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　２７　日現在

機関番号：　１７１０２
研究種目：　若手研究(A)
研究期間：　2013～2016
課題番号：　２５７００００４

研究課題名（和文）複数の遺伝要因及び環境要因の組み合わせを考慮したゲノムワイド相関解析法の開発

研究課題名（英文）A GWAS study considering multiple genic/invironmental factors

研究代表者
西郷　浩人（Saigo, Hiroto）

九州大学・システム情報科学研究院・准教授

研究者番号：９０５８６１２４

交付決定額（研究期間全体）：（直接経費）　　4,700,000円

研究成果の概要（和文）：（１）分子限定法による組み合わせ空間の探索と多重検定の補正：組み合わせの空間の探索により見つかる膨大な数の遺伝子の組み合わせには多くの偽陽性が含まれるので、Taroneの補正を行うことで明らかな偽陽性を探索空間から取り除いた。その結果、従来より高速で高い検出力の実現に成功した。
（２）カーネル法による組み合わせ空間の大きさの見積もり：多項式カーネルとリッジ回帰を利用した計算機実験の結果、遺伝子数に比べてサンプル数がある程度確保出来る時には、６つまでの相互作用の検出が出来ることを示した。また、本手法をマウスデータに適用したところ、文献と一致する原因遺伝子の染色体上での位置の特定に成功した。

研究成果の概要（英文）：(1)Traversing combinatorial space with brand-and-bound search and rigorous multiple testing correction. A combinatorial space spanned by combination of genes contains a large number of false positives. We devised to remove them by employing Tarone’s correction, which resulted in a faster search with higher statistical power.
(2)Estimating the size of combinatorial space by kernel methods: In our computational experiments equipped with polynomial kernels and kernel ridge regression, we have shown that detection of up-to-six degrees interaction was possible. By applying the same method to mouse genotype/phenotype data, we have successfully detected the region in a chromosome that harbors causal genes.

研究分野：　Machine Learning, Bioinformatics

キーワード：　GWAS　SNP　variable selection　kernel methods　hypothesis testing　model selection

１．研究開始当初の背景

近年のゲノムワイド相関解析の進展により疾患遺伝子の特定が進んでいるが、これらの方法の多くは単一の遺伝要因が疾患の原因であるというモデルに基づいた単遺伝子の探索である。一方で、多くの複雑な疾患は複数の遺伝要因と環境要因が合わさることにより起こると考えられている。複数要因の特定が進まない理由は組み合わせの数え上げが計算科学的に困難だからである。

２．研究の目的

京都大学の山中博士が iPS 細胞を完成させるのに 1000 個の遺伝子から 4 個の遺伝子を選ぶ必要があったように、重要な要因の組み合わせを選ぶ問題の重要性は増すばかりである。本提案研究ではデータマイニングと統計の手法を用いてこの問題に取り組む。

３．研究の方法

本提案研究では、病気などの個人差を生み出す遺伝形の組み合わせを探索する方法の開発を目指す。提案手法は一種の分枝限定法であり、ある条件を満たす遺伝形の組み合わせを網羅的に探索する。但し、計算量の爆発を抑える為に探索空間には適切な枝狩りを導入する。このために、統計において知られる Cp, AIC や BIC といった情報量基準を採用して出来るだけ小さな節約モデルを構築する。探索空間の定義には組み合わせを効率的に数え上げるアイテムセットマイニングの技術を拡張することにより高速な実装の開発を目指す。また、これとは独立に遺伝形の作用構造や相互作用数の上限を推定するカーネル法を開発する。これらの推定値は分枝限定法における枝狩り条件に利用することにより更なる探索空間の絞り込みに利用することが出来るので有用である。

４．研究成果

（１）分子限定法による組み合わせ空間の探索と多重検定の補正

組み合わせ空間の探索における問題点は２つある。

① 探索の為に必要な計算量が指数関数的に増加する問題
② 見つかった組み合わせ仮説に対する多重検定の補正問題

本研究では、Terada らの提案した方法に基づき、まず②に着目し、多重検定の際に１回以上の偽陽性が生じる割合である Family-Wise Error Rate(FWER)の上限を制御する方法を採用した[Terada et al.2013]。当手法の利点は、FWER の上界を変化させない組み合わせ仮説は探索の対象から外すことにより、①を実現できることにある。ただし、Terada らは各組み合わせ仮説に付随する応答変数として離散値だけを扱ったのに対して、本研究では連続値を扱えるような拡張を行った。この結果、図１に示すように、従来手法(Bonferroni)に対して、異なる範囲のパラメータ ε にわたって、FWER を制御できる実験結果を得た。(ε は小さいほど区別が難しい。) 特に ε が大きい領域において FWER の制御に差が出ていることがわかる。

