非コード RNA であり、一般にはシード領域と呼ばれる 8 塩基長の領域と相補的な配列を非翻訳領域に持つ mRNA を破壊したり、翻訳を阻害したりすることで遺伝子の発現を抑止する機能があると信じられてきた [2]。しかし、miRNA の標的 mRNA を同定するために行われた多数の miRNA 導入実験においては、発現が「上昇」する mRNA が多数観測されることがままあった。miRNA の本来の機能からはこの現象は説明が難しい。

本研究ではこの現象に、著者が最近提案している「主成分分析を用いた教師なし学習による変数選択」および「テンソル分解を用いた教師なし学習による変数選択」 [3] を用いて挑み、発現が変化する mRNA に、導入する miRNA の種類や、導入する培養細胞の種類に依らない普遍性があることが解ったのでこれを報告する。

[1] 中央大学理工学部物理学科
[a)] tag@granular.com
本研究は原著論文として刊行済みである [1]

2. 方法

2.1 miRNA プロファイル

本研究で用いた mRNA プロファイルは**表 1** のとおりである。使用した培養細胞の種類も導入した miRNA の種類も非常に多様性に富んでいる。

2.2 主成分分析を用いた教師なし学習による変数選択

$x_{ij} \in \mathbb{R}^{N\times M}$ は i 番目の mRNA の j 番目のサンプルにおける発現プロファイルであるとする。x_{ij} は $\sum_{i=1}^{N} x_{ij} = 0$ 及び $\sum_{i=1}^{N} x_{ij}^2 = N$ になるように、標準化されているとする。x_{ij} に mRNA に主成分得点 $u_{\ell i} \in \mathbb{R}^{N\times M}$ が、サンプルに主成分負荷量 $v_{\ell j} \in \mathbb{R}^{M\times M}$ が付与されるように主成分分析を適用する（今の場合全てについて $N > M$ である）。主成分負荷量を検証し「参照群と操作群の間には差があるが、miRNA の種類にはよらない発現プロファイル」を呈している成分 ℓ' を見つけ、これに対応する第 ℓ' 主成分得点を用いて、i 番目の遺伝子に P 値 P_i を付与する。

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell' i}}{\sigma_{\ell'}}\right)^2 \right] \tag{1}$$

$P_{\chi^2}[> x]$ は、引数が x 以上である χ 二乗分布の累積確率であり、$\sigma_{\ell'}$ は標準偏差である。P 値は BH 基準 [4] で多重比

表 1　本研究で使用された 1 1 個の実験のリスト。OE は強制発現
Table 1　Eleven experiments used for this analysis. OE: overexpression.

exp	GEO ID	cell lines (cancer)	miRNA	misc
1	GSE26996	BT549 (breast cancer)	miR-200a/b/c	
2	GSE27431	HEY (ovarian cancer)	miR-7/128	mas5
3	GSE27431	HEY (ovarian cancer)	miR-7/128	plier
4	GSE8501	Hela (cervical cancer)	miR-7/9/122a/128a/132/133a/142/148b/181a	
5	GSE41539	CD1 mice	cel-miR-67,hsa-miR-590-3p,hsa-miR-199a-3p	
6	GSE93290	multiple	miR-10a-5p,150-3p/5p,148a-3p/5p,499a-5p,455-3p	
7	GSE66498	multiple	miR-205/29a/144-3p/5p,210,23b,221/222/223	
8	GSE17759	EOC 13.31 microglia cells	miR-146a/b	(KO/OE)
9	GSE37729	HeLa	miR-107/181b	(KO/OE)
10	GSE37729	HEK-293	miR-107/181b	(KO/OE)
11	GSE37729	SH-SY5Y	181b	(KO/OE)

