Vol.2020-MPS-128 No.1
Vol.2020-BIO-62 No.1
2020/6/29

テンソル分解に基づく教師なし学習による変数選択のRNAi処理を行ったプラナリアのRNA-seq解析への適用

田口 善弘[2,a)]　鹿島 誠[1,b)]

概要：プラナリアは驚異的な自己再生能力を有していることで有名な生命体であるが、どのような分子生物学的な基盤がこのような特異な自己再生能力を担保しているかはよく解っていない。近年の研究によりキー遺伝子が解明され、特定の遺伝子をノックダウンすることで、リーサルではない範囲で自己再生能力特異的な表現型変異が起こせることが解ってきた。本研究ではこれらのキー遺伝子をRNAiで抑止して自己再生の力を奪った場合の時系列的なRNA-seqを敢行して、キー遺伝子のノックアウトでどのような遺伝子が影響を受けるかの観測を行った。RNA-seqのデータセットに「テンソル分解を用いた教師なし学習による変数選択」法を用いることで、RNAi特異的でかつ時系列依存的に発現が変化する転写物群の特定に成功したのでこれについて報告する。

1. はじめに

プラナリアは、扁形動物門ウズムシ綱ウズムシ目ウズムシ亜目に属する、いわゆる扁形動物の一種である。体長は数ミリから数センチ程度であり、目、口、集中神経系は保持するものの、脊髄や肛門はなく、眼もレンズをもっているわけではないので、ボディプラン的には、イソギンチャクなど肛門をもたない（口から食物摂取と排泄を行う）多細胞生物から、昆虫や脊椎動物の様に口と排泄口を別々に持っている（口から摂取して排泄口から排泄する）多細胞生物の中間に位置する生物である。

プラナリアは体をかなり多数の断片に分割されても、各断片が独立に全身を再生できるという、イモリやトカゲの手足や尻尾の再生とは異なった（ヤモリの手足やトカゲの尻尾は失われても再生するが、切断された手足や尻尾から全身が再生するわけではない）強力な自己再生能力をもっていることで有名である。

プラナリアの断片からの全身再生能力は実験的にはかなり古くから知られていたが、いかなる分子科学的な基盤がその再生能力を支えているのかは長い間謎であった。Shibata et al [1] は近年、*piwiA*,*piwiB*,*piwiC* などのいわゆる *piwi* 遺伝子がこのプラナリアの自己再生能力のキー遺伝子であり、これらをノックダウンすることでプラナリアの自己再生能力をリーサルではない範囲で抑止できることを見出した。また、これらのキー遺伝子と相互作用する遺伝子が非常に多数あることも特定したが、具体的に、これらのキー遺伝子のノックダウンがどのような機構でプラナリアの自己再生能力の抑止を行っているのかよく分からなかった。

本研究ではキー遺伝子を含む、相互作用遺伝子をRNAiでノックダウンしたあとの時系列的なRNA-seq計測を低コストハイスループットなRNA-Seq手法であるLasy-Seq [2]によって行った。「テンソル分解を用いた教師なし学習による変数選択」法 [3] を用いることで、表現型（自己再生能力）の喪失と相前後する形でノックダウン特異的な発現変化を行う遺伝子群の特定に成功したのでその結果について報告する。

2. 材料と方法

2.1 プラナリアのノックダウン実験

プラナリアは2日間の間隔を空けて給餌を介したRNAiを二度、施した。RNAiの対象遺伝子は *piwiA*,*piwiB*,*piwiC* のキー遺伝子の他、PiwiBと相互作用する可能性があるタンパク質をコードする遺伝子である *hp1-A*, と *ima1-A*、及び、コントロールとしてGFP, また、自己再生機能と関係するが、PIWI経路と関係がない遺伝子として *hdac* と *cdc1* を選んだ。2度めのRNAi処理後、1個体ずつサンプリングを行い1 6日間に渡ってRNA-seqを行った。各実験条件でのバイオロジカルレプリケイト数は6である。

