

令和 5 年 6 月 15 日現在

機関番号：14201

研究種目：基盤研究(C)（一般）

研究期間：2020～2022

課題番号：20K06610

研究課題名（和文）T細胞受容体の抗原エピトープを予測する機械学習システムの開発、検証、及び、応用

研究課題名（英文）Development, validation and application of machine learning system for TCR epitope prediction

研究代表者

寺口 俊介（Teraguchi, Shunsuke）

滋賀大学・データサイエンス学系・准教授

研究者番号：00467276

交付決定額（研究期間全体）：（直接経費） 3,300,000円

研究成果の概要（和文）：免疫システムは様々な病原体に個別に対応するために、免疫レパトアと呼ばれる膨大な種類の分子センサーを備えているが、どのセンサーがどの病原体（抗原エピトープ）を認識するかを予測することは困難である。本研究課題では、最新の実験技術を用いて、これらの対応付けのための実験データを取得するとともに、機械学習を用いてそのような予測を行うためのシステム開発を行った。加えて、外部の専門家とともに、特定の病原体における臨床症状の違いが、細胞レベルの免疫レパトアとどのように関わっているかを調べる実験研究も行った。

研究成果の学術的意義や社会的意義

免疫レパトアの各センサーと対応する病原体の対応が正確に予測できるようになると、将来的に血液中の免疫細胞が持つ情報から、その個人の現在の病原体に関する罹患情報はもちろん、これまでの病気の来歴や、将来的な病気への対応力など様々な情報を得ることができると考えられ、直接的な臨床応用の可能性が考えられる。また、学術的にも、同じ実験データから遥かに多くの知見を得ることができると期待される。本研究で取得したデータや予測システムはそのような将来的な応用の基礎となるものである。

研究成果の概要（英文）：The immune system is equipped with a vast array of molecular sensors, called immune repertoire, to individually respond to various pathogens. However, predicting the specific sensor that recognizes a particular pathogen (antigen epitope) remains challenging. In this research, we utilized state-of-the-art experimental techniques to collect empirical data on the associations between molecular sensors and pathogens, and developed a machine learning-based epitope prediction system for these molecular sensors. In addition, we collaborated with external experts to conduct experimental research investigating how the variations in clinical symptoms caused by specific pathogens relate to the immune repertoire at the cellular level.

研究分野：バイオインフォマティクス

キーワード：エピトープ予測 機械学習 免疫レパトア 1細胞遺伝子発現解析 SARS-CoV-2

## 1. 研究開始当初の背景

脊椎動物の免疫システムは免疫レパトアと呼ばれる膨大な種類の免疫受容体を備えており、これらが特定の抗原を認識することで、クローン拡大を起こし、効果的に病原体感染に対応する。また、過去に感染した病原体への再暴露において、迅速な応答を起こすことで、再感染を防ぐといった獲得免疫系の基盤をなしている。免疫レパトアは B 細胞が発現している抗体/BCR レパトアと、T 細胞が発現している TCR レパトアに分けられる。実際に各 BCR, TCR が認識する病原体分子の領域は抗原エピトープと呼ばれている。近年のシークエンシング技術の発展により、免疫レパトアの情報を一細胞レベルで大量に入手することが可能になってきた。このような免疫細胞受容体の情報から、それが認識する抗原エピトープを正確に予測できれば、ワクチン開発、がん治療、自己免疫疾患治療をはじめとした様々な医学応用や、免疫システムのメカニズムを一細胞レベルの解像度で理解する免疫学の基礎研究に決定的な役割を果たすことができると考えられる。しかし、個体が発現する免疫レパトアは膨大なため、特定の病原体/抗原エピトープに対しそれを認識する免疫受容体のいくつかを個別の実験で同定することは可能なものの、逆に、ある免疫受容体が認識する抗原エピトープを特定することは困難であった。

この問題に対するアプローチの一つは、近年様々な分野で活用が進む機械学習を用いることである。研究開始時点でも、既存のデータを用いて、免疫レパトアの配列情報から、対応する受容体の抗原エピトープを予測する試みは存在していたが、実用的な精度が得られるのは十分なデータが存在する特定の病原体/抗原エピトープに対してのみであった。