較補正され、補正 P 値が 0.01 以下の遺伝子が選択される。

2.3　テンソル分解を用いた教師なし学習による変数選択

$x_{ijk} \in \mathbb{R}^{N\times M\times K}$ は i 番目の mRNA の j 番目及び k 番目の実験条件のサンプルにおける発現プロファイルであるとする。x_{ijk} は $\sum_{i=1}^{N} x_{ijk} = 0$ 及び $\sum_{i=1}^{N} x_{ijk}^2 = N$ になるように、標準化されているとする。x_{ijk} に HOSVD [3] を適用し、テンソル分解

$$x_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1,\ell_2,\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \qquad (2)$$

を得る。典型的には i として導入した miRNA の種類、$j=1$ としてモック、あるいは、コントロール、$j=2$ として miRNA 導入をしたサンプルが選ばれる。k は導入された miRNA の種類に対応する。典型的には $\ell_2 = 2$ がコントロールと導入サンプルで逆符号、すなわち、導入によって発現が変化したプロファイルを表し、$\ell_3 = 1$ が miRNA の種類によらない、すなわち、シード配列に依らない mRNA の発現プロファイルを表す。そこで $G(\ell_1, 2, 1)$ の中でもっとも絶対値が大きい ℓ_1 を同定し、

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}}\right)^2 \right] \qquad (3)$$

で i 番目の遺伝子に P 値を付与する。P 値を多重比較補正して、補正 P 値が 0.01 以下の遺伝子を選択するのは同じである。

2.4　エンリッチメント解析

同定された遺伝子をエンリッチメントサーバーである Enrichr [5] にアップロードしてどのような生物学的な機能がある遺伝子選ばれたのかを調べた。

3.　結果

表 1 にある mRNA プロファイルのどれに「主成分分析を用いた教師なし学習による変数選択」と用いて、どれに「テンソル分解を用いた教師なし学習による変数選択」を用いるのかは一概に言えず、場合による。大まかに言えば、対照群（今の場合はモック miRNA を導入した場合など）と操作群（今の場合は miRNA を導入した場合）の数が同じである場合や一対一対応している場合は後者を用いることが容易である（テンソルのあるモードを導入した miRNA に対応させ、対称群か操作群かの違いを長さ 2 の別のモードに対応させることができるので）。そうでない場合には前者を用いることが多いが、前者は行列の場合、つまり、mRNA 対実験条件、の形式しかある変えないため、一概には言えない。詳しくは原著論文 [1] を参照してほしいが、No.1,3,5 の mRNA プロファイルに前者が、他の mRNA プロファイルには後者が適用された。

11 個の mRNA プロファイルの全ての結果をここで説明することは紙面の都合があり出来ないが、いくつかの例を説明する。他の例については原著論文 [1] を参照いただきたい。

図 1 は表 1 の実験 1 に主成分分析を適用した時に得られる第二主成分負荷量である。見てのとおり、操作群（左の 6 つ）と参照群（右の 6 つ）の間は逆符号であるにも関わらず、それぞれの群の中では miRNA 依存性がないプロファイルが得られている。このことから、この変化は miRNA のシード配列によらない普遍的な、しかし、miRNA 導入によって引きこされた発現であることがわかる。そこでこの実験 1 に対しては、$\ell' = 2$ とし、第二主成分得点 u_{2i} を用いて式 (1) により、P 値を付与する。補正 P 値が 0.01 以下という基準で選んだところ、実験 1 については 2 3 2 個の遺伝子が選択された。

図 2 は表 1 の実験 4 に HOSVD を適用した結果である。x_{ijk} は i が mRNA、j が参照群と操作群の区別、k が（レプリケイトを含めた）導入 miRNA の種別を表している。u_{2j} が参照群と操作群で逆符号であること、u_{1k} が mRNA の発現が導入した miRNA の種類に依らないことを表現しているので、$G(\ell_1, 2, 1)$ の中で絶対値が大きな ℓ_1 を選択す

