様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２１　日現在

機関番号：　１２６０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１４４３３

研究課題名（和文）シーケンスデータに基づく、免疫レパートリ構造の統計的解析手法の構築

研究課題名（英文）Statistical methods for immuno-repertoire analysis based on TCR sequence data

研究代表者
小林　徹也（Kobayashi, Tetsuya）

東京大学・生産技術研究所・准教授

研究者番号：９０５１３３５９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：獲得免疫を構成する T 細胞などの多様(レパトア) ををを解析するため、次世代シーケンサーにより得られた 個々の T 細胞のレセプター (TCR) 配列の情報を元に、免疫レパトアが持つ隠された構造を明らかにする情報技術を構築を行った。
TCR配列の相同性を保持したまま、データを低次元に射影し、データの低次元構造を取り出した。また低次元空間内での密度分布間の差異を元に異なるサンプル間の差異や、差異を作り出している重要な配列群を同定することに成功した。我々の方法で得られたサンプル間の関連は、先行研究と整合した結果を与えるだけでなく、より高次元性・疎性が高いデータに関して、安定していることが確認された。

研究成果の概要（英文）：In order analyze the diversity of the T cells (repertoire), based on the information of the receptor (TCR) sequence of individual T cells obtained by the next generation sequencer, we constructed computational methods to clarify the hidden structure of immune repertoires.
A low dimensional structure of the data was extracted by projecting the data into a low dimensional space while keeping the similarity relation among TCR sequences. We also succeeded in quantifying the differences between different repertoire samples and in identifying the responsible sequences that make up the sample differences, by using the information-theoretic measure between the density distributions in the low dimensional space. We also confirmed that the relationship between samples obtained by our method not only provides consistent results with previous studies but also stable results even for higher dimensional and more sparse data sets. We also examined the effectiveness of hierarchical statistical models.

研究分野：　定量生物学

キーワード：　T細胞　T細胞受容体　免疫レパートリ　免疫レパトア　適応免疫　複雑系　次世代シーケンサー

１．研究開始当初の背景
我々の体内のＴ細胞は数千万種もの膨大な多様性を持って構成されると推測され、この多様性こそが未知の外敵に対して生体を防御する獲得免疫系の一旦を担っている。この無数の免疫細胞が集団として形成する免疫状態(レパトア)を知ることは、免疫現象の根幹の解明とともに、様々な免疫疾患治療にも重要な情報を提供すると期待される。
近年次世代シーケンサー技術を応用して、個々のＴ細胞からその TCR 配列を計測することが可能になった。しかし、それでも観測できる細胞数は限られており、またさらに理論的に可能な TCR 配列の組み合わせの数と比較して、シーケンスできる配列のみならず、実際に体内に存在する Ｔ 細胞の数も少ない。これは可能な配列で構成される高次元空間に実際のデータやＴ細胞集団はスパース(疎)に存在することを意味する。これらの疎性を回避して、データの持つ構造を捉えることが、Ｔ細胞レパトアを適切に理解し、解析するために不可欠である。

２．研究の目的
本研究では、次世代シーケンサーによって得られた、個々のＴ細胞のレセプター(TCR)配列情報をもとに、免疫レパトアが持つ隠された低次元構造を明らかにする情報技術の構築を目的とする。具体的には、高次元性を持つ TCR 配列集団情報に潜在する構造を、低次元化によって発見的に検証する解析手法と、構造を仮定した統計的階層モデルに基づいて解析する方法とを合わせて構築・検討する。

３．研究の方法
高次元で疎な TCR シーケンスデータの低次元構造を同定するため、低次元空間へのデータの埋め込みと、階層的統計モデルに基づく手法の２つを構築し、その有効性の検証を行う。
低次元化に関しては、TCR の個々のシーケンス長が可変であるため、まずシーケンス間の相同性をすべてのシーケンスの組み合わせに関して計算し、その相同性行列構造を保って低次元空間に埋め込む方法を検討する。一方、階層的統計モデルを基礎とする手法では、配列観測頻度などに隠れ状態を仮定し、MCMC などを利用して、隠れ状態を抽出する方法などを検討する。これらを公開されているＴ細胞レパトアの実データなどに適用し、その結果を検討することによって、データ解析法としての有効性を多角的に検討する。

