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　６ 年　 ５ 月 ２７ 日現在

機関番号：１３９０１
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７８５２

研究課題名（和文）大規模プロテオーム解析による俯瞰的がんシステム理解に向けた解析基盤の構築

研究課題名（英文）Establishing an Analytical Framework for a Comprehensive Understanding of Cancer Systems through Large-Scale Proteomic Analysis

研究代表者
宇野　光平（Uno, Kohei）

名古屋大学・医学系研究科（保健）・助教

研究者番号：５０８７３５８５

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：大規模プロテオーム解析では欠測値が生じるため欠測値補完が必要となるが、その欠測値補完をした後のデータには分散が過小に推定されるというバイアスが生じる。これはバイオマーカー探索の偽陽性確率を高めるため問題がある。そこで欠測値補完後の分散バイアスを補正する新たなアルゴリズムを開発した。そして、シミュレーションによってその有用性を示すとともに、実際のプロテオームデータに対して適用することで、提案手法の有用性を示した。本研究の成果をまとめると、以下のようになる。
[1] 欠測値補完手法の性能を評価する新たな指標を開発した。
[2] 分散バイアスを補正する新たなアルゴリズムを開発した。

研究成果の学術的意義や社会的意義
オミクス解析による疾患のバイオマーカー探索は重要な研究である。セントラルドグマの最終生産物であるプロテオームデータを用いたバイオマーカー探索も極めて重要であり、欠測値補完は慣習的に用いられている。しかし、その分散バイアスによる偽陽性については見過ごされている。そのため、本研究はそのバイアスに対しての警鐘であると同時にその対処方法を提示しており、バイオマーカー探索に貢献していると考えられる。そして、バイオマーカー探索は疾患研究において重要な役割を果たすことから社会的にも意義があると考えられる。

研究成果の概要（英文）：There are missing values in large-scale proteome analysis, thus missing value imputation should be necessary. However, missing imputed data have a bias that the variances are underestimated. This is a serious ploblem because it increases the rate of false positives in biomarker discovery. Therefore, we developed a new algorithm to correct the variance bias after missing value imputation. We demonstrated its usefulness through simulations and demonstrated the usefulness of the proposed method by applying it to proteome data. The results of this research can be summarized as follows:[1] We developed a new index to evaluate the performance of missing value imputation methods.[2] We developed a new algorithm to correct variance bias.

研究分野：統計科学

キーワード：プロテオーム解析　データ科学　多変量解析　欠測値補完

１．研究開始当初の背景

これまでバイオマーカー探索はがんゲノムを中心に進められてきたが、液相クロマトグラフィー質量分析法（LC-MS/MS）の進歩によって大規模なプロテオーム解析が可能となったことで、LC-MS/MS を利用したバイオマーカー候補の報告が年々増加している。これらの研究は主に疾患を持つ被験者群と対照群のタンパク質存在量を比較することによって探索する発現変動解析をもとにバイオマーカー候補を同定している。

しかし大規模なプロテオーム解析における発現変動解析には、欠測値問題というボトルネックが存在する。LC-MS/MS では測定限界や実験環境などを起因とした欠測値が多数発生する。統計解析手法や機械学習手法の多くは欠測値があると適用できないため、データの欠測値を補完することが必要になる。従来のプロテオーム解析においては、k 近傍法や特異値分解、ベイズ主成分分析やランダムフォレストなどによる欠測値補完手法を用いていた。

ところが、プロテオームデータは欠測率が非常に高く、欠測値補完によって得られた補完後データにバイアスが生じ、その後の発現変動解析の解析結果が歪められることがわかった。このバイアスによって、データの欠測率が高くなるほど補完後データの分散が低下するため発現変動解析における偽陽性の可能性が高まる。補完バイアスがその後の解析結果に悪影響を及ぼすことから、バイアスのない欠測値補完手法の開発が急務である。