一方、近年、1細胞レベルの免疫レパトアシークエンシングを利用して、様々なエピトープに対し、それらを認識する免疫レパトア情報を一度に大量に取得する手法(1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシング)が商業化された。この手法を活用した研究が進むことによって、機械学習の訓練データとして利用できるデータ数が大きく拡大するとともに、データの偏りの問題も大きく改善すると期待された。

## 2. 研究の目的

本研究の目的は、機械学習による抗原エピトープ予測システムを開発し、実験による検証と応用を進めることである。研究代表者が研究開始時に所属していた研究室は、免疫レパトアに対するタンパク質立体構造モデリング技術に強みを持っていた。そのため、既存の抗原エピトープ予測システムではほとんど利用されてこなかったタンパク質立体構造情報を追加することで、予測システムの性能向上、さらに可能であれば、汎化性能を持たせていくことを狙っていた。

同時に、既存のデータだけでは、システム開発に十分なデータ量には遠いため、最新の実験手法である、1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシングにより、機械学習のための訓練データを増やしていくことも目的の一つであった。また、共同研究者とのコラボレーションにより、免疫学の基礎研究や臨床研究への応用も計画していた。特に、本研究期間中は、新型コロナウイルス(SARS-CoV-2)が大きな社会的脅威となっており、このウイルス自体やそれに対する免疫応答を理解することが喫緊の課題となっていた。そのため、本研究では、免疫レパトアの観点から新型コロナウイルスを理解することも研究目的の一つに加えることとした。

## 3. 研究の方法

### (1) 機械学習による抗原エピトープ予測システムの開発

訓練データとしては、VDJdbのようなパブリックなデータに加えて、1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシングにより、当研究グループや共同研究者らと取得したデータを利用する。機械学習システムとしては、クラスタリングを用いるもの、ドッキングシミュレーションを用いるもの、深層学習を用いたものなど、複数のデザインを検討する。

### (2) 1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシング

近年、盛んにおこなわれている1細胞遺伝子発現シークエンシングでは、個々の細胞の遺伝子発現データを取得することで、細胞個別の状態を一度に取得し、サンプルに含まれる細胞群の全体像を詳細に捉えることができる。これに加え、本研究で利用した実験系では、特定のエピトープを認識する1細胞免疫レパトアの配列情報を同時に読むことができる。これにより、機械学習システムにおける新たな訓練データを取得するとともに、対象となる免疫システムを多面的な観点から読み解くことができる。このようなデータを詳細に分析することで、免疫学の基礎研究や臨床応用研究を進める。

### (3) 新型コロナウイルスに関わる免疫レパトアの研究

本研究の主要テーマである免疫レパトアは、新型コロナウイルス感染症(COVID-19)を理解するうえでも重要な研究対象である。このため、本研究期間に行った様々な基礎、臨床応用研究においても、COVID-19患者やワクチン接種者のサンプルを利用することで、その免疫応答の解析を行った。

#### 4. 研究成果

##### (1) 機械学習による抗原エピトープ予測システムの開発

本研究では、複数の抗原エピトープ予測システムのデザインの研究を行ったが、その一つは免疫レパトアの配列、及び、(予測された)立体構造の類似性に基づくクラスタリングシステムである。一般に、タンパク質の立体構造予測を行う際には、複数の予測候補が作られるが、これを同時に利用することで、精度の向上を目指す研究を行った。精度の向上はわずかであったが、この研究の際の実装時の工夫により、システムの10倍以上の速度向上を行うことができた。

B細胞の免疫レパトアにおいて、エピトープ領域はわからなくとも抗原となるタンパク質が知られている場合、抗原タンパク質のモデルと抗体のモデルを用いてドッキングシミュレーションを行うことで、エピトープを予測する戦略も有効である。我々のグループでは、最初に抗原と抗体に対し配列情報からエピトープ領域とパラトープ領域をそれぞれ予測した上で、これらの領域に対してドッキングシミュレーションを行い、その結果に基づいてエピトープ領域とパラトープ領域の予測をアップデートする機械学習システムを構築し、Webサーバーとして公開を行った(Davila et al. 2022)。