[1] 青山学院大学
Aoyama Gakuin University
[2] 中央大学
Chuo University
[a)] tag@granular.com
[b)] kashima.biology@gmail.com

Vol.2020-MPS-128 No.1
Vol.2020-BIO-62 No.1
2020/6/29

2.2 RNA-seq 解析

RNA-seq で得られた fastq ファイルは、先行研究で Trinity [4] を使って得られたコンティグに対して BWA [5] でマッピングすることで各コンティグへのマッピングを行い、マッピング数を改めて salmon [6] を使用して各コンティグの発現量とすることで発現量解析を行った。

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

得られた発現プロファイルは以下の形式でテンソルの形にフォーマットされた。$x_{ij_1j_2j_3} \in \mathbb{R}^{N\times 8\times 15\times 6}$ は i 番目のコンティグ（$N = 278167$）の j_1 番目の RNAi(j_1 =1:*cdc1*, 2:*gfp*, 3:*hdac*, 4:*hp1-A*, 5:*ima1-A*, 6:*piwiA*, 7:*piwiB*, 8:*piwiC*), j_2 番目の日付 (異常値があった 1 4 日めを除いたため、j_2=14,15 はそれぞれ 15,16 日目に相当)、j_3 番目のバイオロジカルレプリケイト ($1 \le j_3 \le 6$) での発現量である。$x_{ij_1j_2j_3}$ は $\sum_i x_{ij_1j_2j_3} = 0$ 及び $\sum_i x^2_{ij_1j_2j_3} = N$ というように規格化されているとする。$x_{ij_1j_2j_3}$ に Higher Order Singular Value Decoposision (HOSBD) [3] を適用してテンソル分解

$$x_{ij_1j_2j_3} = \sum_{\ell_1=1}^{8}\sum_{\ell_2=1}^{15}\sum_{\ell_3=1}^{6}\sum_{\ell_4=1}^{N} G(\ell_1\ell_2\ell_3\ell_4)u_{\ell_1 j_1}u_{\ell_2 j_2}u_{\ell_3 j_3}u_{\ell_4 i} \quad (1)$$

を得る。ここで $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{N\times 8\times 15\times 6}$ はコアテンソル、$u_{\ell_1 j_1} \in \mathbb{R}^{8\times 8}, u_{\ell_2 j_2} \in \mathbb{R}^{15\times 15}, u_{\ell_3 j_3} \in \mathbb{R}^{6\times 6}, u_{\ell_4 i} \in \mathbb{R}^{N\times N}$ は特異値行列で、全て直交行列である。

テンソル分解を用いた教師なし学習による変数選択法 [3] では、まず、望ましい時間依存性がある特異値ベクトル $\boldsymbol{u}_{\ell_2}$ を選び、次に、望ましい RNAi 依存性（自己再生能力と関係していると思われる、*hp1-A* ($j_1 = 4$), *ima1-A* ($j_1 = 5$), *piwiA* ($j_1 = 6$), *piwiB* ($j_1 = 7$), 及び, *piwiC* ($j_1 = 8$) とそれ以外で差がある）特異値ベクトル $\boldsymbol{u}_{\ell_1}$ を選ぶ。最後に、バイオロジカルレプリケイト依存がない特異値ベクトル $\boldsymbol{u}_{\ell_3}$ を選ぶ。

次に、コンティグ選択に使う特異値ベクトル $\boldsymbol{u}_{\ell_4}$ を選ぶために、選択した ℓ_1, ℓ_2, ℓ_3 に対して $G(\ell_1, \ell_2, \ell_3, \ell_4)$ の絶対値が大きくなるような ℓ_4 を選択する。

最後に、選択された特異値ベクトル $u_{\ell_4 i}$ がガウス分布することを仮定して、χ 二乗分布を用いて i 番目のコンティグに P 値を

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}}\right)^2\right] \quad (2)$$

という式で付与する。但し、$P_{\chi^2}[> x]$ は引数が x 以上になる場合の χ 二乗分布の累積確率であり、σ_{ℓ_4} は標準偏差である。得られた P_i は多重比較補正（Benjamini and Hochberg [7]）したあと、補正 P 値が 0.01 以下のコンティグを選択する。