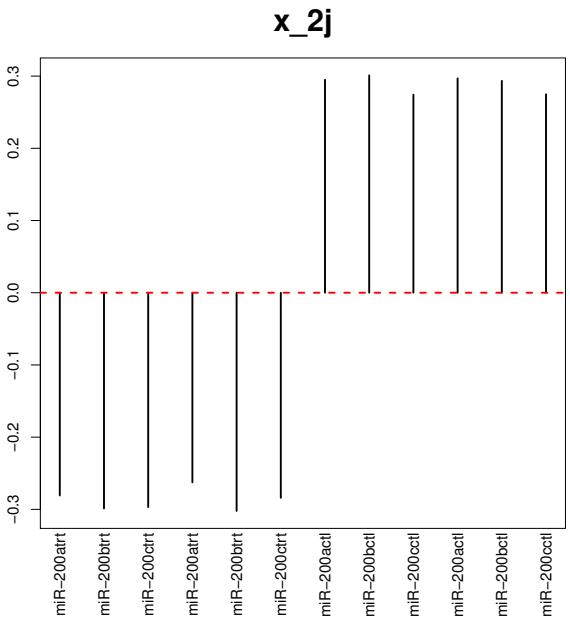

図 1 表 1 の実験 1 に PCA を適用して得られた第二主成分負荷量
Fig. 1 The second PC loading obtained by applying PCA to exp. 1 in Table 1

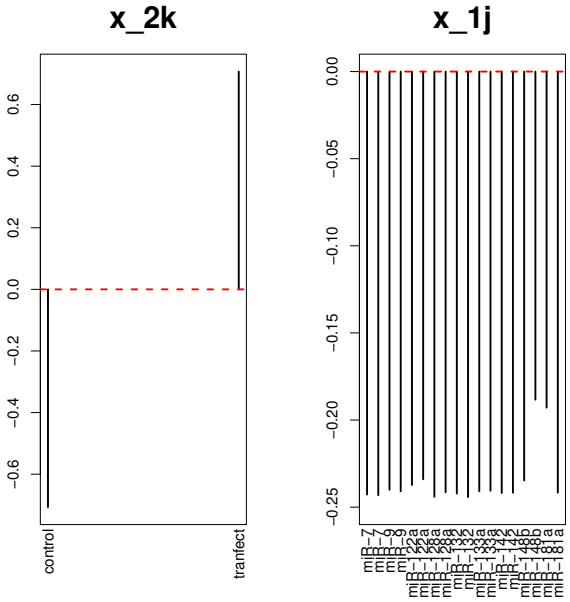

図 2 表 1 の実験 4 に HOSVD を適用して得られた u_{2j}（左）と u_{1k}（右）。
Fig. 2 u_{2j}(left) and u_{1k} right) when HOSVD was applied to exp. 4 in Table 1

ることで mRNA の選択に用いる $u_{\ell_1 i}$ を決める。その結果 $G(6,2,1)$ が絶対値最大であることが解ったので u_{6i} を用いて、式 (3) で i 番目の mRNA に P 値を付与した。補正 P 値が 0.01 以下という基準で選んだところ、実験 4 については 4 4 1 個の遺伝子が選択された。

表 2 は、1 1 個の実験に対して遺伝子選択を行った結果の比較である。まず、実験に依らず、ほぼ数百個の遺伝子が選択されていることがわかる。さらに、選択された遺伝子はお互いに大きく重なっていることが解る。異なった培養細胞に異なった miRNA を導入しているにも関わらず、共通した遺伝子が選択されていることからなんらかの普遍的な原因がそこにはあることが期待される。

こんなに簡単に共通の遺伝子が選択されるなら、こんな複雑なことをしなくてももっと簡単に 1 1 個の実験に対して共通の遺伝子がみつかるのではないか、という疑問が湧くであろう。そこで一番簡単な手法として t 検定を試してみた。t 検定では、二群の比較しかできないので、単純に操作群と対称群の二群比較を行って P 値を計算し、BH 基準で多重比較補正して補正 P 値が 0.01 以下の遺伝子を選択した。**表 3** がその結果である。実験ごとにサンプル数が大きく異なっていることもあり、1 1 個の実験のうち、5 個の実験で一つも遺伝子が選べていない一方、何千個も選ばれてしまっている実験が 3 つもあるという状態で、共通の遺伝子が有意に選択されているかどうかの検証など全くできない有様である。表 2 だけをみるといとも簡単そうに見えるが、実は、サンプルサイズによらず、同じ基準でほぼ同じ数の遺伝子を選択するという事自体、非常に難しいのである。これができるだけでも、「主成分分析を用いた教師なし学習による変数選択」や「テンソル分解を用いた教師なし学習による変数選択」がいかに優れた手法であるかは明らかだろう。

