テンソル分解を用いた変数選択法による百日咳ワクチンに関するデータ解析
Data analysis on pertussis vaccine using tensor decomposition-based unsupervised feature extraction

物理学専攻　梅木 悠加
department of physics　Haruka Umeki

1.はじめに

先行研究による解析から，百日咳ワクチンに関連する免疫システムは徐々に明らかになりつつある[1]．しかし，今だその全体像は不明であり，さらなる解析が求められている．本研究ではまず初めに遺伝子発現データの解析を行った．これにより，先行研究に加え，百日咳ワクチンの効果的な免疫に関連すると思われる遺伝子を同定することが出来た[2]．だが，先行研究で測定されたのは遺伝子発現データのみでなく，同時に抗体データも採取されており，もし統合解析が叶えばさらなる発見を得られる可能性があった．しかしながら，多くの生命科学データで見られるように，上記のデータも考慮する遺伝子などの変数に対し十分なサンプル数を得られておらず，先行研究や本研究の初期段階においてもそれらの統合解析は困難であった．ところが，最近，テンソル分解を用いた教師なし学習による変数選択法が拡張された，カーネルテンソル分解に基づく教師なし特徴抽出法[3]が提案された．そこで本研究では上記の拡張手法を用いることで，第二段階として百日咳ワクチンデータの統合解析を行い，先行研究に付随したさらなる知見の獲得を目指した[4].

2.解析手法

本研究では，遺伝子発現プロファイル解析は“Unsupervised Feature Extraction Applied to Bioinformatics, A PCA Based and TD Based Approach”(Taguchi, Y. 2020)[5]で提案されている手法を元に，最近，提案された同手法に対する改良として“Adapted tensor decomposition and PCA based unsupervised feature extraction select more biologically reasonable differentially expressed genes than conventional methods”(Taguchi, Y. and Turki, T. 2022)[6]で用いられている標準偏差の最適化手法を合わせて使用し，解析を行った．またマルチオミクス解析では，上記の手法が拡張された“Novel feature selection method via kernel tensor decomposition for improved multi-omics data analysis ”(Taguchi, Y. and Turki, T. 2022)[3]で提案されている手法を元に解析を行った．なお本解析は，統計解析ソフト R を用いて行われた．また Bioconductor にて一般公開されている，テンソル分解を用いた教師なし学習による変数選択法と，カーネルテンソル分解に基づく教師なし特徴抽出法のためのコード含むライブラリ，TDbasedUFE[7]に沿って解析が行われた.

データについて．遺伝子発現データは，アメリカ国立生物工学情報センター（NCBI）のデータベース Gene Expression Omniubus（GEO）からデータシリーズ : GSE152683 をダウンロードした[1]．また抗体データは既存研究の Supplemental Data3[1]から取得された．これらのデータからサンプルと時刻，ワクチンの種類が共通するものを抽出し, 57820 の遺伝子（または 75 の抗体），20 人のワクチン接種者，5 つの時刻（接種後 0 日，1 日，3 日，7 日，14 日），2 つのワクチン（非細胞性ワクチン（aP ワクチン）と全細胞性ワクチン（wP ワクチン））からなる 4 階のテンソルに整形し，解析に用いた.

2.1.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た遺伝子発現プロファイル解析

遺伝子データに，テンソル分解として Tucker 分解を適応すると，x_{ijkl}は次のようにテンソル分解された.

$$x_{ijkl} = \sum_{l_1=1}^{N}\sum_{l_2=1}^{20}\sum_{l_3=1}^{5}\sum_{l_4=1}^{2} G(l_1,l_2,l_3,l_4)\, u_{l_1 i} u_{l_2 j} u_{l_3 k} u_{l_4 l} \tag{2.1}$$

ここで，$u_{l_1 i} \in \mathbb{R}^{N\times N}$，$u_{l_2 j} \in \mathbb{R}^{20\times 20}$，$u_{l_3 k} \in \mathbb{R}^{5\times 5}$，$u_{l_4 l} \in \mathbb{R}^{2\times 2}$は特異値行列．またテンソルは$\sum_i x_{ijkl} = 0$，$\sum_i x_{ijkl}^2 = N$となるように標準化されている．さら