3. 結果

3.1 時間依存性:u_{ℓ_2}

まず、最初に興味ある時間依存性をもつ特異値ベクトル、$\boldsymbol{u}_{\ell_2}$ を探索した。既報 [1] によれば、プラナリアは処置後７日程度で自己再生能を喪失するが、自己再生に本質的な役割を果たすと思われる新生細胞（幹細胞）自体は処置後２週間程度まで維持されていた。従って、自己再生能の喪失は２段階で起きることが期待された。**図 1** は u_{2j_2} の j_2

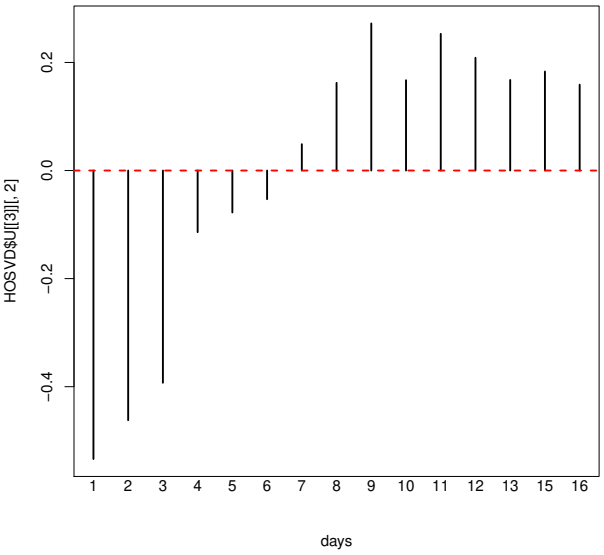

図 1 u_{2j_2} の j_2（時間）依存性
Fig. 1 Time (j_2) dependece of u_{2j_2}

（時間）依存性である。期待通り、最初の７日で急速に変化が起き、その後１６日まで一定値を維持している。従って、この時間依存性は実験で見出された２段階の表現型の現出（自己再生能の喪失）によく対応しているものと期待された。

3.2 RNAi 依存性:u_{ℓ_1}

次に RNAi 依存性について調べた。**図 2** は u_{2j_1} の j_1（RNAi）依存性である。実は *piwiA* についてはノックダウンしても明確な自己再生能喪失が観測されていないため、*piwiB*,*piwiC*,*ima1-A*,*hp1-A* 対 *piwiA*,*hdac*,*cdc1*,*gfp* という結果は妥当である。そこでこの u_{2j_1} を望ましい *RNAi* 依存性を満たす特異値ベクトルとして採用した。

3.3 バイオロジカルレプリケイト 依存性:u_{ℓ_3}

最後に、バイオロジカルレプリケイト依存性を考えた。バイオロジカルレプリケイトについては依存性があっては困るので依存性のない特異値ベクトル $\boldsymbol{u}_{\ell_3}$ を見つける必要