ここでやめてしまってもいいのだが、さらにダメ押しをするために、有意性にはこだわらず、単純に P 値が小さい順に、各実験で遺伝子を表 2 にある数だけ強引に選択し、重なりがどれだけあるかを見てみた。これで表 2 と同等程度に共通の遺伝子が選ばれれば、「主成分分析を用いた教師なし学習による変数選択」や「テンソル分解を用いた教師なし学習による変数選択」の優位性は単純により小さな P 値をだせるというだけで（勿論、それ自体重要なことだが）、遺伝子の重要度のランキングには優位性がないことになる。**表 4** がその結果である。オッズ比がゼロ、つまり全く重なりがない場合も多数あり、そもそも、オッズ比が 1 を超えている（つまり、偶然以上に重なりが多い）ペアが殆どないことが解る。つまり、1 1 個の独立した実験でお互いに重なり合った遺伝子を選択するというタスクはとんでもなく困難なのである。

ここで、t 検定よりもっと複雑な方法を用いれば良くなるのでは、という可能性も捨てきれないが、基本的に SAM [6] や limma [7] といった方法は「改良された t 検定」にすぎず、特に、遺伝子に付与される P 値の順位（大小関係）が劇的に変化することは考えにくい（t 検定で P 値が小さいものは SAM や limma でも P 値が小さいという強い傾向がある）ことを考えると、表 4 に比べて劇的に良くなると考えるのは難しい。そこでこれ以上、深追いはしないことにする。

表 2 １１実験に対するフィッシャーの正確確率検定の結果。上半分：P 値。下半分：オッズ比。

Table 2 Fisher's exact tests for coincidence among 11 miRNA transfection experiments. Upper triangle: P-value, lower triangle: odds ratio.

exp.		1	2	3	4	5	6	7	8	9	10	11
	#	232	711	747	441	123	292	246	873	113	104	120
1	232		4.14e-19	6.59e-22	3.96e-41	4.12e-71	9.41e-70	2.90e-60	1.34e-17	1.15e-27	6.84e-26	2.66e-07
2	711	7.68		0.00	1.89e-18	4.93e-27	5.59e-20	2.69e-32	4.62e-13	9.23e-16	8.66e-12	1.37e-03
3	747	8.30	345.52		3.63e-20	7.96e-21	5.70e-12	1.82e-27	9.52e-12	1.18e-14	1.01e-12	3.90e-06
4	441	18.23	5.19	5.34		6.14e-41	1.01e-34	1.44e-69	4.61e-11	2.16e-30	4.09e-28	1.35e-10
5	123	53.86	9.04	7.27	17.48		2.9e-179	1.27e-63	6.24e-15	3.16e-25	2.37e-17	4.69e-09
6	292	61.50	8.15	5.52	17.71	204.39		3.53e-53	2.57e-15	6.65e-22	1.65e-12	5.60e-05
7	246	20.27	5.35	4.67	12.39	20.11	22.03		6.91e-42	1.77e-36	4.50e-31	2.78e-14
8	873	18.61	7.22	6.51	8.29	15.61	18.53	20.73		1.81e-07	1.37e-06	2.76e-02
9	113	39.34	9.87	8.77	25.98	32.44	34.90	21.94	16.02		3.7e-125	9.27e-18
10	104	40.29	8.22	8.27	26.64	23.34	20.86	21.56	15.18	517.87		6.82e-16
11	120	10.15	3.19	4.43	9.19	11.55	8.11	8.28	4.92	19.57	18.70	

#: the number of genes selected for each of 11 experiments via TD- or PCA-based unsupervised FE.

