様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　３０　日現在

機関番号：　７２６０２
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３００５４

研究課題名（和文）大規模ゲノムデータから遺伝子間相互作用を検出する統計的方法の開発

研究課題名（英文）Development of statistical methods for detecting gene-gene interactions from large-scale genomic data

研究代表者

牛嶋　大（USHIJIMA, Masaru）

公益財団法人がん研究会・ゲノムセンター・研究員

研究者番号：６０３２８５６５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：網羅的なSNP情報から遺伝子間相互作用を探索する際に、簡単な統計量を用いることによって従来より20倍程度高速に探索できるアルゴリズムを開発した。同様の方法を用いて、遺伝子発現データからの遺伝子間相互作用の検出、3遺伝子の相互作用を検出するアルゴリズムも開発した。応用として、Lassoクラスタリング法を用いて乳がんマイクロアレイデータの解析を行い、新規のサブタイプ分類や関連遺伝子の探索を行った。

研究成果の概要（英文）：We developed a statistical method for detecting gene-gene interactions using simple statistics from genome-wide SNP data. This program is about 20 times faster than the program we developed previously. In addition, we developed analytical programs which detect gene-gene interactions from microarray data and three-way interactions from SNP data in the similar way. As an application of our method, we performed Lasso clustering to the breast cancer microarray data to detect a novel subtype and related genes.

研究分野：　生物統計学

キーワード：　多重検定　SNP　ゲノムワイド関連解析　遺伝子間相互作用　マイクロアレイ　クラスタリング

１．研究開始当初の背景

DNA マイクロアレイが網羅的遺伝子発現解析に用いられるようになって約 10 年が経過し、癌をはじめとしてさまざまな疾患に対して薬剤治療効果、良悪性診断、サブタイプ分類などと、遺伝子発現との関連について調べられてきた。同様に一塩基多型（SNP: Single Nucleotide Polymorphism）に対しても、一度に 50 万個の SNP が容易に調べられるようになった。その結果、単独の遺伝子や SNP が表現形に直接寄与するということは珍しく、複数の遺伝子が相互作用して表現形に影響を与える場合がほとんどであることが明らかになってきた。歴史的には、複数遺伝子による作用としては、集団遺伝学の分野で古くからエピスタシス（epistasis）という概念がある。本来は「ある遺伝子座が別の遺伝子座に及ぼす相乗的効果」を意味していたが、現在では遺伝子間相互作用とほぼ同義で使われ、統計学的には２因子交互作用に対応する。このような遺伝子間相互作用を網羅的かつ高速に検出する方法の開発が期待されている。

２．研究の目的

(1) 世界的に用いられている一般的な方法として、まず単独の SNP で表現形と有意に関連するものを抽出し、次に抽出した SNP の中から交互作用を示すものを探索する方法がある。この方法では、下記の図１（上）の中の最も高いバーである相乗的効果を探索するために、その横の各 SNP の有意なリスクに基づき探索が行われる。しかし、図１（下）のような相乗的作用のみが存在する場合、各 SNP の効果が両方とも有意で無いため、この SNP の組み合わせは検出できない。探索すべき組み合わせを統計学的に述べれば、「主効果が無く、かつ、交互作用のみが存在する変数の組み合わせ」を膨大な数の中から高速にスクリーニングすることが本研究での最大の目的となる。

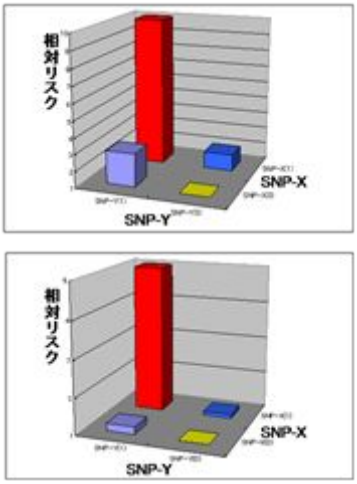

図１ 一般的な相乗的エピスタシス（上）
主効果のない相乗的エピスタシス（下）

(2) マイクロアレイによる遺伝子発現データ解析では、SNP 解析のために開発した遺伝子間相互作用の探索アルゴリズムを応用する。検定の多重性を考慮した上で 2 遺伝子間相互作用をすべての遺伝子の組み合わせから高速に検出する方法の開発を行う。さらに、遺伝子発現データとして連携研究者の三木教授が既に取得している乳がん組織のマイクロアレイデータを用いてサブタイプ解析へ応用する。

３．研究の方法

(1) SNP 解析において、既に代表者が共同研究において特許申請（特願 2009-070753）を行ったエピスタシス検出アルゴリズムに対し、検定の多重性を考慮した上で統計的な有意性の評価を理論的およびデータ解析的に行う。解析にはすでにがん研究会で取得した乳がん患者 60 例の SNP データ（約 90 万 SNP）と薬剤副作用の情報を用いた。

(2) マイクロアレイデータに基づく遺伝子間相互作用の高速な探索法の開発へと応用する。さらに、3 遺伝子の遺伝子間相互作用の解析法への拡張と、解析ソフトウェアの開発を行う。次世代シーケンサーからの結果にも対応可能にするために、マルチコア CPU に対応するための並列計算法などアルゴリズムの高速化に関しては全期間にわたって継続的に行ない、フリーソフトウェアの構築を行う。