新たなT細胞受容体エピトープ予測システムのプロトタイプとして、複数のアーキテクチャを検証した。最終的に、大規模言語モデルをはじめ、現在様々な分野で活用が進むアテンション機構に基づく深層学習モデルを採用した。このアーキテクチャの下で、単純な抗原エピトープ予測だけでなく、複数の抗原エピトープ予測や、エピトープが結合しているMHCの種類やアリル等の個別の予測など多数のタスクを同時に学習することで、性能向上を行うことができた。しかしながら、その一方で、研究終了時点でのパブリックな訓練データの総数は、研究計画時の想定程増えておらず、アーキテクチャの改良と訓練データの増加にもかかわらず、T細胞受容体エピトープ予測精度の大幅な向上には至らなかった。これは、他の研究グループを見ても同様の状況である。

また、この間の関連研究分野の大きな進展に、Google傘下のDeepMind社による深層学習モデルであるAlphaFoldによるタンパク質立体構造予測精度の劇的な向上があった(Jumper et al. 2021)。本研究では、本来所属研究室が強みとしていたホモロジーモデリングによる立体構造予測モデルを利用することを想定していたが、AlphaFoldはホモロジーモデリングでは実現できない高精度な立体構造予測を可能にした。実際、前述のドッキングシミュレーションにおいても、そこで用いられていたホモロジーモデルをAlphaFoldによるより精度の高いモデルに置き換えることで、精度の向上が可能であることを実証した(Xu et al. 2022)。

しかもAlphaFoldは免疫レパトアのモデリングに特化しているわけではないため、これを免疫レパトア用にチューニングすることで更なる性能向上も期待できる。実際、そのような試みは行われており(Bradley, 2023)、本研究でもこのモデルを稼働させ、本研究で利用可能か検証を行った。結果としては、現時点では実行スピードが極めて遅く、実用的な利用はできなかった。これらの問題は徐々に解消されていくと考えられるので、このようなモデルの利用は今後の課題である。

##### (2) 1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシング

本研究では、共同研究者らと連携し、健常者、及び、感染症の患者から血液サンプルを取得し、多数の1細胞シークエンシングを行った。特に、本研究では、これまで日本では例の少なかった1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシングを用いることで、日本人に特有のHLAアリルと結合したエピトープを認識する貴重なT細胞受容体の配列情報を得ることができた。一方で、この実験は、一回あたりのコストが高いにもかかわらず、シークエンシングを行い最終的に解析を完了するまで、どの程度のデータがどのようなクオリティで得られるかわからないという欠点が存在する。本研究グループが中心となって行った3度の実験では、基本的に同等の条件での実験を行ったものの、実際に取得できたデータでは、単純な細胞数にして、約3,000、600、7,000細胞と大きなばらつきがあり、また、それら得られたデータ自体のクオリティも大きく異なっていた。このため、本来予定していた通りの実験結果の解釈は難しくなってしまった。一方で、共同研究として行った2回の1細胞シークエンシングでは、このような問題は起きず、再現性の良いデータが取れている。現在、こちらの研究結果の論文化を進めている。

また、複数の1細胞遺伝子発現データを解析する際には、生物学的な差異とは無関係に各実験のバッチごとに特有の変動が現れるバッチ効果の問題が知られており、遺伝子発現データ解析の障害になっている。この問題を軽減するためのバッチ補正の新しいアルゴリズムも共同研究者とともに提案した(Loza et al. 2022)。

##### (3) 新型コロナウイルスに関わる免疫レパトアの研究

COVID-19パンデミックは、多くの1細胞遺伝子発現解析研究の対象となり、本研究課題期間中にも、多数の公開データが公表された。上述の本研究グループが行った1細胞免疫レパトア/エピトープ/遺伝子発現シークエンシングも、COVID-19の患者(回復者)とワクチン接種者のサンプルを用いたもので、新型コロナウイルスに対する免疫応答の理解を試みたものであった。我々の実験は、日本人に多くみられるHLAと結合した様々なT細胞エピトープを利用し、異なる認識特性を持つ対SARS-CoV-2 T細胞受容体の情報を取得した点で、ユニークなものであると考

