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　４ 年　 ５ 月 ２９ 日現在

機関番号：１４４０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2021
課題番号：２０Ｋ２１８３４

研究課題名（和文）遺伝統計学と最適化理論の学際連携による大規模ゲノム情報の再解釈

研究課題名（英文）Re-annotation of large-scale human genome data by integration of statistical genetics and operations research

研究代表者

岡田　随象（Okada, Yukinori）

大阪大学・医学系研究科・教授

研究者番号：７０７２７４１１

交付決定額（研究期間全体）：（直接経費）　　5,000,000円

研究成果の概要（和文）：大規模ゲノム情報は、①：長大なゲノム配列の中で一部の遺伝子変異のみが関与する「スパース性」、②：遺伝子変異数がサンプル数と比べて著しく大きい「P>>N問題」、が特徴である。遺伝統計学で扱われるヒト集団ゲノム情報は単純なグラフ構造として記述でき、即ち組合せ最適化問題として解釈可能である。本研究は、遺伝統計学と最適化理論の学際連携を通じて、大規模ヒトゲノム・臨床情報の組合せ最適化問題としての再定義を目指すものである。

研究成果の学術的意義や社会的意義
遺伝統計学と最適化理論は共通した理論的背景を有するも、異なる学問分野として捉えられてきた。本研究は、大規模疾患ゲノム情報をシンプルな行列・グラフ情報として捉えることにより、る数理理論の応用研究の題材となることを示し、学際連携の新たな可能性を切り拓いたものとして学術的な意義を有すると考えられる。今後、解析アルゴリズムのスケーラビリティーの獲得と高速計算化を進めることで、より汎用性の高い情報解析ツールとしての実装が可能になると期待される。

研究成果の概要（英文）：Large-scale human genome data has two major characteristics; (i) limited fractions of the human genome variations only affects human disease risk, (ii) numbers of the human genome variations are much larger than those of samples (i.e., P>>N problem). Statistical genetics handles human population genome data as input, which can be described as simple graphs. We considered that these graphs can be solved by operations researches. This project aims re-annotation of large-scale human genome data by integration of statistical genetics and operations research through iterdisciplinary cooperative studies.

研究分野：遺伝統計学

キーワード：遺伝統計学　最適化理論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

次世代シークエンサーに代表されるゲノム配列解読機器の発達により、大容量のヒトゲノム情報が日常的に出力される時代が到来した。一方で、これらの大規模ヒトゲノム情報を適切に解釈し、個人の疾患発症予測や最適な治療法の提供といった、ゲノム個別化医療へとつなげる具体的な道筋は未だ見いだされていない。

大規模疾患ゲノム情報は、①：長大なゲノム配列の中でごく一部分の遺伝子変異のみが疾患感受性に関与しているという「スパース性」、②：遺伝子変異数（約 1000 万か所）がサンプル数（数百~数万人）と比べて著しく大きいという「P>>N 問題」、の 2 点が特徴である。問題解決のために、ヒトゲノム解析を対象とする遺伝統計学において、線形/非線形の機械学習やペナルティ付き回帰モデルの適用が試みられてきた。しかし、ゲノム個別化医療の社会実装を可能にする高精度の発症予測モデルには至らず、さらなる手法開発が望まれていた。

最適化理論とは、ある制約のもとで最も良い解を見つけるための数理科学方法論として開発されてきた。情報科学や都市・金融工学など多彩なリアルワールドデータへの応用研究が盛んである。特に、グラフなど組合せ構造を持つ最適化問題に対しては、数理的構造に着目した効率的なアルゴリズム設計がなされてきた文献１。ヒト集団ゲノム情報は比較的単純なグラフ構造として記述でき、大規模ヒトゲノム情報解析は組合せ最適化問題として再解釈が可能であるが、異分野をつなぐ学際的研究活動が必要となるため、試みられてこなかった。

２．研究の目的

本研究の目的は、組合せ最適化問題として大規模ゲノム情報を再解釈し、社会実装への道を切り拓く点にある。大規模疾患ゲノム情報は、①：長大なゲノム配列の中で一部の遺伝子変異のみが疾患に関与する「スパース性」、②：遺伝子変異数がサンプル数と比べて著しく大きいという「P>>N 問題」、が特徴である。これまで遺伝統計学に基づく問題解決が試みられてきたが、高精度の発症予測モデルには至らず、更なる手法開発が望まれていた。最適化理論は、ある制約のもとで最も良い解を見つけるための数理科学的手法であり、組合せ構造を持つ最適化問題を中心に応用研究がなされてきた。ヒト集団ゲノム情報は単純なグラフ構造として記述でき、即ち組合せ最適化問題として再解釈が可能である。本研究は、学際連携による遺伝統計学と最適化理論の邂逅を通じて、大規模ヒトゲノム・臨床情報を組合せ最適化問題として再定義するアルゴリズム構築を目指す。