Vol.2020-MPS-128 No.1
Vol.2020-BIO-62 No.1
2020/6/29

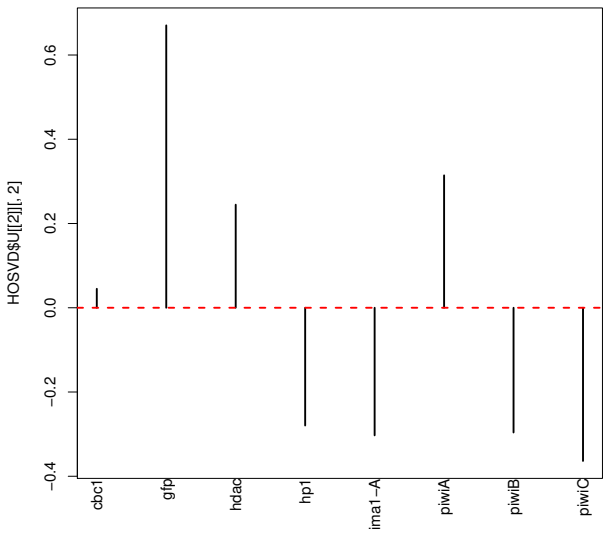

図 2 u_{2j_1} の j_1（RNAi）依存性
Fig. 2 RNAi (j_1) dependece of u_{2j_1}

がある。**図 3** は u_{1j_3} の j_3（バイオロジカルレプリケイト）

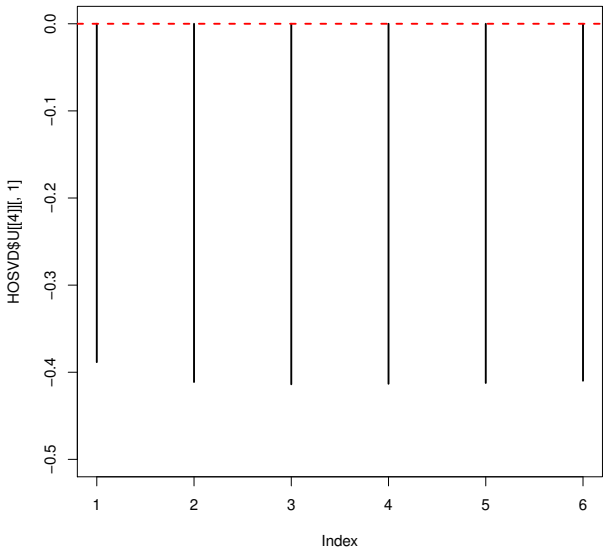

図 3 u_{1j_3} の j_3（バイオロジカルレプリケイト）依存性
Fig. 3 Biological replicates (j_3) dependece of u_{1j_3}

依存性である。明らかに u_{1j_3} はバイオロジカルレプリケイト依存性を喪失している。そこで、ここでは、u_{1j_3} を望ましいバイオロジカルレプリケイト依存性をもつ特異値ベクトルとして選択することにした。

3.4 コンティグ特異値ベクトル:u_{ℓ_4}

コンティグの選択に用いるコンティグ特異値ベクトル,$\boldsymbol{u}_{\ell_4}$ を選択する。そのためには絶対値が大きい $G(2,2,1,\ell_4)$ を選択する必要がある（**表 1**）。明らかに $G(2,2,1,3)$ が大き

表 1 $G(2,2,1,\ell_4)$
Table 1 $G(2,2,1,\ell_4)$

ℓ_4	$G(2,2,1,\ell_4)$
1	-54.425757
2	-13.935676
3	-108.537776
4	76.835795
5	9.117833
6	63.888098
7	-13.673652
8	-8.094344
9	-22.639488
10	10.015876

な絶対値を持っているので、u_{3i} を（2）式に代入して P_i を計算した。多重比較補正した P_1 が 0.01 以下のコンティグを選んだところ、１２９個のコンティグを選ぶことが出来た。

4. 議論

選択されたコンティグが実際に、特異値ベクトルで観測されたような時間依存性や RNAi 依存性をもっているかどうかを確認する。まず、時間依存性について確認した。選ばれた１２９個のコンティグに対して RNAi ごとに $\{x_{ij_1j_2j_3}|j_2 \leq 8\}$ と $\{x_{ij_1j_2j_3}|j_2 > 8\}$ の間で t 検定を行って P 値を計算し BH 基準で多重比較補正した後、補正 P 値が 0.05 以下のコンティグがあるかどうかをしらべた（**表 2**）。*piwiB*,*piwiC*,*ima1-A*,*hp1-A* で時間依存性が見られるコン

表 2 選択された１２９個のコンティグの発現量の時間依存性

Table 2 Time dependece of selected 129 contigs

	補正 P 値	
RNAi	> 0.05	< 0.05
hp1-A	75	54
ima1-A	123	6
piwiB	96	33
piwiC	126	3
hdac	77	52
piwiA	66	63