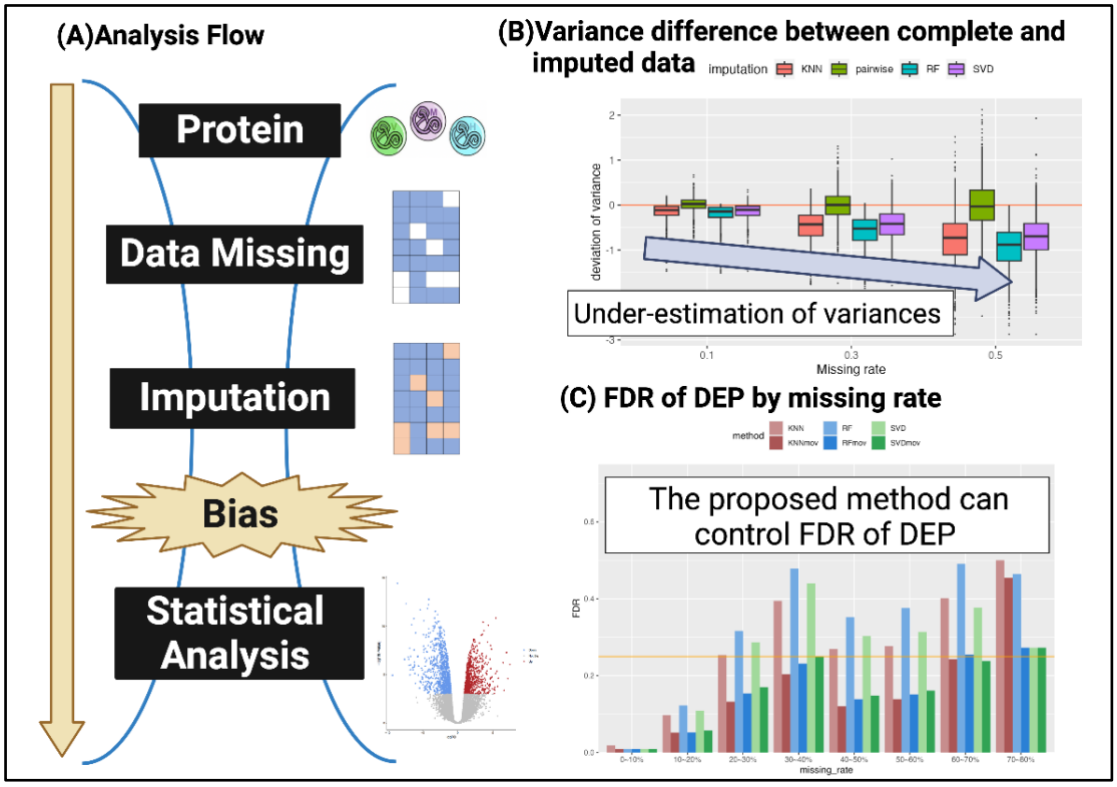

図　1　研究の概要図（A）プロテオーム解析における解析フロー：発現変動解析などの統計解析手法を適用する以前に欠測値補完が必要となるが、補完時にバイアスが生じるため解析結果の結果に歪みが生じる。（B）完全データの分散と補完後データの分散との差：既存の欠測値補完手法を適用して得られた補完後データ分散と完全データ分散の差を比較したところ、欠測率が高くなるほど分散が過小推定されることがわかった。（C）欠測値補完後の発現変動解析における誤検出率の比較：既存の欠測値補完手法および提案アルゴリズムによる補正後のデータを用いて発現変動解析を行った際の FDR を比較した結果、提案手法による補正後は FDR が抑えられることがわかった。

２．研究の目的

本研究の目的はバイアスのない欠測値補完手法を開発することで、プロテオーム解析における欠測値補完バイアスによるボトルネックを解消することである。特に発現変動解析に影響を与える恐れのある分散の過小推定バイアスがない手法を開発する。また、これまでバイアスが見過ごされてきた原因として、欠測値補完手法の性能を評価する指標がデータ点の再現にのみ着

目しており、平均や分散といった構造の再現を評価できていなかったことが挙げられる。そこで、構造の再現を評価する新たな指標を開発する。

３．研究の方法

シミュレーションによって、既存の欠測値補完手法には分散を過小推定し、発現変動解析の偽発見率が増加することを確認する。そして過小推定された分散を補正するアルゴリズムを開発し、シミュレーションおよび実際のプロテオームデータへの適用によって性能を検証する。

４．研究成果

完全データと補完データの分散の再現を評価するための指標である MDV を開発した。P個のタンパク質のうち、欠測値が 1 つでもあるタンパク質の集合を$\boldsymbol{M_p}$とすると、このとき MDV は

$$MDV = median(\boldsymbol{b}_{var}) = median([\dots, var(\boldsymbol{x}_j^{imp}) - var(\boldsymbol{x}_j^{comp}), \dots]) \quad (j \in M_p)$$