表 3 t 検定による遺伝子選択。実験番号は表 2 と同じ。コロンの前後の数字はそれぞれ対称群と操作群である

Table 3 The number of genes selected by t test. The numbering of experiments are the same as those in Table 2. Two numbers besides colon are the number of control and treated samples, respectively.

Experiments	1	2	3	4	5	6	7	8	9	10	11
Samples	6:6	3:4	6:4	18:18	2:2	16:16	19:19	18:18	6:12	6:12	4:4
Selected genes	11060	0	0	0	0	35	280	55	5949	5730	0

次にある疑問としては表 2 で共通に選ばれた遺伝子は、方法論的なバイアスで選ばれたのであり、生物学的には何か共通のことは起きてはいないのでは、という可能性が考えられる。この疑念を払拭するため、選ばれた遺伝子をエンリッチメント解析し、どの様な生物学的な意味があるかを検証した。**表 5** は Encihr の "ENCODE and ChEA Consensus TFs from ChIP-X" カテゴリによる選択された遺伝子を標的とする転写因子の実験ごとの上位２０位までの結果である。１１個の実験に対して最低でも１１個、最大では９７個もの転写因子が、選択された遺伝子を標的とする転写因子として選択されている。まず、この事実が、各実験ごとに選ばれた遺伝子が生物学的に無意味なものであることを強く否定している。１１個の実験に対して選択された遺伝子は全て生物学的に意味があるものだといって構わないだろう。また、EKLF,MYC,NELFA,E2FL の４つの転写因子は１１個の実験全てで共通に選択されている。少ない場合にはわずか１１個の転写因子しか選ばれない場合もあることを考えると、４個もの転写因子が、１１個の実験全て選ばれることは偶然ではほぼ不可能である。このことは、単に個々の実験で選ばれた遺伝子が生物学的に意味があるものであるということ以外に、表 2 に見るような共通に選ばれた遺伝子にも生物学的な意味があるということを強く示唆する。

一方で我々は同じく Enrichr の "TargetScan microRNA" を調べた。miRNA の種類に依らず、共通の遺伝子が選ばれている以上、これらが miRNA の標的として発現変化していることは考えにくいが、それでも、個々の実験で、導入した miRNA の標的遺伝子が含まれていないとは断言できない。その結果、１１個の実験のうち、実験２と３以外は特定の miRNA の標的になっている遺伝子が有意に多く含まれているということは一切ないことが解った。この事実は、miRNA 導入実験で普遍的に見られる発現が変化する遺伝子が、導入された miRNA の直接の標的ではないことの傍証になる。

実際にはこの他にも多数のエンリッチメント解析を行ったが [1]、紙面の関係上、割愛する。

4. 議論

miRNA の機能はシード領域の配列と相補的な配列をもつ mRNA の発現抑制にあるはずである。個々の実験で、導入した miRNA の種類によらず発現が変化する mRNA のセットがみつかったばかりではなく、それらが互いに異なった実験ごとに共通しているとなると、何らかの非常に普遍的は生物学的な枠組みがないかぎりあり得ないと思われる。その理由として、我々は導入した miRNA によるタンパク装置の奪取を考えた。miRNA はシード領域に相補的な mRNA に結合した後、翻訳抑制や、破壊を行う必要があるが、そのためには、様々なタンパクの補助を必要とする。また、DNA から転写された miRNA はそのままでは機能せず、様々なプロセッシングを経なくては標

表 4　１１実験に対する t 検定選択遺伝子の一致度。他の表記は表 2 と同じ。
Table 4　The coincidence of genes selected by t test between 11 experiments. Notations are the same as those in Table 2.