３．研究の方法

(1)：組合せ最適化問題としてのヒト集団ゲノム情報の再定義

研究分担者：垣村尚徳（慶應義塾大学）の主導の元、組合せ最適化問題として、ヒト集団ゲノム情報の再定義を行った。集団中のヒトゲノム情報は、ヒトゲノム配列に数百万箇所に存在する遺伝子変異（例：一塩基多型、SNP）の各部位について、各サンプルを行、遺伝子変異を列とする行列で記述される。両親由来の 2 通りの遺伝情報を有する 2 倍体であるため、各遺伝子変異アレルの保有数に基づき、行列の構成要素は 0，1，2 のいずれかの整数（もしくは推定確立を反映した小数）で記述される。しかし、遺伝子変異分布のスパース性よりデータ構造の冗長性が増し、また異なる遺伝子領域における遺伝子変異の照合に制約が生じていた。本研究では、集団内における遺伝子変異の時系列に沿った拡散過程を反映した樹形図（ancestral recombination graph; ARG）として、ヒトゲノム情報の再構築を実施した。2。構築されたグラフ構造における各ノード（＝遺伝子変異）間の縦断的なつながりを

一定の制約のもとで許容し（例：3 遺伝子変異までの相互作用を許容)、各ノードの医学生物学的機能注釈に基づく方向性を持った重みづけ（例：ゲノムワイド関連解析により同定された疾患感受性遺伝子変異、蛋白質アミノ酸配列置換変異、遺伝子発現制御変異）を与えることで、高次元の相互作用を考慮した樹形図としてヒト集団ゲノム情報を再定義した。

(2)：国際共同研究による大規模ゲノム・臨床情報の収集

研究代表者を中心に、大規模疾患ゲノム情報・臨床情報の収集とモデル化を行う。バイオバンク組織より分譲された大規模ゲノム・臨床データに加え、所属施設・関連附属病院にて構築された疾患ゲノム情報を対象とした。国際バイオバンク連携を通じた 50 万人のデータも対象とした。全ゲノムシークエンスを効率的な遺伝子変異同定が可能となる中程度の深度（20×～）で実施し、研究計画の進展に伴う新たなゲノムデータ構築を行った。

(3)：ゲノム個別化医療の実装を可能にするアルゴリズム構築（令和 3 年度）

再定義を行った組合せ最適化問題としてのヒト集団ゲノム情報のモデリングを、収集した大規模疾患ゲノム情報に対して適用した。大規模疾患ゲノム情報の樹形図としてのグラフ構造化と、複数の疾患や形質を対象とした予測モデルの構築を行った。

４．研究成果

学際連携による遺伝統計学と最適化理論の融合を通じて、大規模ヒトゲノム情報の組合せ最適化問題としての再定義とグラフ構造としての解釈を試みた。さらに、個人の疾患発症予測精度の向上と本邦のゲノム個別化医療の社会実装を可能にするアルゴリズム構築を行った。並行して、解析対象となるヒトゲノム情報の構築を全ゲノムシークエンス解析を中心に進めた。ARG については国際バイオバンク連携を通じて日本人集団と欧米人集団の双方において実装例が可能となっった。今後、数十万人規模での大規模 ARG 推定とヒト疾患感受性遺伝子変異情報との統合解析を進めていく結果である。

特に、ヒトゲノム領域全域に分布する多数の遺伝子変異情報を個人のジェノタイプ情報に基づき統合して疾患発症リスクを推定留守 Polygenic Risk Score（PRS）に注力した成果が得られた。ゲノムワイド関連解析を通じて得られるゲノムワイド関連統計量に基づく PRS 推定を題材として、最適化理論を活用した予測精度の向上に関する解析を実施した。本研究を通じて日本人集団のゲノム・臨床情報を対象に様々な手法を用いて PRS を計算したところ、ベイズ推定など基づき領域内の LD 構造を最適化した上で対象遺伝子変異を抽出する手法群において PRS 予測精度の向上が認められることが明らかになった。更に、ヒト白血球型抗原（human leukocyte antigen; HLA）などヒトゲノム配列上で複雑なゲノム構造を有する領域については、最適化理論を活用したハプロタイプ構造の推定と、ハプロタイプに基づく疾患リスク推定が有用なことを示した。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　2件／うちオープンアクセス　3件）

１．著者名 Naito Tatsuhiko、Suzuki Ken、Hirata Jun、Kamatani Yoichiro、Matsuda Koichi、Toda Tatsushi、Okada Yukinori	４．巻 12
２．論文標題 A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes	５．発行年 2021年
３．雑誌名 Nature Communications	６．最初と最後の頁 1639
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-021-21975-x	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 Shi Huwenbo、Gazal Steven、Kanai Masahiro、Koch Evan M.、Schoech Armin P.、Siewert Katherine M.、Kim Samuel S.、Luo Yang、Amariuta Tiffany、Huang Hailiang、Okada Yukinori、Raychaudhuri Soumya、Sunyaev Shamil R.、Price Alkes L.	４．巻 12
２．論文標題 Population-specific causal disease effect sizes in functionally important regions impacted by selection	５．発行年 2021年
３．雑誌名 Nature Communications	６．最初と最後の頁 1098
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-021-21286-1	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 該当する

１．著者名 Atkinson Elizabeth G.、Maihofer Adam X.、Kanai Masahiro、Martin Alicia R.、Karczewski Konrad J.、Santoro Marcos L.、Ulirsch Jacob C.、Kamatani Yoichiro、Okada Yukinori、Finucane Hilary K.、Koenen Karestan C.、Nievergelt Caroline M.、Daly Mark J.、Neale Benjamin M.	４．巻 53
２．論文標題 Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power	５．発行年 2021年
３．雑誌名 Nature Genetics	６．最初と最後の頁 195～204
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41588-020-00766-y	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 該当する

〔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	垣村　尚徳 (Kakimura Naonori) (30508180)	慶應義塾大学・理工学部（矢上）・准教授 (32612)	

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