で定義される。ただし$\boldsymbol{b}_{var}$は$P_{M_p} \times 1$のベクトルで各要素は補完データの分散から完全データの分散を引いた値であり、P_{M_p}は集合$\boldsymbol{M_p}$の濃度とする。

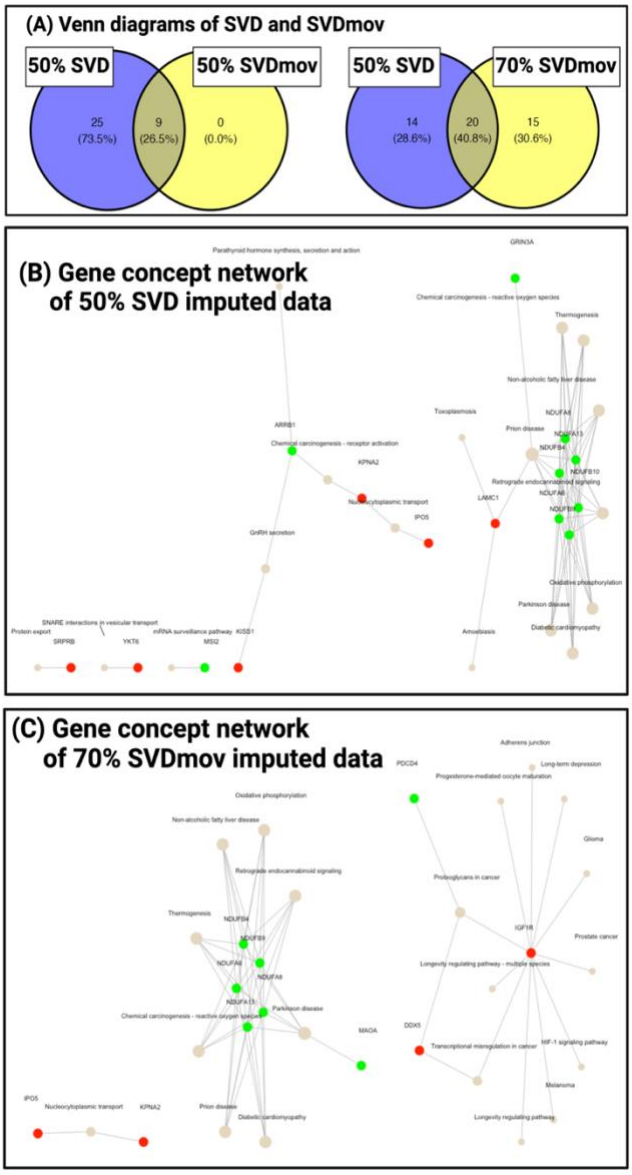

図 2 実データへの適用結果 (A)有意に変動していたタンパク質の数: 欠測率 50%以下と 70%以下のタンパク質を用いて発現変動解析を行い有意に変動していたタンパク質の数をベン図にした。(B)既存の補完手法を用いた場合の gene concept network。(C)提案手法による補正をした場合の gene concept network。

くわえて、補完データの分散バイアスを補正するアルゴリズム MOVAMI を開発した。MOVAMI は補完データの分散にバイアス補正項を加える。あるタンパク質jの分散を補正するとき、補正分散s_j^{mov}は、$s_j^{mov} = var(x_j^{imp}) + h$で推定される。ただし式中の$h$は非負であり、事前に定める必要がある。補完データの分散と完全データの分散を用いた MDV をhとすることが理想だが、実際には完全データを観測することはできない。そこで、欠測値補完による分散バイアスが生じないペアワイズ分散を完全データの分散の代用とすることで、hを決定する。ペアワイズ分散とはペアワイズ共分散行列の対角要素であり、各タンパク質の観測値のみを用いて分散を推定することを意味している。ペアワイズ分散をs_j^{pair}としたとき、以下の式でhを推定する。

$$\hat{h} = median(\dots, var(x_j^{imp}) - s_j^{pair}, \dots)$$

そして提案手法を、公開されているプロテオームデータにて、がん治療の予後良好群、予後不良群に分けて発現変動解析を行い有意に変動したタンパク質を同定した。図 2 のように補正手法では 70%の欠測率まで解析対象として採用し、その結果、従来は欠測率が高いため解析不可としていたタンパク質によるネットワークを推定することができた。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 宇野光平，松井佑介
２．発表標題 欠測値補完の分散バイアス補正：FDR抑制による未解析プロテオームデータの活用
３．学会等名 日本計算機統計学会第37回大会
４．発表年 2023年

〔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

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