４．研究成果
[Ｔ細胞レパトアの低次元に基づく解析]
TCR 配列間の相同性を計算する方法として、Smith-Waterman 法を用いて相同性を計算した。Smith-Waterman Algorithm では、置換や欠損などのコストを表すパラメータに依存してその結果が定量的に変わりうる。幾つかの代表的なパラメータを試し、適切なパラメータの探索を行った。
次に、得られた相同性行列に基づき、データを低次元空間に埋め込む方法として、多次元尺度法(MDS)を中心に、t-SNE, Spectral Embedding(SE)、Isomap など複数の方法を検討した。MDS および t-SNE が、サンプル間のクラスタ構造の再現性などの観点から、TCR 配列の埋め込み方法として他の方法よりもよい性質をもつことを見出した。
次に、低次元空間に埋め込みされたデータを異なるサンプルで得られたレパトアデータの間で比較するため、分布間の距離を Jensen-Shannon divergence(JSD)を用いて定量化する方法を検討した。離散データ点から密度分布を推定するため、カーネル密度法による密度分布の推定方法を適用した。また、データ点がない部分に対応するため、一様のベースライン測度を推定に加える対応をした。得られた推定密度分布の間の JSD を計算することで、サンプル間の相同性行列を作成した。
この相同性行列に基づき、レパトアの多様性が大きく制限された変異マウスの公開データを解析した結果、配列観測頻度に基づく先行研究のクラスタリングと極めて近い結果が再現できることを確認した。我々の手法は配列のシーケンス情報だけを用いているため、この結果は、Ｔ細胞レパトアにおいて、配列の観測頻度と配列自体の間に何らかの関連性が存在することを示唆している。得られた相同性やクラスタリング結果の信頼性は、ブートストラップ法を用いて統計的にその有意性を評価した。
これらの結果は、プレプリントを bioXiv に公開し(https://doi.org/10.1101/128025)、現在投稿準備を進めている。
さらに予備的ではあるが、本手法をより現実的で大きなデータセットに適用して有効性を検討するために、新たにマウス TCR レパトアのシーケンス解析を行い、その解析を進めた。また最近公開されたヒトのレパトアデータなどへも解析方法を試し、先行する手法と比べて、データの高次元性・スパース性に起因する問題をうまく回避して、安定なサンプルの比較ができていることを示唆する結果を今のところ得ている。

[階層的統計モデルによる方法]
Ｔ細胞レパトアの持つ潜在構造を階層的統計モデルを仮定して推定するため、Latent Dirichlet Allocation(LDA)の検討を行った。LDA を含め通常の機械学習方法の多くは、固定長のベクトルを入力データとして仮定するので、直接 TCR の配列は用いず、個々の TCR 配列の生成にもちいられた V(D)J 遺伝子の使用頻度分布、特に VJ 遺伝子の分布 p(V,J)に着目して解析を行った。LDA と MCMC による隠れ状態の推定を行った結果、VJ 使用頻度分

布が 3 つの隠れ状態(トピック)でうまく表現できることが確認した。3 のうち一つはノイズ成分を表しているものの、残り 2 つは VJ 遺伝子のゲノム上の位置と相関があることが見いだせた。
TCR 遺伝子はゲノム上の VJ 遺伝子がランダムに組み合わさって形成されるが、この結果は VJ 遺伝子のゲノム上の物理的な位置が、作られる TCR 配列の頻度に影響していることを示唆している。実際に TCR 遺伝子の組み換えは複数回生じ、その際に遺伝子のゲノム上の位置が影響しうることが分子生物学的知見から明らかになっている。これらに基づき、ゲノム上の組み換えプロセスを単純な数理モデルで表現し、TCR シーケンスデータで観測されたものと同様の傾向が再現できるかを検討した。簡単なモデルでは定性的な傾向は再現できるものの、定量的な面では十分な再現性が得られない、という結果を得た。この定量的な齟齬が何らかの未知の生物学的知見を示唆するのか、それともより適切なモデル化で解消しうるものなのかは今後の検討課題である。これらの内容は、27 年度免疫学会で報告し、プレゼンテーション賞を受賞した。

[今後の課題] 本研究により、特に低次元化に基づく方法が、高次元かつ疎性を持つ TCR データの解析に極めて有効であることが確認できた。今後この方法をより現実的かつ応用性の高いデータセットなどに適用し、その有効性や予測性を確認してゆくことが課題となる。また他方で、低次元化の方法は、既存の手法を組み合わせたヒュリスティックスであり、どのような情報が高次元空間から低次元に保存して射影されているかは明らかではない。より理論的な背景のある低次元手法を考案することは、データを適切に解析するために不可欠である。また、個々のシーケンス間の相同性計算やその後の低次元化は、大きなデータセットに関して計算コストのかかるプロセスである。このプロセスをより計算機的に効率化する方法について検討を加える必要があると考えている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

① Taishin Akiyama, Ryosuke Tateishi, Nobuko Akiyama, Riko Yoshinaga, Tetsuya J. Kobayashi, Positive and negative regulatory mechanisms for fine-tuning cellularity and functions of medullary thymic epithelial cells, Frontier Immunology, 2015, Vol.6:461.
②秋山泰身, 小林徹也, システム生物学と免疫系の自己-非自己識別, 医学の歩み, 2016, Vol.259(8), 839-842.