	1	2	3	4	5	6	7	8	9	10	11
1		4.96e-04	8.49e-01	2.59e-01	6.35e-01	1.00e+00	5.40e-01	1.00e+00	4.08e-01	6.45e-01	6.68e-01
2	2.56		6.40e-69	1.38e-02	1.25e-01	1.55e-01	9.36e-03	1.00e+00	1.00e+00	3.76e-01	1.00e+00
3	0.80	10.49		8.65e-01	5.28e-01	3.76e-01	2.47e-01	7.79e-01	7.75e-01	5.30e-01	1.00e+00
4	1.55	1.90	0.89		6.58e-01	1.00e+00	4.31e-01	1.26e-01	2.71e-01	2.56e-01	1.00e+00
5	0.00	0.00	0.36	1.39		1.13e-22	1.00e+00	3.86e-01	1.00e+00	1.00e+00	1.00e+00
6	0.77	1.83	0.32	0.72	27.05		3.71e-01	1.00e+00	1.00e+00	1.00e+00	1.00e+00
7	1.16	0.48	0.71	1.22	0.67	0.31		4.47e-01	1.83e-01	7.60e-02	2.04e-01
8	0.64	1.00	1.17	2.15	2.09	0.00	0.46		1.59e-01	4.54e-01	1.27e-03
9	0.00	0.81	0.60	0.00	0.00	0.00	0.25	2.91		1.18e-03	4.07e-01
10	0.00	0.32	0.35	1.75	0.00	0.00	0.00	1.68	5.56		6.37e-01
11	1.31	0.78	0.88	0.97	0.00	0.00	1.69	6.87	0.00	0.00	

表 5　１１実験に対する転写因子同定（上位２０位まで）。標的が選択遺伝子と被っている転写因子を選んだ。EKLF, MYC, NELFA, と E2F1 が全１１実験で共通に選ばれた。
Table 5　In each of 11 experiments, 20 top-ranked significant TFs whose sets of target genes significantly overlap with the set of genes selected for each experiment were identified. Then, EKLF, MYC, NELFA, and E2F1 turned out to be among the 20 top-ranked significant TFs for all 11 experiments.

			EKLF		MYC		NELFA		E2F1	
exp	#1	#2	OL	adj. P-value	OL	adj. P-value	OL	adj. P-value	OL	adj. P-value
1	232	30	40/1239	2.16e-07	53/1458	6.22e-12	59/2000	8.94e-10	61/1529	1.30e-15
2	711	77	94/1239	1.51e-10	106/1458	1.01e-10	134/2000	3.26e-11	100/1529	6.14e-08
3	747	97	100/1239	2.28e-11	98/1458	2.96e-07	152/2000	1.93e-15	108/1529	3.89e-09
4	441	43	83/1239	4.77e-18	99/1458	2.08e-22	105/2000	1.06e-15	85/1529	9.29e-14
5	123	45	26/1239	2.16e-06	25/1458	9.12e-05	31/2000	4.26e-05	28/1529	7.41e-06
6	292	19	51/1239	2.38e-09	65/1458	5.54e-14	63/2000	2.72e-07	69/1529	8.31e-15
7	246	11	37/1239	5.11e-05	48/1458	5.97e-08	46/2000	1.45e-03	64/1529	7.02e-16
8	873	55	188/1239	8.33e-52	189/1458	8.58e-42	222/2000	3.52e-39	157/1529	8.24e-23
9	113	36	24/1239	4.47e-06	30/1458	2.86e-08	32/2000	2.33e-06	40/1529	6.84e-15
10	104	22	27/1239	1.16e-08	25/1458	4.83e-06	36/2000	1.07e-09	35/1529	3.63e-12
11	120	22	21/1239	8.02e-04	27/1458	2.25e-05	29/2000	4.68e-04	25/1529	3.57e-04

#1: the number of genes selected for each of 11 experiments via TD- or PCA-based unsupervised FE, #2: the number of TFs whose sets of target genes significantly (adjusted P-values < 0.01) overlap with the set of genes selected for each experiment. OL: overlaps, (the number of genes coinciding with the genes selected for each experiment)/(genes listed in Enrichr as TF target genes).