に$G(l_1,l_2,l_3,l_4) \in \mathbb{R}^{N\times 20\times 5\times 2}$はコアテンソルであり，$l_1,l_2,l_3,l_4$に対する重み$G$を表している．

これらを用いて目的の遺伝子を探索する．まず遺伝子選択に用いる$u_{l_1 i}$を特定するために，jに依存しない値，つまりワクチン接種者に依存しない$u_{l_2 j}$を選択した．また 0 日目と区別できる$u_{l_3 k}$，さらに aP ワクチンと wP ワクチンを区別できる$u_{l_4 l}$を選択した(図 1)．

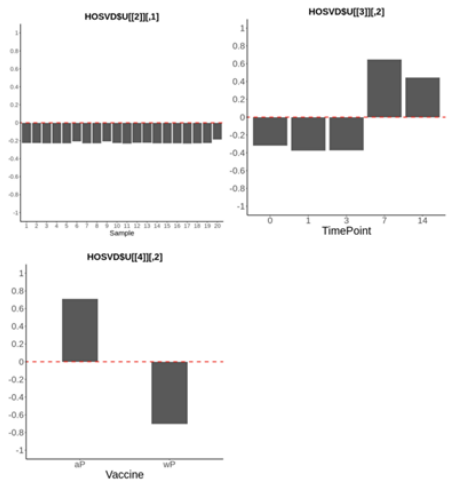

図 1 要件を満たす特異値ベクトル（遺伝子プロファイル解析）

これらの要件を満たすことは，ブースターワクチン接種後，変化した遺伝子の中で，ワクチン間の差を表す遺伝子を選択するために必要である．次に，要件を満たすl_2, l_3, l_4が与えられたときに絶対値が最も大きい$G(l_1,l_2,l_3,l_4)$を特定した（表 1）．

表 1 計算された$G(l_1,1,2,2)$の一部抜粋

l_1	$G(l_1,1,2,2)$	l_1	$G(l_1,1,2,2)$
1	61.721381	6	8.654405
2	32.344485	7	7.146653
3	14.301662	8	3.577691
4	45.188808	9	13.080684
5	13.108894	10	2.017871

このようにしてl_1を選択することによって，要件を満たすj, k, lの依存性と最も関連する$u_{l_1 i}$を特定できる．そして得られた$u_{l_1 i}$はガウス分布に従うという帰無仮説のもとで，各遺伝子にカイ二乗分布を用いて

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{l_1 i}}{\sigma_{l_1}}\right)^2\right] \tag{2.2}$$

のように P 値を付与した．ここで，σ_{l_1}は付与される P 値のヒストグラムの分散を最小にする標準偏差を用いている．

また，$P_{\chi^2}[>x]$は累積カイ二乗分布で値がx以上の確率を表している．これにより，要件を満たすj，k，lの依存性と有意に関連する遺伝子を選択することができる．ただし，遺伝子数が多く，偽陽性を含む可能性があるため，これに加えて Benjamini-Hochberg 法で多重比較補正してから補正 P 値が 0.01 より小さいものを選択すると，4670 個の遺伝子が選ばれた．

2.2.カーネルテンソル分解に基づく教師なし特徴抽出法を用いたマルチオミクス解析

マルチオミクスデータとして，遺伝子発現データと抗体データを用いた．するとi_kは 2 種のマルチオミクスデータ，j_1はサンプル，j_2は時刻，j_3はワクチン種として，

$$x_{i_k j_1 j_2 j_3} \in \mathbb{R}^{N_k\times 20\times 5\times 2} \tag{2.3}$$

と表すことが出来る．ここで，$\sum_{i_k=1}^{N_k} x_{i_k j_1 j_2 j_3} = 0$，$\sum_{i_k=1}^{N_k} x_{i_k j_1 j_2 j_3}^2 = N_k$，$N_k = 57820(k=1$は遺伝子数$)$，$N_k = 75(k=2$は抗体数$)$となっている．