えられる。

また、我々は、コロナウイルスワクチン抗原の創出を目指した研究において、広いコロナウイルス種に対し感染防御が可能な抗体レパトアを誘導する抗原デザインを提唱した。抗原タンパク質表面にグリカン修飾を導入することで、変異株や近縁のコロナウイルスに保存性の高い部位にエピトープを限定させる手法で、抗体産生応答が特定のレパトアに限定されることを見出した(Shinnakasu et al. 2021)。将来的な *in silico* ドッキングの検証にこの技術が利用できると考えている。それ以外の研究では、全身性エリテマトーデス(systemic lupus erythematosus [SLE])に関連する抗体レパトアを同定し、マウスでの自己抗体産生の再現 (El-Hussien et al. 2022)や、アレルギー性副鼻腔炎患者の鼻腔粘膜におけるアレルゲンの同定(Haruna et al. 2022)を行った。これらの成果で得られたデータは、*in silico* ドッキングの訓練データとして有用である。

#### <引用文献>

- Davila Ana, Xu Zichang, Li Songling, Rozewicki John, Wilamowski Jan, Kotelnikov Sergei, Kozakov Dima, Teraguchi Shunsuke, Standley Daron M (2022) "AbAdapt: an adaptive approach to predicting antibody-antigen complex structures from sequence" *Bioinformatics Advances*, Volume 2, Issue 1, 2022, vbac015
- J. Jumper et al. (2021) "Highly accurate protein structure prediction with AlphaFold" *Nature* volume 596, pages 583-589 (2021)
- Xu Zichang, Davila Ana, Wilamowski Jan, Teraguchi Shunsuke, Standley Daron M (2022) "Improved Antibody Specific Epitope Prediction Using AlphaFold and AbAdapt" *ChemBioChem* 2022, 23, e202200303
- P. Bradley (2023). "Structure-based prediction of T cell receptor:peptide-MHC interactions" *eLife* 2023;12:e82813.
- Loza Martin, Teraguchi Shunsuke, Standley Daron M, Diez Diego (2022) "Unbiased integration of single cell transcriptome replicates" *NAR Genomics and Bioinformatics*, Volume 4, Issue 1, March 2022, lqac022
- Shinnakasu, R., S. Sakakibara, H. Yamamoto, P. H. Wang, S. Moriyama, N. Sax, C. Ono, A. Yamanaka, Y. Adachi, T. Onodera, T. Sato, M. Shinkai, R. Suzuki, Y. Matsuura, N. Hashii, Y. Takahashi, T. Inoue, K. Yamashita and T. Kurosaki (2021). "Glycan engineering of the SARS-CoV-2 receptor-binding domain elicits cross-neutralizing antibodies for SARS-related viruses." *J Exp Med* 218(12).
- El Hussien, M. A., C. Y. Tsai, Y. Satouh, D. Motooka, D. Okuzaki, M. Ikawa, H. Kikutani and S. Sakakibara (2022). "Multiple tolerance checkpoints restrain affinity maturation of B cells expressing the germline precursor of a lupus patient-derived anti-dsDNA antibody in knock-in mice." *Int Immunol* 34(4): 207-223.
- Haruna, S., K. Takeda, M. A. El-Hussien, Y. Maeda, M. Hayama, T. Shikina, K. Doi, H. Inohara, H. Kikutani and S. Sakakibara (2022). "Local production of broadly cross-reactive IgE against multiple fungal cell wall polysaccharides in patients with allergic fungal rhinosinusitis." *Allergy* 77(10): 3147-3151.