ティグが見つかっているのは妥当だが、*hdac*,*piwiA* でも時間依存性があるコンティグがみつかっている。このことは、表現系に影響しないまでもこれらのコンティグが自己再生に関係するコンティグに影響を与えている可能性を示唆する。*piwiA* は同じ *piwi* であり、多くの RNA と相互作用することは知られていながらノックダウンが自己再生に直接影響を与えてはいなかったのだが、今回、具体的に発現量に影響があるコンティグが *piwiB* や *piwiC* と共通で

あることがわかったことにより、この関係の理解がよりすすむものと思われる。また、*piwiB* に比べると *piwiC* をノックダウンした時の方が時間依存性があるコンティグの割合が少なくなっている。これについては *piwiC* をノックダウンした場合には自己再生能力が一週間後ではなく二週間後であるという報告と整合的である。また、*ima1-A* と *piwiC* の結果がよく似ているが、これについては今後の考察を待ちたい。

次に、各時刻で、*hp1-A*,*ima1-A*,*piwiB*,*piwiC* と *hdac*,*piwiA* の間で有意差があるかを確認した（**表 3**）。解析に用いたすべての時刻でこの 2 群の RNAi の間に有意差があるものが見つかったことが確認された。

表 3 選択された 1 2 9 個のコンティグの発現量の RNAi 依存性

Table 3 RNAi dependece of selected 129 contigs

	補正 P 値	
days	> 0.05	< 0.05
1	95	34
2	105	24
3	97	32
4	52	77
5	58	71
6	92	37
7	56	73
8	69	60
9	96	33
10	44	85
11	67	62
12	89	40
13	78	51
15	68	61
16	68	61

以上の結果からテンソル分解を用いた教師なし学習による変数選択法はバイオロジカルレプリケイトの間では差がないが、時間とともに単調に、かつ、RNAi 依存的に発現が変化しているコンティグを選択するという複雑なプロセスをうまくこなしていることが解る。

現在、選択されたコンティグの生物学的な意味を検討中である。

5. おわりに

一般に複数の実験条件（今回の場合は、RNAi 依存性と時間依存性）が組み合わさって実験を解析する場合には複雑なモデルを設定しないとうまく観測を表現できないことが多かった。しかし、テンソル分解を用いることでモデル化を必要とせず、複雑な要求（今回の場合は、バイオロジカルレプリケイト間で差がないが、時間依存性と RNAi 依存性のある発現量をもっているコンティグの選択）を自然に満たすことができることが解った。今後のこの方法でより詳細な解析を進めてプラナリアの自己再生能力の分子遺伝学的な背景を理解することを目指す予定である。

謝辞 本研究は科研費番号 20K12067、20H04848、19H05270 の科研費の支援の元に行われた。

参考文献

[1] Shibata, N., Kashima, M., Ishiko, T., Nishimura, O., Rouhana, L., Misaki, K., Yonemura, S., Saito, K., Siomi, H., Siomi, M. and Agata, K.: Inheritance of a Nuclear PIWI from Pluripotent Stem Cells by Somatic Descendants Ensures Differentiation by Silencing Transposons in Planarian, *Developmental Cell*, Vol. 37, No. 3, pp. 226 – 237 (online), DOI: https://doi.org/10.1016/j.devcel.2016.04.009 (2016).

[2] Kamitani, M., Kashima, M., Tezuka, A. and Nagano, A. J.: Lasy-Seq: a high-throughput library preparation method for RNA-Seq and its application in the analysis of plant responses to fluctuating temperatures, (online), DOI: 10.1101/463596 (2018).

[3] Taguchi, Y.-h.: *Unsupervised Feature Extraction Applied to Bioinformatics, A PCA Based and TD Based Approach*, Springer International (2020).

[4] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N. and Regev, A.: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nature Protocols*, Vol. 8, No. 8, pp. 1494–1512 (online), DOI: 10.1038/nprot.2013.084 (2013).

[5] Li, H. and Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, Vol. 25, No. 14, pp. 1754–1760 (online), DOI: 10.1093/bioinformatics/btp324 (2009).

[6] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression, *Nature Methods*, Vol. 14, No. 4, pp. 417–419 (online), DOI: 10.1038/nmeth.4197 (2017).

[7] 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 (1995).