〔学会発表〕（計　12 件）

① Tetsuya J. Kobayashi, Systems-Biological Approaches for TCR Repertoire Analysis, The 44th Annual Meetings of The Japanese Society for Immunology, 札幌コンベンションセンター, 2015:11/15
② Yotaro Katayama, Ryo Yokota, Tetsuya J. Kobayashi, Statistical analysis shows non-randomness of V/J gene selection in human TCR., The 44th Annual Meetings of The Japanese Society for Immunology, 札幌コンベンションセンター, 2015:11/15
③ Kazumasa Kaneko, Ryo Yokota, Taishin Akiyama, Tetsuya J. Kobayashi, Modeling and inferring dynamics of T cell population in thymus, The 2016 (26th) annual meeting of the Japanese Society for Mathematical Biology, Shiiki Hall, Kyushu Univ, Fukuoka, Japan, 2016:9/7-9
④ Ryo Yokota, Yuki Kaminaga, Tetsuya J. Kobayashi, Quantification of the intersample difference in T cell population, 生物物理学会 第 54 回年会, つくば国際会議場,2016:11/25-27
⑤ Kazumasa Kaneko, Ryo Yokota, Taishin Akiyama, Tetsuya J. Kobayashi,Modeling and inferring dynamics of T cell population dynamics in thymus, 生物物理学会, 第 54 回年会,つくば国際会議, 2016:11/25-27
⑥ Ryo Yokota, Tetsuya J. Kobayashi, Quantification of intersample difference in TCR sequences,第 45 回日本免疫学会学術集会, 沖縄コンベンションセンター, 2016:12/5-7
⑦ Kazumasa Kaneko, Ryo Yokota,Taishin Akiyama, Tetsuya J. Kobayashi, Modeling and inferring dynamics of T cell population dynamics in thymus. 第 45 回日本免疫学会学術集会, 沖縄コンベンションセンター, 2016:12/5-7
⑧金子和正, 横田亮, 秋山泰身, 小林徹也, 免疫細胞の分化プロセスの力学系によるモデリング 生命情報科学若手の会, 第 8 回研究会, 北海道伊達市大滝セミナーハウスおよび北海道大学, 2016:10/12-14
⑨横田亮, 神永祐貴, 小林徹也, サンプル間における T 細胞受容体の定量的レパトア解析, 定量生物学の会　第八回年会, 基礎生物学研究所, 2017:1/8-9
⑩金子和正, 横田亮, 秋山泰身, 小林徹也, 力学系モデルによる免疫細胞の分化制御プロセスの理論解析 定量生物学の会, 第八回年会, 基礎生物学研究所, 2017:1/8-9
⑪小林 徹也, 定量免疫学に向けて, 第 1 回理論免疫学ワークショップ 2017, JMS アステールプラザ 広島, 2017:1/19-20

⑫金子和正，横田亮，秋山泰身，小林徹也，胸腺における上皮細胞との相互作用による T 細胞分化制御の理論解析，第１回理論免疫学ワークショップ 2017，JMS アステールプラザ 広島，2017:1/19-20

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計 0　件）

○取得状況（計 0　件）

〔その他〕
ホームページ等 N/A

６．研究組織
(1)研究代表者
小林 徹也（Kobayashi, Tetsuya J.）
東京大学・生産技術研究所・准教授
研究者番号：90513359

(2)研究分担者 N/A

(3)連携研究者
堀 昌平（Hori, Shohei）
国立研究開発法人理化学研究所・統合生命医科学研究センター・チームリーダー
研究者番号：　50392113

横田 亮（Yokota, Ryo）
東京大学・生産技術研究所・特任研究員
研究者番号：　80733154

(4)研究協力者
堅山 耀太郎（Katayama, Yotaro）
神永 祐貴（Kaminaga, Yuki）
金子 和正（Kaneko, Kazumasa）
梶田 真司（Kajita, Masashi）