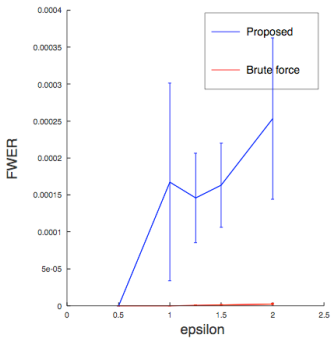

図 1、従来手法(Beonferroni)と提案手法の FWER による比較［Inokuchi et al. 2016］

さらに、提案手法では従来手法と比べて探索空間の枝刈りをすすめた結果、効率的な探索に成功している。図２では、入力のサイズに関わらず、提案手法が高速であることが分かる。

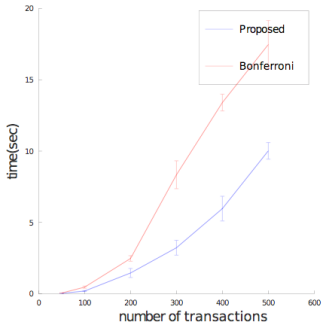

図２、従来手法と提案手法の計算時間の比較［Inokuchi et al. 2016］

また、FWER の正しい補正により、偽陰性の数も減らすことに成功している。図３では、従来手法よりも検定の検出力(Power)が高いことが分かる。特に、 ε の大きい領域においてその差は顕著である。

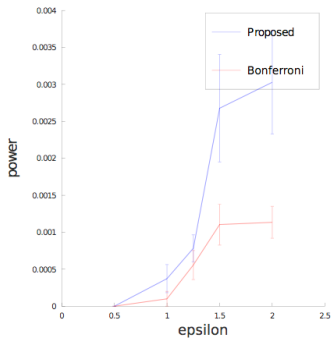

図３、従来手法と提案手法の検出力(Power)の比較［Inokuchi et al. 2016］

（2）カーネル法による組み合わせ空間の大きさの見積もり

入力の変数（遺伝子）の数が大きい時は、組み合わせの空間の探索は非常に困難である。そのような状況を配慮して本研究で提案したのは、カーネルリッジ回帰と交差検証を利用した原因遺伝子を含む領域の推定である。この際、与えられた範囲内であれば、2 つ以上の遺伝子の相互作用も考慮できるようにした点が独創的である。

図 4 にシミュレーションの結果を示す。個体数(n=1000)に対して、遺伝子の数(p=100)程度であれば、比較的高次の相互作用数の推定にも成功していることが分かる。

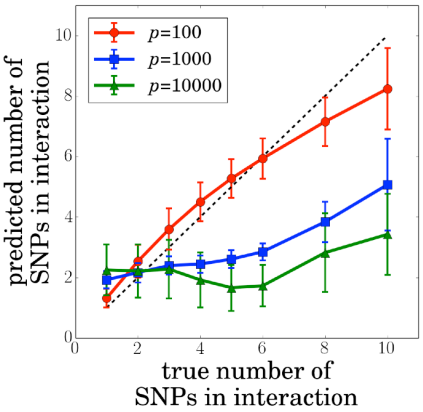

図 4、真の相互作用次数に対して予測された相互作用次数(n=1000)[Kodama and Saigo 2016]

また、開発した手法をオックスフォードマウスデータに対して適用した際の結果を図 5 に示す。「fear conditioning time freezing cue」という表現型に関与する遺伝子を含む 15 番染色体を正しく予測していることがわかる。

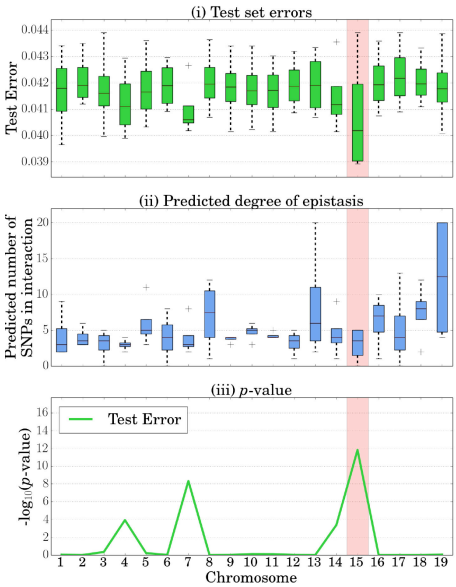

図 5、マウスデータにおける表現型の原因遺伝子位置の予測結果[Kodama and Saigo 2016]

