遺伝子の働き具合の違いを調べる

門田 幸二

はじめに

人体は様々な組織や器官から構成されています。ヒトゲノム中の特定の領域から転写された転写物は数万種類以上存在し、ゲノム中に存在する転写物全体をトランスクリプトームといいます。転写物の実体はリボ核酸（RNA）であり、同じ転写物でも転写される量（発現量）は体内の組織（器官）ごとに異なります。同じ組織でも人によって若干異なりますし、癌と正常のように状態によっても異なります。本稿では、これまで取り組んできた「働き具合の異なる遺伝子（発現変動遺伝子）」の検出に関するトピックを紹介します。

組織特異的遺伝子の検出

特定の組織のみで高発現または低発現となる遺伝子の検出は、様々な組織で取得された発現データを入力として行います（図 1）。iPS 細胞樹立[1]の鍵となった胚性幹細胞のみで特異的に発現する候補遺伝子のスクリーニング作業とも通じるものであり、欲しいのは赤枠内で示すような結果です。私たちはまず、上田氏によって提案された赤池情報量規準に基づく複数外れ値の簡易検出法[2]が赤枠の問題にそのまま適用できることを見出しました[3]。しかしながら、

	皮膚	大腸	肝臓	膵臓	心臓	肺	脳	血液	筋肉	眼球	…	欲しい結果
遺伝子*1*	2	2	1	3	3	100	0	1	2	2		肺 ↑
遺伝子*2*	50	45	33	53	54	44	39	47	31	54		まんべんなく
…												
遺伝子*i*	100	100	101	100	100	102	102	2	101	100		血液 ↓
…												
遺伝子*k*	50	400	51	50	50	50	51	50	51	1		大腸 ↑ 眼球 ↓

図 1.　様々な組織で取得された遺伝子発現データおよび解析結果（赤枠）のイメージ。ROKU 法は、数万行×数十サンプルの数値行列を入力として、遺伝子ごとに特異的高発現および低発現組織を外れ値として同定するだけでなく、全体的な組織特異性の度合いでランキングすることができます。

この方法は遺伝子ごとに実行するため、全体的な組織特異性の度合いを知る術がないという問題が残されていました。この問題の解決策として、簡単なデータ変換を施した後エントロピーを計算することで、直感をうまく反映した特異性の度合いに基づくランキングが可能になりました[4]。これら2つの要素技術からなる組織特異的発現遺伝子検出法 ROKU に関する一連の仕事は 2006 年までのものですが、今日でも多くのユーザに利用されています[5-6]。

群間で発現の異なる遺伝子の検出

がん組織と正常組織のような比較したい群間で発現の異なる遺伝子を検出したい場合、通常は同一群内の患者（個体）ごとのばらつきを考慮すべく反復データを取得してから解析を行います。図2の仮想データを、「列Ａ～Ｅが計5名のがんサンプル群、列Ｆ～Ｊが計5名の正常サンプル群」からなる2群間比較用だとすると、遺伝子3のような明瞭に異なる（統計的に有意な）遺伝子群を検出するのが目的になります。同様にして、「列Ａ～Ｃが偽薬投与群、列Ｄ～Ｆが薬剤Ｋ投与群、そして列Ｇ～Ｊが薬剤Ｌ投与群」からなる3群間比較用だとすると、遺伝子 i や k の検出が目的となります。

もちろん本稿は異分野の読者向けですので、図2は発現変動遺伝子が正しく発現変動遺伝子として、そうでないものが正しくそうでないものとして簡単に認識できるように前処理が施された後のものを示しています。前述の2群間比較で考えると、遺伝子3が発現変動遺伝子、遺伝子2がそうでないものということになります。答えがわかっている状態で眺めると簡単ですが、当然ながら実際には答えがわからない状態からスタートします。実際のサンプル内（列内）の数値情報は、ほとんどの遺伝子は発現していないか低発現であり、ごく一部の遺伝子が高発現という「べき分布」のような形状となっ

	A	B	C	D	E	F	G	H	I	J
遺伝子1	2	2	1	3	3	100	0	1	2	2
遺伝子2	50	45	33	53	54	44	39	47	31	54
遺伝子3	20	51	38	41	32	2287	2429	2091	2016	2381
..										
遺伝子i	601	583	678	21	42	25	30	36	28	41
遺伝子k	50	52	51	50	50	50	400	453	390	469