４．研究成果

(1) 研究代表者らがこれまで開発したゲノムワイド SNP 情報に基づくエピスタシス効果の探索アルゴリズムを改良した。特に高速化を図るため、主効果が無い場合のデータ構造を明らかにし、主効果が存在する場合は解析対象から除外するアルゴリズムを組みこむ、という工夫を行った。2 遺伝子間の交互作用は複雑な相関構造をもつために FDR を理論的に評価することが難しいことから、提案した統計量に対して計算機シミュレーションによって FDR を評価する方法の開発を行った。その結果、以前開発した方法と比較して約 20 倍高速にエピスタシスを検出することが可能となった。また検出されたエピスタシスに対して検定の多重性を考慮した q 値を与えて評価することが可能となった。

(2) この方法を応用し、3 遺伝子の遺伝子間相互作用を検出するアルゴリズムの開発を行った。統計量を工夫することで高速化はされたものの、組み合わせが膨大であるため 2 遺伝子のときのように PC で実行できるような解析時間にはならなかった。並列化によってスパコンでの解析は可能であったが、プログラムの改良によって高速化の余地があるものと考えられる。

(3) マイクロアレイを用いた遺伝子発現解析についても SNP と同様に遺伝子間相互作用を検出するアルゴリズムを開発した。がん患者を 2 群に分け、ロジスティック回帰モデルを用いて交互作用項が有意となる組み合わせを検出することが可能となった。解析の結果得られた遺伝子の組み合わせについて GO (Gene Ontology)を調べたところ、同じ GO タームを含む遺伝子の組み合わせが相対的に多くなることが確認された。

(4) 応用として、乳がん患者のマイクロアレイデータを用いてサブタイプに分類し、分類と関連する遺伝子および遺伝子間相互作用を探索する方法の開発を行った。回帰分析と変数選択を同時に行う Lasso の手法をクラスタリングに応用した Lasso クラスタリングの方法を用いて乳がん 417 例のサブタイプ分類と遺伝子選択を行った。417 例をトレーニングデータとテストデータに分け、トレーニングデータで選ばれた遺伝子でテストデータの分類を行う、ということを繰り返してクラスタリングの精度を評価したところ、90%近い正答率を得ることができた。

(5) さらに、公開されている乳がんマイクロアレイデータを用いて、サブタイプの一つである basal-like 症例の分類を試みた。約 3600 例のマイクロアレイデータを分類して basal-like 症例を選び出し、選ばれた 750 例を用いて Lasso クラスタリングを行った。その結果、basal-like 乳がんは 5 つの群に分類された。得られた 5 つの群に対して臨床情報を調べたところ、DRFS (無遠隔再発生存期間) において予後が良好な群が存在することが明らかになった (図 2)。

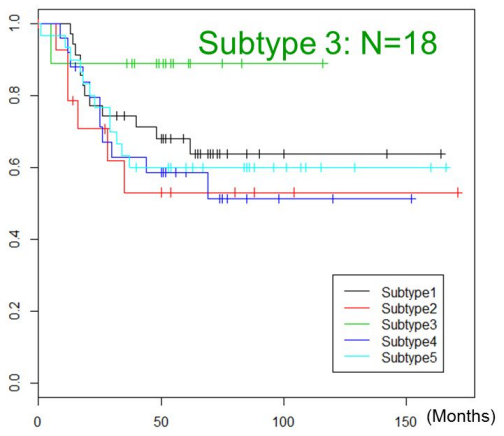

図 2 Lasso クラスタリングによって分類された 5 つの群に対する DRFS (無遠隔再発生存期間)

(6) 本研究の解析に使用したプログラムについて、ソフトウェアとして公開するための整備を行った。メインのプログラムは C++言語で記述しており、64 ビットの Linux 上でのテストは完了している。Windows でも動作するようプログラムの準備を進めており、テストが完了し公開の準備が整ったところで公開する予定である。理論的には 3 遺伝子の遺伝子間相互作用の解析も可能であるが、PC で実行するのは現実的でないため、公開するプログラムの機能には含まれていない。また、検定の多重性を考慮した有意性の指標である q 値の計算についても計算時間がかかることから公開プログラムとしては実装されていない。

５．主な発表論文等

〔雑誌論文〕（計１件）

① 松浦　正明、牛嶋　大、抗がん剤副作用予測システム、血液腫瘍、査読無、Vol. 70、2015、pp. 533-538

〔学会発表〕（計５件）

① Ushijima M, Miyaguchi K, Mori S, Miki Y, Matsuura M、Development of Lasso clustering method with application to basal-like breast cancer microarray data、10th AACR-JCA Joint Conference、2016 年 2 月 17 日、Maui, HW, USA.

② Ushijima M, Eguchi S, Komori O, Miki Y, Matsuura M、Lasso clustering method for classification of cancer subtypes using microarray data 、 27th International Biometric Conference、2014 年 7 月 7 日、Florence, Italy.

③ 牛嶋　大、三木　義男、松浦　正明、Development of clustering-based gene selection method with application to breast cancer subtype data、第 72 回日本癌学会学術総会、2013 年 10 月 5 日、パシフィコ横浜（横浜市）

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：

種類：
番号：
取得年月日：
国内外の別：

６．研究組織
(1)研究代表者
牛嶋　大（USHIJIMA, Masaru）
がん研究会・ゲノムセンター・研究員
研究者番号：　６０３２８５６５

(2)研究分担者
なし

(3)連携研究者
松浦　正明（MATSUURA, Masaaki）
帝京大学・大学院公衆衛生学研究科・教授
研究者番号：　４０１７３７９４

三木　義男（MIKI, Yoshio）
東京医科歯科大学・難治疾患研究所・教授
研究者番号：　１０２８１５９４

(4)研究協力者
旦　慎吾（DAN, Shingo）
がん研究会・がん化学療法センター・副部長
研究者番号：　７０３３２２０２