ここで線形カーネルとして，

$$K^k\left(x_{i_k j_1 j_2 j_3}, x_{i_k j_1' j_2' j_3'}\right) = \sum_{i_k=1}^{N_k} x_{i_k j_1 j_2 j_3} x_{i_k j_1' j_2' j_3'}$$

を用い，

$$x_{k j_1 j_2 j_3 j_1' j_2' j_3'} = K^k\left(x_{i_k j_1 j_2 j_3}, x_{i_k j_1' j_2' j_3'}\right) \in \mathbb{R}^{2\times 20\times 5\times 2\times 20\times 5\times 2} \tag{2.4}$$

で表されるテンソルに再整形した．そして，$x_{k j_1 j_2 j_3 j_1' j_2' j_3'}$に HOSVD を適応すると

$$x_{k j_1 j_2 j_3 j_1' j_2' j_3'} = \sum\sum\sum\sum\sum\sum\sum G(l_1,l_2,l_3,l_3,l_5,l_6,l_7)\, u_{l_1 j_1} u_{l_2 j_2} u_{l_3 j_3} u_{l_4 j_1'} u_{l_5 j_2'} u_{l_6 j_3'} u_{l_7 k} \tag{2.5}$$

とテンソル分解することが出来る．ここで，$G(l_1,l_2,l_3,l_3,l_5,l_6,l_7) \in \mathbb{R}^{10\times 5\times 2\times 10\times 5\times 2\times 2}$，$u_{l_1 j_1}, u_{l_4 j_1'} \in \mathbb{R}^{20\times 10}$，$u_{l_2 j_2}, u_{l_5 j_2'} \in \mathbb{R}^{5\times 5}$，$u_{l_3 j_3}, u_{l_6 j_3'} \in \mathbb{R}^{2\times 2}$，$u_{l_7 k} \in \mathbb{R}^{2\times 2}$である．

そして，興味のある特異値ベクトルを探索する．$u_{l_1 j_1}, u_{l_4 j_1'}$は$j_1, j_1'$に対応する各サンプルに依存しない特異値ベクトルを選択し，また時刻 0 日目と区別できる特異値ベクトル$u_{l_2 j_2}, u_{l_5 j_2'}$を選択し，さらに aP ワクチンと wP ワクチンを区別できる$u_{l_3 j_3}, u_{l_6 j_3'}$を選択した．最後にマルチオミクスデータ間で共通する$u_{l_7 k}$を選択した．結果的に，$l_1 = l_4 = 1$，$l_2 = l_5 = 2$，$l_3 = l_6 = 2$，$l_7 = 1$が以上の条件を満たした（図 2）．

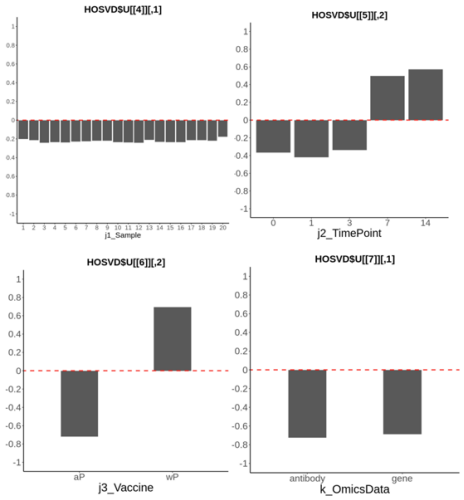

図 2 要件を満たす特異値ベクトル（マルチオミクス解析）

これらを用いて，

$$u_{122i_1} = \sum_{j_1=1}^{20}\sum_{j_2=1}^{5}\sum_{j_3=1}^{2} x_{i_1 j_1 j_2 j_3} u_{1 j_1} u_{2 j_2} u_{2 j_3} \quad (2.6)$$

$$u_{122i_2} = \sum_{j'_1=1}^{20}\sum_{j'_2=1}^{5}\sum_{j'_3=1}^{2} x_{i_2 j'_1 j'_2 j'_3} u_{1 j'_1} u_{2 j'_2} u_{2 j'_3} \quad (2.7)$$