## 5. 主な発表論文等

〔雑誌論文〕 計7件（うち査読付論文 7件/うち国際共著 0件/うちオープンアクセス 6件）

1. 著者名 Loza Martin, Teraguchi Shunsuke, Standley Daron M, Diez Diego	4. 巻 4
2. 論文標題 Unbiased integration of single cell transcriptome replicates	5. 発行年 2022年
3. 雑誌名 NAR Genomics and Bioinformatics	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nargab/lqac022	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 Davila Ana, Xu Zichang, Li Songling, Rozewicki John, Wilamowski Jan, Kotelnikov Sergei, Kozakov Dima, Teraguchi Shunsuke, Standley Daron M	4. 巻 2
2. 論文標題 AbAdapt: an adaptive approach to predicting antibody?antigen complex structures from sequence	5. 発行年 2022年
3. 雑誌名 Bioinformatics Advances	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioadv/vbac015	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 Teraguchi Shunsuke, Saputri Dianita S., Llamas-Covarrubias Mara Anais, Davila Ana, Diez Diego, Nazlica Sedat Aybars, Rozewicki John, Ismanto Hendra S., Wilamowski Jan, Xie Jiaqi, Xu Zichang, Loza-Lopez Martin de Jesus, van Eerden Floris J., Li Songling, Standley Daron M.	4. 巻 18
2. 論文標題 Methods for sequence and structural analysis of B and T cell receptor repertoires	5. 発行年 2020年
3. 雑誌名 Computational and Structural Biotechnology Journal	6. 最初と最後の頁 2000 ~ 2011
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csbj.2020.07.008	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 Xu Zichang, Davila Ana, Wilamowski Jan, Teraguchi Shunsuke, Standley Daron M.	4. 巻 23
2. 論文標題 Improved Antibody Specific Epitope Prediction Using AlphaFold and AbAdapt**	5. 発行年 2022年
3. 雑誌名 ChemBioChem	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/cbic.202200303	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

1. 著者名 Haruna Soichiro, Takeda Kazuya, El Hussien Marwa Ali, Maeda Yohei, Hayama Masaki, Shikina Takashi, Doi Katsumi, Inohara Hidenori, Kikutani Hitoshi, Sakakibara Shuhei	4. 巻 77
2. 論文標題 Local production of broadly cross reactive <scp>IgE</scp> against multiple fungal cell wall polysaccharides in patients with allergic fungal rhinosinusitis	5. 発行年 2022年
3. 雑誌名 Allergy	6. 最初と最後の頁 3147 ~ 3151
掲載論文のDOI (デジタルオブジェクト識別子) 10.1111/all.15413	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 El Hussien Marwa Ali, Tsai Chao-Yuan, Satouh Yuhkoh, Motooka Daisuke, Okuzaki Daisuke, Ikawa Masahito, Kikutani Hitoshi, Sakakibara Shuhei	4. 巻 34
2. 論文標題 Multiple tolerance checkpoints restrain affinity maturation of B cells expressing the germline precursor of a lupus patient-derived anti-dsDNA antibody in knock-in mice	5. 発行年 2021年
3. 雑誌名 International Immunology	6. 最初と最後の頁 207 ~ 223
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/intimm/dxab111	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Shinnakasu Ryo, Sakakibara Shuhei, Yamamoto Hiromi, Wang Po-hung, Moriyama Saya, Sax Nicolas, Ono Chikako, Yamanaka Atsushi, Adachi Yu, Onodera Taishi, Sato Takashi, Shinkai Masaharu, Suzuki Ryosuke, Matsuura Yoshiharu, Hashii Noritaka, Takahashi Yoshimasa, Inoue Takeshi, Yamashita Kazuo, Kurosaki Tomohiro	4. 巻 218
2. 論文標題 Glycan engineering of the SARS-CoV-2 receptor-binding domain elicits cross-neutralizing antibodies for SARS-related viruses	5. 発行年 2021年
3. 雑誌名 Journal of Experimental Medicine	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) 10.1084/jem.20211003	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

〔学会発表〕 計0件

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	榊原 修平  (Sakakibara Shuhei)  (10618838)	大阪大学・免疫学フロンティア研究センター・特任助教    (14401)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------