的 mRNA と結合してこれらを破壊することができないが、そのプロセッシングにもタンパク装置は必要である。要するに、miRNA がその機能を発揮するには補助的なタンパク装置の助力が必要である。だが、ただの RNA に過ぎない miRNA に対して、ちゃんとしたタンパクを多数作ることは簡単ではない。外部から大量の miRNA が人工的に導入されれば既存のタンパク装置は導入した miRNA に乗っ取られ、既存の miRNA は標的 mRNA の発現抑止を行えなくなり、その結果、既存 miRNA の標的 mRNA の発現量はむしろ上がることになる。このプロセスは、導入した miRNA の種類によらず、既存の miRNA の標的 mRNA すべての発現に影響するので、今回見出されたような、導入 miRNA の種類によらない発現変化を起こす mRNA の出現を引き起こしうる。

実際、従来から、導入 miRNA によって多数の mRNA の発現が「上昇」することは知られていたが、miRNA 導入実験の目的はあくまで、シード領域と相補的な配列をもつ mRNA の抑止であり、発現が上昇する mRNA は興味の対象外であったため、積極的に解析がされてこなかった。

この仮説を実験なしに確かめるのは難しいが、我々はいくつかの傍証となる解析を行った。**図 3** は各遺伝子を標的とする miRNA の数と発現変化の関係である。前述の仮説が正しければ、標的としている miRNA の数が多いほど、タンパク機構の奪取によって、標的としている miRNA が失われることで、発現が上昇する割合は大きいはずである。図 3 はこの関係を実数とランクで解析したものである。相関係数は非常に小さいが有意性を示す P 値は非常に小さい。この結果は仮説とは矛盾しない。

もう一つの傍証は DICER と呼ばれる、miRNA をプロセッシングするタンパクとの関係である（**表 6**）。大量の miRNA が導入されれば DICER はその処理に追われ、本来の機能を果たすことができなくなる。結果的に、これは DICER タンパクをノックアウト（KO）したのと同じ効果をもたらすはずだ。Enrichr には DICER を KO した時の網羅的な遺伝子の発現変化の実験が１６実験収録されているが、この実験のうち、miRNA 導入で普遍的に変化する

表 6 GEO DICER KO: 1 6個の DICER タンパク KO 実験のうちのいくつと選択遺伝子が被っているか。IP:DICER タンパクとの結合が有意であるかをフィッシャーの正確確率検定で調べた。

Table 6 GEO DICER KO: the number of experiments among the 16 experiments included in Enrichr whose set of listed genes significantly overlapped with the set of genes identified in each of the 11 experiments. IP: Fisher's exact test for the overlap between the set of genes that bind to Dicer in immunoprecipitation (IP) experiments and the set of genes selected in each of the 11 experiments.

Experiments		1	2	3	4	5	6
GEO DICER KO	up	12/16	12/16	12/16	12/16	14/16	11/16
	down	13/16	12/16	12/16	13/16	14/16	10/16
IP	*P*-value	2.49e-23	7.22e-22	1.31e-17	5.55e-29	5.21e-35	1.78e-20
	odds	47.4	20.6	15.9	38.7	64.2	41.2
Experiments			7	8	9	10	11
GEO DICER KO	up		12/16	14/16	12/16	13/16	12/16
	down		12/16	12/16	14/16	14/16	10/16
IP	*P*-value		4.72e-32	4.29e-16	2.19e-11	3.96e-10	4.64e-08
	odds		37.0	41.4	42.6	39.6	27.3