また、図 6 に示すように、染色体の中でも原因遺伝子の場所を正しく絞り込むことに成功している。

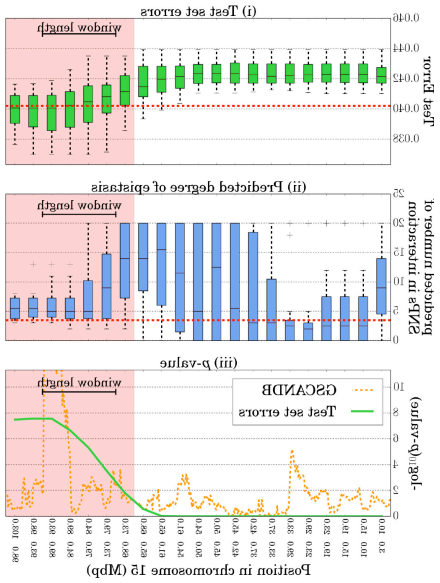

図 6、マウスデータの 15 番染色体における原因遺伝子位置の予測結果 [Kodama and Saigo 2016]

このように本手法の利点は、原因である遺伝子（変数）の領域と相互作用の最大次数を提示出来る点である。一方で弱点は、染色体上の位置等の何らかの事前知識が必要な点である。

〔参考論文〕
A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. PNAS, 110(32):12996-13001, 2013.

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 4 件）
①Ismail, H.D., Saigo, H., Bahadur, K.C.D., RF-NR: Random forest based approach for improved classification of Nuclear Receptors , to appear in IEEE Transactions on Computational Biology and Bioinformatics, 2017, 査読有り

② Kodama, K., Saigo, H.: KDSNP: a Kernel-based approach to Detecting high-order SNP interactions Journal of Bioinformatics and Computational Biology 14(5), 1644003, 2016, 査読有り

③Suryanto, C. H., Saigo, H., Fukui, K.: Structural Class Classification of 3D Protein Structure Based on Multi-View 2D Images IEEE Transactions on Computational Biology and Bioinformatics , Vol.PP, Issue:99 (08 2016), 査読有り

④Shao, Z., Hirayama, Y., Yamanishi, Y., Saigo, H.: Mining discriminative patterns from graph data with multiple labels and its application to QSAR Journal of Chemical Information and Modeling , 55(12), 2519-27 (12 2015), 査読有り

〔学会発表〕(計 5 件)
①井ノ口敬章, 永野竜輝, 西郷浩人, 応答変数が連続値の際の組み合わせ仮説に対する多重検定補正法, 第 103 回人工知能基本問題研究会, 湯布院公民館（大分県·由布市）, 2017.03.13.

②Tabei, Y., Saigo, H., Yamanishi, Y., Puglisi, S., Scalable Partial Least Squares Regression on Grammar- Compressed Data Matrices , ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2016), 1875-1884, San Francisco, U.S., August 13-17, 2016.

③ Kodama, K., Saigo, H.: KDSNP: a Kernel-based approach to Detecting high-order genetic Epistasis International Conference on Genome Informatics (GIW2016), Shanghai, China, October 3-5, 2016.

④Ismail, H.D., Saigo, H., Bahadur, K.C.D., RF-NR: Random forest based approach for improved classification of Nuclear Receptors , International Conference on Genome Informatics & International Conference on Bioinformatics (GIW/InCoB2015), Tokyo, Japan. (9 2015)

⑤山口拓郎, 西郷浩人, カイ二乗検定による p 値の下限値を利用した遺伝子相互作用の効率的な数え上げ, 第 42 回バイオ情報学研究会, 沖縄科学技術先端機構（沖縄県·国頭郡）, 2015.06.23.

〔図書〕(計 1 件)
①西郷 浩人, QSAR モデルの構築; 機械学習と部分構造マイニングによるアプローチ, 日本化学会情報化学部会誌, 2013, 4

〔産業財産権〕
なし

〔その他〕
ホームページ等
http://www.i.kyushu-u.ac.jp/~saigo/

６．研究組織
(1)研究代表者
西郷 浩人 (SAIGO, Hiroto)
九州大学·大学院システム情報科学研究院·准教授
研究者番号：90586124