を計算する．そして遺伝子と抗体それぞれに以下の式により P 値が付与された．

$$P_{i_k} = P_{\chi^2}\left[> \left(\frac{u_{122 i_k}}{\sigma_{122}}\right)\right] \quad (2.8)$$

計算された P 値は BH 基準で補正され，$P_{i_k} < 0.01$（遺伝子発現，抗体）に関連するi_kが選択された．

2.3 エンリッチメント解析

選択された遺伝子は，Enrichr[8]にアップロードされた．Enrichr は，複数のライブラリに対してエンリッチメント解析を行い，結果は，補正 P 値が小さい順にランキングされて出力された．そして，有意な P 値が付与された生物学的用語を参照することによって見つかった遺伝子リストの特徴を概観した．

3.結果・考察

まず，遺伝子発現プロファイル解析では 4670 個の遺伝子が同定された．これら遺伝子についてエンリッチメント解析を行うと，感染症や免疫システムに関する用語がヒットしており，さらに具体的な経路について調べると，Th1/Th17/好中球/TLR4/NF-$\kappa\beta$などに関連した用語がヒットしていた．

次にマルチオミクス解析では，3200 個の遺伝子と 14 個の抗体が同定された．まず遺伝子についてエンリッチメント解析を行うと，Th1 応答に関連する用語や Th17 に関連する用語など，遺伝子発現プロファイル解析のときに比べて選択される遺伝子が 1000 個ほど減少しているのにも関わらず類似したエンリッチメント解析結果を得られた．また抗体については，遺伝子のときに用いたエンリッチメント解析のような，適したデータベースが見つからず，特徴的なものを個別に調べた．まず，選ばれた抗体の特徴を概観した．過去の研究において，すべての IgG サブタイプの重鎖をコードする遺伝子の発現は Tdap 接種から 7 日目にピークを迎えるとの報告があったため[1]，7 日目の抗体発現に対してのヒートマップを作成した（図 3）．

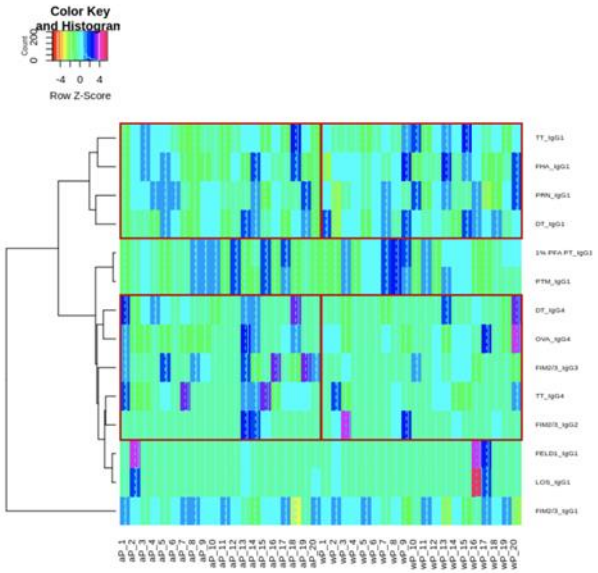

図 3 接種後 7 日目における抗体発現のヒートマップ

すると，7 日目の時点で aP ワクチンに比べて wP ワクチンにおいて多く発現しているとみられる 4 つの IgG1 抗体を特定した．また，aP ワクチンでプライミングされたサンプルで比較的高い発現を示しているように見える 3 つの IgG4 抗体と FIM2/3-IgG2，FIM2/3-IgG3 も同定された．

同定されたこれらの遺伝子と抗体について，それぞれ既存研究を参照すると，Th1/Th17/好中球/TLR4/NF-$\kappa\beta$などは，主に wP ワクチンによって誘導が報告されている免疫機構に合致しており[9]，また抗体は，一部

関連が未知の抗体を含むものの，wP 接種者では IgG1 が支配的となり，aP 接種者では特に IgG4 の生産が増加するという特徴を捉えていた[10]. そのため，これらの結果から，クラス間の差異を表す特徴として妥当性のある結果を得られていると考えられた.

