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

令和　　元 年　　６ 月 １０ 日現在

機関番号：８２４０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００００６４

研究課題名（和文）大規模ゲノムデータの相関構造を考慮した遺伝的予測モデリング

研究課題名（英文）Genetic prediction modeling that accounts for correlation structure in large-scale genomic data

研究代表者

植木　優夫（Ueki, Masao）

国立研究開発法人理化学研究所・革新知能統合研究センター・研究員

研究者番号：１０５１５８６０

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：SNPアレイおよびWGSデータは、超高次元データであり、数十万～数千万のバリアントが含まれる。そのような大規模なゲノム情報を用いて、疾患発症リスクを正確に推定するための数学モデルが必要とされている。事前に、疾患に関与する可能性のあるバリアントをスクリーニングする方法がしばしば利用されるが、ほとんどの方法は周辺的な関連に基づき、ゲノムデータ中の相関構造を無視したものである。連鎖不平衡による相関構造のため、周辺スクリーニングで取り出すことが困難な遺伝的因子を効率的に抽出する方法を開発した。さらに、ゲノムワイドスクリーニングにおける計算コストを削減可能な方法を開発した。

研究成果の学術的意義や社会的意義
SNPアレイやWGSデータなどの網羅的な遺伝情報と生活習慣などの情報を統合し、疾患発症リスクを高精度で算出できれば、個々に最適化された医療（個別化医療）の実現に近づくことができる。しかしながら、これらのゲノム情報は非常に大規模かつ高次元であり、単純な回帰モデルの適用は困難となる。バリアント候補をスクリーニングする次元削減がしばしば行われるが、相関構造を考慮しないことによる予測精度の低下が懸念される。本課題では、相関構造によりスクリーニングにかからないバリアントを利用するための方法を開発した。また、近年のサンプルサイズの大規模化に伴う計算量増大は深刻であるが、統計理論によって高速化を実現した。

研究成果の概要（英文）：The SNP array and WGS data are ultra-high-dimensional data, which include hundreds of thousands to tens of millions of variants. Mathematical models to accurately estimate the risk of developing a disease is required using such large-scale genomic information. Although screening of variants having an effect on disease is often performed in advance, most methods use a marginal association signal that ignores the correlation structure in genomic data. We developed a method to efficiently extract genetic factors that are difficult to retrieve from marginal screening because of correlation structure due to linkage disequilibrium. In addition, we developed a method that can reduce the computational cost in genome-wide screening.

研究分野：統計科学

キーワード：遺伝的予測モデル　ゲノムデータの相関構造　ゲノムワイド関連解析　大規模ゲノムデータ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

次世代シークエンス技術の登場により、密な SNV(一塩基バリアント)の網羅的な利用が可能となった。個人の SNV 情報から疾患発症リスクを定量可能な数学的モデルが確立できれば、個別化医療の実現に大きく近づく。しかしながら現在のところ、多くの複合疾患において、実用に耐える予測モデルは得られていない。このことは、失われた遺伝率問題によるものである。大量の SNV のうち効果をもつものは一部であると考えられるが、効果をもつ SNV を適切に峻別できていないことがひとつの要因として挙げられる。効果をもたない冗長な SNV はノイズとなり、予測精度の大幅低下を招く。一般には、数千万もの SNV(=p)に対して、検体数(=n)は数千～数万という深刻な p >> n 条件にあり、効果をもつ SNV を正確に分離することは現在の技術では困難である。予測精度の向上が求められている。

２．研究の目的

従来より、GWAS(ゲノムワイド関連解析)において単点解析と呼ばれる各 SNV ごとの周辺回帰（対象の SNV 以外を無視した一変量回帰）が遺伝子発見に利用されてきた。周辺回帰は、p >> n 状況下においても常に実行可能であり、予測モデリングにおいても中心的な役割を果たしている。代表的な手法としては、得られた各 SNV の回帰係数を足しあわせた遺伝子スコアが挙げられる。遺伝子スコアは単点解析の結果を利用することから、全ゲノムを一度だけ探索(=p 個の SNV の探索)すれば良いため計算が高速であるが、各 SNV を独立に扱うという点で、遺伝子スコアは相関構造を無視していることになる。特に、密な SNV には連鎖不平衡による高度な相関構造が存在する。相関構造を考慮することでより高精度な予測モデリングが可能となると考えられた。ただし、全ゲノムデータは大規模であるため、実行可能性を有し、なおかつゲノムデータに存在する相関構造を考慮できる統計手法開発を目的とする。

３．研究の方法

ゲノムデータを用いた予測モデリングの出発点となる単点解析を利用した統計解析法について数理的に解析し、新たな手法の開発を行う。特に、ゲノムデータ間の相関構造と予測対象の変数との関係の数理的解析、ならびに単点解析の数学的性質について研究を行う。既存手法や新規開発手法について、実際の SNV データを想定した数値実験によって性能を検証する。また実際の全ゲノムデータ（Alzheimer's Disease Neuroimaging Initiative (ADNI)データ）へと適用を行う。さらに、開発したアルゴリズムを実装したソフトウェアプログラムを作成する。近年のデータ大規模化に対応するため、計算効率化にも取り組む。