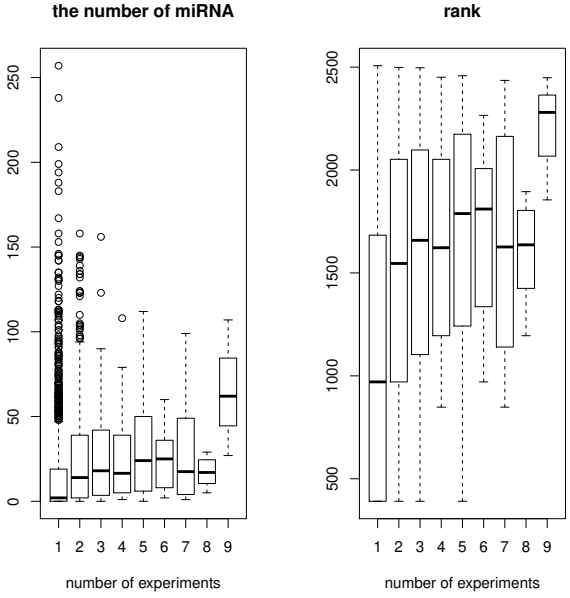

図 3 1 1実験で選ばれた遺伝子を標的とする miRNA の数と、1 1個中何個の実験で選ばれたかの箱ひげ図。左：実数、ピアソン相関係数は 0.13 で、P 値は 3.9×10^{-11}。右：順位。スペアマンの相関係数は 0.29 で、P 値は 2.2×10^{-16} 以下。

Fig. 3 A boxplot of the number of miRNAs that target individual genes as a function of the number of experiments that select individual genes within 11 experiments (most frequently selected genes were selected in nine experiments). Left: raw numbers (Pearson's correlation coefficient = 0.13, $P = 3.9 \times 10^{-11}$), right: ranks of numbers (Spearman's correlation coefficient = 0.29, $P < 2.2 \times 10^{-16}$)

遺伝子群と被っているものがどれくらいあるか調べた。その結果、大部分の実験が有意な相関を1 1個の miRNA 導入実験と持っていることが分かった。この事実は、miRNA 導入実験によって生じる、普遍的な発現変化が導入 miRNA による、タンパク装置の奪取に依って起きる、という仮説と矛盾しない。また、DICER と結合する mRNA というものが存在するが、これらの遺伝子も、miRNA 導入で普遍的に変化する遺伝子群と激しく被っており、これらも仮説とは矛盾しない結果となっている。

5. おわりに

本研究では、純粋にバイオインフォマティクスを用いた既存のデータの再解析だけで、何か新しい生物学的な機構の提案をできないかという観点から、miRNA 導入実験はシード配列の相補性とは無関係な、普遍的な発現変化を引き起こすことを示した。現在、原著論文 [1] の出版から1 4ヶ月が経過しているが、まだコンピュータ・サイエンスの論文 [8] に一度引用されただけで、生物学の分野からは完全に黙殺されている。こんな解析をしただけで実験家が実験をしてくれるわけもないのだとは思うが、なんらかの形で純粋にバイオインフォマティクスで提案された機構を実験的に検証できるような枠組みが将来はできることを祈っている。

参考文献

[1] Taguchi, Y.-H.: Tensor Decomposition-Based Unsupervised Feature Extraction Can Identify the Universal Nature of Sequence-Nonspecific Off-Target Regulation of mRNA Mediated by MicroRNA Transfection, *Cells*, Vol. 7, No. 6 (online), DOI: 10.3390/cells7060054 (2018).
[2] 落合孝広，山本雄介（編）：医学のあゆみ マイクロ RNA 研究の進歩 2019 年 269 巻 5 号 5 月第 1 土曜特集 [雑誌]，医歯薬出版 (2019).
[3] Taguchi, Y.-h.: *Unsupervised Feature Extraction Applied to Bioinformatics*, Springer International, Switzerland (2019).
[4] Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300 (online), available from ⟨http://www.jstor.org/stable/2346101⟩ (1995).
[5] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W. and Ma'ayan, A.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, *Nucleic Acids Research*, Vol. 44, No. W1, pp. W90–W97 (online), DOI: 10.1093/nar/gkw377 (2016).
[6] Tusher, V. G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, Vol. 98, No. 9, pp. 5116–5121 (online), DOI: 10.1073/pnas.091062498 (2001).
[7] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K.: limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research*, Vol. 43, No. 7, pp. e47–e47 (online), available from ⟨https://doi.org/10.1093/nar/gkv007⟩ (2015).
[8] Sani, L., Pecori, R., Mordonini, M. and Cagnoni, S.: From Complex System Analysis to Pattern Recognition: Experimental Assessment of an Unsupervised Feature Extraction Method Based on the Relevance Index Metrics, *Computation*, Vol. 7, No. 3 (online), DOI: 10.3390/computation7030039 (2019).