4.結論

百日咳は現行のワクチンの効果不足による再流行が起こっている．そのため新しく効果の高いワクチンの開発が必要とされており，そのためには有益な免疫メカニズムの詳細な理解が必要である．本研究では，aP ワクチンと wP ワクチンで異なる発現を示す遺伝子や抗体をテンソル分解を用いた教師無し学習による変数選択法を用いて同定した．遺伝子を単独で解析した場合も，抗体データと統合的に解析した場合も，選択された遺伝子は，Th1/Th17/好中球/TLR4/NF-$\kappa\beta$ などと関連しており，これらは主に wP ワクチンでプライミングされた被験者で誘導が報告されている特徴である．wP ワクチンの特徴は，百日咳防御に重要な役割を果たすと考えられているもので，第 3 世代ワクチンの開発においてもターゲットとされることが多い．そのため wP ワクチンに関連する遺伝子をリストアップすることは次世代ワクチンの研究開発において有意義であると考えられる．また，抗体についても，それぞれのクラスにおける IgG サブクラスの偏りを検出できていると考えられる．乳児期にプライミングされるワクチンの種類によって生産される抗体とその免疫システムはまだ完全には解明されていない．そのため，これらの結果も有用な免疫効果を促進する抗体システムの理解に活用の可能性が考えられる.

参考文献

[1] da Silva Antunes, R., Soldevila, F., Pomaznoy, M., Babor, M., Bennett, J., Tian, Y., Khalil, N., Qian, Y., Mandava, A., Scheuermann, R. H., et al.: A system-view of Bordetella pertussis booster vaccine responses in adults primed with whole-cell versus acellular vaccine in infancy, *JCI insight*, Vol. 6, No. 7 (2021).

[2] 梅木悠加，田口善弘：テンソル分解による百日咳の不活化した全菌体ワクチンと無細胞ワクチンの差を示す遺伝子の推定，研究報告バイオ情報学（BIO），Vol. 2022-BIO-70, No. 52, pp.1–6 (2022).

[3] Taguchi, Y. and Turki, T.: Novel feature selection method via kernel tensor decomposition for improved multi-omics data analysis, *BMC medical genomics*, Vol. 15, No. 1, pp. 1–12 (2022).

[4] 梅木悠加，田口善弘：カーネルテンソル分解ベースの教師なし特徴抽出を用いた百日咳ワクチンデータにおけるヒト遺伝子発現と抗体量の統合解析，研究報告バイオ情報学（BIO），Vol.2024-BIO-77, pp. 1–5 (2024).

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

[6] Taguchi, Y. and Turki, T.: Adapted tensor decomposition and PCA based unsupervised feature extraction select more biologically reasonable differentially expressed genes than conventional methods, *Scientific Reports*, Vol. 12, No. 1, p. 17438 (2022).

[7] Taguchi, Y.: *TDbasedUFE: Tensor Decomposition Based Unsupervised Feature Extraction* (2023), https://bioconductor.org/packages/TDbasedUFE.

[8] 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., et al.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, *Nucleic acids research*, Vol. 44, No. W1, pp. W90–W97 (2016).

[9] Wilk, M. M., Borkner, L., Misiak, A., Curham, L., Allen, A. C. and Mills, K. H.: Immunization with whole cell but not acellular pertussis vaccines primes CD4 TRM cells that sustain protective immunity against nasal colonization with Bordetella pertussis, *Emerging microbes & infections*, Vol. 8, No. 1, pp. 169–185 (2019).

[10] Diavatopoulos, D. A. and Edwards, K. M.: What Is Wrong with Pertussis Vaccine Immunity? Why Immunological Memory to Pertussis Is Failing, *Cold Spring Harb Perspect Biol*, Vol. 9, No. 12 (2017).