４．研究成果

(1)個人のゲノムデータから疾患発症をモデル化できる数学モデルが待望されているが、疾患発症に関与するバリアントはごく一部であり、多くの冗長な情報を含むスパースな超高次元データである。冗長なデータはノイズとして統計モデルの性能を大幅に低下させる。開発した遺伝的予測手法である smooth-threshold multivariate genetic prediction (STMGP)法を R パッケージ stmgp に実装し、CRAN(The Comprehensive R Archive Network)に公開した。

(2)ゲノムワイド関連研究(GWAS)における標準手法である単点解析(つまり各バリアントを一つづつ検査する方法)は、統計遺伝学的解析において中心的な役割を果たしている。単点解析は、高次元小標本のデータであっても実施できる反面、各バリアントを独立として扱うために、バリアント間の相関構造を無視した解析となる。この相関構造は連鎖不平衡によって生じており、大規模ゲノムデータにおいて相関構造が観察される。従来のゲノムワイド関連解析の研究計画は、独立な領域に単一の遺伝的要因が存在することを仮定していたが、複数の遺伝的要因が連鎖不平衡領域に存在する場合には、単点解析などの既存の遺伝的関連解析手法では検出が困難となる状況が起こりうることを、回帰モデルに基づく数理的解析を通じて見出した。このように遺伝的効果が隠される状況においても、遺伝的要因を検出するために、新たに、双方向グラフ上の最短経路を利用して検出を行う統計手法を開発した。

(3)GWAS データや、さらに大規模かつ網羅的な WGS データが取得されるようになっている。それに加えて、多種多様な臨床情報も同時に取得されることも多い。これら超高次元データの統計解析は喫緊の課題である。近年では、多くの被験者を収集するゲノムコホートが世界各地で設立されつつあり、サンプルサイズが非常に大きいことによって統計解析にかかる計算コストも深刻な問題となっている。そのため、複雑な統計手法を利用することはしばしば困難であり、適用可能な手法は限定される。しかしながら、高精度な予測を行うためには、既存の手法では未だ不十分であり、さらなる研究が必要となる。数十万～数千万のバリアントを同時に解析することは難しく、代わりに、各バリアントごとに行う単点解析が利用されることが多い。各バリアントごとに一変量回帰を適用した事前スクリーニングは、予測モデルの次元削減としてよく行われるが、ゲノムコホートのような大規模データでは未だに計算量の問題が大きい。そこで、単点解析に用いる統計量をより詳細に理論的に検討し、計算量を抑えつつ検出力を本質的に改善する方法を考案した。尤度比検定や、Wald 検体に比べて高速計算が可能なスコア検定に着目し、偏相関係数との同値性から Fisher の Z 変換を適用し、検出力を向上させる手法を開発

した。数値実験と実データ適用によって提案法の有用性を確認することができた。

(4)大規模集団コホートを用いて、ゲノム情報や様々な臨床・健康情報の間の関係を調べる研究が各国で進められている。SNP(単一塩基多型)や全ゲノムインピュテーションデータ、全ゲノムシークエンスデータは、数十万から数千万もの変数からなる超高次元なゲノムデータであり、数万人もの参加者に付随する健康情報との関連性を調査することが求められる。GWAS は、各バリアントと興味ある疾患や量的形質間の関連を統計的に調べる手法であるが、超高次元かつ大サンプルサイズのデータにおいては、統計解析に要する計算コストは多大であり非常に深刻である。各バリアントと各形質との間の関係を、単純に一変量回帰によって調べるスクリーニングが GWAS では標準的に用いられているが、大規模サンプルでの計算は高コストとなる。この状況は、高次元の共変量を調整する際に特に深刻となる。尤度比検定やワルド検定に比べて計算が高速となるスコア検定について研究を行い、GWAS でのゲノムワイド探索において、ゲノム情報に含まれる欠測パターンの違いに起因する計算負荷が存在していたが、このボトルネックを漸近理論を用いて解消することで、スコア検定のさらなる高速化を達成することに成功した。本手法は、GWAS のみならず、遺伝子×環境相互作用解析(GxE)にも利用可能な一般的な形式となっている。数値実験と実データ適用によって提案法の有用性を確認することができた。