図 2.　群間比較用の発現データのイメージ。「A～E 列 vs. F～J 列」のような 2 群間比較の場合には遺伝子 3 が、そして「A～C 列 vs. D～F 列 vs. G～J 列」の 3 群間比較のような場合には遺伝子 i や k が発現変動遺伝子として同定されることになります。

ています。そして図２では最大でも４桁の数値の範囲（ダイナミックレンジ）になっていますが、実際には５～６桁になります。また、全遺伝子中に占める発現変動遺伝子の割合が60%ほどに達するものもあります[7)]。

データの前処理の観点で考えると、たとえ60%ほどであったとしても「列Ａ～Ｅのがん群で高発現となる発現変動遺伝子の数」と「列Ｆ～Ｊの正常群で高発現となる発現変動遺伝子の数」が同程度であれば、実質的には問題になりません。サンプルごとの代表値（平均値や中央値）を揃える正規化がうまく機能するからです。しかしその数の偏りが大きくなるほど、データの正規化が困難になってきます。この場合のデータ正規化の目的は、「発現変動遺伝子でないものの分布を揃える」ことですが、「発現変動遺伝子が存在する（正確には偏りがある）のでデータの正規化がうまくできない」という問題に直面します。一見するとお手上げ状態ですが、通常の手順「データ正規化　→　発現変動遺伝子検出」で得られた発現変動遺伝子数の分布は、ある程度その偏りが反映されます。これは、得られた発現変動遺伝子群を除去した後、もう一度データの正規化を行えばより正確な正規化係数が得られるということを意味します[8)]。この戦略を実装した方法はデータ解析環境Ｒで実行可能なパッケージとして提供されており、非常に多くの研究者に利用されています[9)]。

おわりに

データの前処理や正規化に関する研究は地味ですが、解析結果に大きなインパクトを与える重要な位置を占めています[10)]。今回紹介した知見は、原理的にスパースでダイナミックレンジの広いデータ全般にも応用できると期待しています。赤緑青それぞれ０～255の数値範囲で取り扱う最近流行りの画像解析（物体の認識や分類）分野でも…もしかしたら有効かもしれません。

参考文献

1. Takahashi K, Yamanaka S, *Cell*, 126：663-676, 2006.
2. 上田太一郎 , 応用統計学 , 25：17-25, 1996.
3. Kadota K, Nishimura SI, Bono H, et al., *Physiol. Genomics*, 12：251-259, 2003.
4. Kadota K, Ye J, Nakai Y, et al., *BMC Bioinformatics*, 7：294, 2006.
5. Fukunaga T, Iwakiri J, Ono Y, et al., *Front Genet*, 10：462, 2019.
6. Mohammad S, Page SJ, Wang L, et al., *Nat Neurosci*., 23：533-543, 2020.
7. Zeisel A, Muñoz-Manchado AB, et al., *Science*, 347：1138-1142, 2015.
8. Kadota K, Nishiyama T, Shimizu K, *Algorithms Mol Biol*., 7：5, 2012.
9. Sun J, Nishiyama T, Shimizu, K, et al., *BMC Bioinformatics*, 14：219, 2013.
10. Vieth B, Parekh S, Ziegenhain C, et al., *Nat Commum*., 10：4667, 2019.

門田　幸二（かどた・こうじ）

［**専門**］バイオインフォマティクス、トランスクリプトーム解析
［**主たる著書・論文**］
①共著：RNA-Seq データ解析　WET ラボのための鉄板レシピ , 羊土社 , 2019. ISBN: 978-4-7581-2243-6
②共著：よくわかるバイオインフォマティクス入門 , 講談社 , 2018. ISBN: 978-4-06-513821-2
③単著：シリーズ Useful R 第 7 巻 トランスクリプトーム解析 , 共立出版 , 2014. ISBN: 978-4-320-12370-0
［**現在の所属**］
①東京大学　大学院情報学環・学際情報学府　総合分析情報学コース
②東京大学　大学院農学生命科学研究科　アグリバイオインフォマティクス教育研究ユニット
③東京大学　微生物科学イノベーション連携研究機構
［**所属学会**］日本バイオインフォマティクス学会