５．主な発表論文等

〔雑誌論文〕（計　7 件）

1. Numakura Chikahiko, Tamiya Gen, Ueki Masao, Okada Tomoo, Maisawa Shun-ichi, Kojima-Ishii Kanako, Murakami Jun, Horikawa Reiko, Tokuhara Daisuke, Ito Koichi, Adachi Masanori, Abiko Takahiro, Mitsui Tetsuo, Hayasaka Kiyoshi. Growth impairment in individuals with citrin deficiency, 査読有, 2019, 42:501-508, Journal of Inherited Metabolic Disease. doi:10.1002/jimd.12051
2. Sakurai Rieko, Ueki Masao, Makino Satoshi, Hozawa Atsushi, Kuriyama Shinichi, Takai-Igarashi Takako, Kinoshita Kengo, Yamamoto Masayuki, Tamiya Gen. Outlier detection for questionnaire data in biobanks. International Journal of Epidemiology, 査読有, 印刷中, doi:10.1093/ije/dyz012
3. Obara Taku, Ishikuro Mami, Tamiya Gen, Ueki Masao, Yamanaka Chizuru, Mizuno Satoshi, Kikuya Masahiro, et al. Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. Scientific Reports, 2018, 査読有, 8, doi:10.1038/s41598-018-33110-w
4. Shimamura Kaito, Ueki Masao, Kawano Shuichi, Konishi Sadanori. Bayesian generalized fused lasso modeling via NEG distribution. Communications in Statistics- Theory and Methods, 査読有, 印刷中. doi: 10.1080/03610926.2018.1489056
5. Sato Shuntaro, Ueki Masao. Fast score test with global null estimation regardless of missing genotypes. PLoS ONE 13: e0199692. 2018. doi: 10.1371/journal.pone.0199692
6. Ueki Masao. Enhancing power of score tests for regression models via Fisher transformation. Journal of the Japanese Computational Statistics, 査読有, 30.2:37-53, 2018. doi:10.5183/jjscs.1702001_234
7. Ueki Masao, Kawasaki Yoshinori, Tamiya Gen. Detecting genetic association through shortest paths in a bidirected graph. Genetic Epidemiology, 査読有, 40:481-97, 2017. doi:10.1002/gepi.22051

〔学会発表〕（計 14 件）

1. 植木優夫. 全ゲノム配列情報を用いた疾患発症予測に向けて. 科研費研究集会「予測モデリングとその周辺-機械学習・統計科学・情報理論からのアプローチ-」 東京. 2018
2. 植木優夫. ゲノムデータ解析の理論と実践. リスク解析戦略研究センターシンポジウム. 東京. 2018
3. 植木優夫, 川﨑能典, 田宮元, 最短経路を利用した遺伝関連解析, 科研費研究集会・「生命・自然科学における複雑現象解明のための統計的アプローチ」, 滋賀, 2018
4. Ueki M, Tamiya G. Rapid and accurate genetic predictive modelling for large-scale genetic study, IASSL 3rd INTERNATIONAL CONFERENCE, Colombo, Sri Lanka, 2017
5. 植木優夫, 田宮 元, スパースモデリングによる大規模ゲノムコホート解析, 科研費集会・「スパースモデリングの深化と高次元データ駆動科学の創成」最終成果報告会, 東京, 2017
6. Ueki M, Kawasaki Y, Tamiya G. Detecting genetic association through shortest paths in a bidirected graph, 日本分類学会・IFCS2017, Tokyo, 2017
7. Ueki M, Kawasaki Y, Tamiya G. Detecting genetic association through shortest paths in a bidirected graph, ISI-ISM-ISSAS Joint Conference Tokyo 2017, Tokyo, 2017

8. 植木優夫，川崎能典，田宮元，最短経路を利用した遺伝関連解析，科研費研究集会「スパースモデリングの深化と高次元データ駆動科学の創成」2017 年度第 1 回公開シンポジウム，東京，2017
9. Ueki M. Rapid and accurate genetic predictive modeling for large-scale genetic study, ISI-ISM-ISSAS Joint Conference, Delhi, 2017
10. 植木優夫．複数遺伝子の複合的効果を考慮したゲノムデータ解析法．科研費（基盤 S）＆ 第 8 回生物統計ネットワークシンポジウム 「統計科学が切り拓く個別化医療：方法論・実践のフロンティア」，福岡，2017
11. 植木優夫，田宮元．Smooth-threshold multivariate genetic prediction with unbiased model selection. 2016 年度公開シンポジウム．新学術領域研究「スパースモデリングの深化と高次元データ駆動科学の創成」，東京，2016
12. Ueki M, Tamiya G. Smooth-threshold multivariate genetic prediction with unbiased model selection. Annual Meeting of the American Society for Human Genetics; Vancouver, 2016
13. Ueki M, Tamiya G. Smooth-threshold multivariate genetic prediction with unbiased model selection. 2016 International Conference for JSCS 30th Anniversary in Seattle, Seattle, 2016
14. Ueki M, Tamiya G. Smooth-threshold multivariate genetic prediction with unbiased model selection. The 13th International Congress of Human Genetics; Kyoto, 2016

〔図書〕（計 3 件）
1. 植木優夫，田宮元．新版 医学統計学ハンドブック（丹後俊郎 ・松井茂之 編） 25.1 章 朝倉書店 2018
2. 植木優夫，田宮元．遺伝統計学と疾患ゲノムデータ解析-病態解明から個別化医療，ゲノム創薬まで-(遺伝子医学 MOOK33 号) (遺伝子医学 MOOK 33) (岡田 随象 編) 第 1 章 1 節 メディカルドゥ 2018
3. 冨田誠，植木優夫 「ゲノムデータ解析」共立出版 第 3 章 2016

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等
R package: stmgp
https://cran.r-project.org/web/packages/stmgp/index.html

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：

所属研究機関名：

部局名：

職名：

研究者番号（8 桁）：

(2)研究協力者

研究協力者氏名：

ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。