

**厚生労働科学研究費補助金（政策科学総合研究事業）
（臨床研究等 ICT 基盤構築・人工知能実装研究事業）
分担研究報告書**

課題名 : 新薬創出を加速する症例データベースの構築・拡充/創薬ターゲット推定アルゴリズムの開発
研究分担者名 : 山西 芳裕
国立大学法人九州工業大学大学院 情報工学研究院 生命化学情報工学研究系 教授

研究要旨

本年度は、臨床情報と多階層オミックス情報を統合解析するための機械学習手法のプロトタイプを開発した。上記の手法を本事業で収集した臨床情報と多階層オミックス情報に適用し、IPF の創薬標的探索やバイオマーカー探索を行った。これまでに開発したツール化合物探索 AI を、特発性肺線維症（IPF）の創薬標的探索に応用し、IPF の創薬標的候補を予測して、最終候補を絞り込んだ。

A. 研究目的

医薬ビッグデータに基づいて、対象疾患の創薬標的候補の「確からしさ」の検証実験に利用可能なツール化合物を探索するインシリコ手法を開発する。様々な疾患に関するマルチオミックスデータや分子ネットワークデータ、既承認薬、開発中止化合物、合成化合物、天然化合物など大規模な化合物の構造データや実験データを収集する。疾患データと化合物データの融合解析を行う統計手法や、多様なオミックス関連データを有効活用して化合物を効率的にスクリーニングできる機械学習の手法を開発する。最終的に、特発性肺線維症に対して見出された創薬標的分子候補を制御するツール化合物の候補をインシリコ予測する。

B. 研究方法

- ① 臨床情報と多階層オミックス情報を統合解析するための機械学習手法を開発する。
- ② 上記の手法を本事業で収集した臨床情報と多階層オミックス情報に適用し、IPF の創薬標的探索を行う。
- ③ 前年度までに開発したツール化合物探索 AI を創薬標的探索に応用し、IPF の創薬標的候補を提示する。

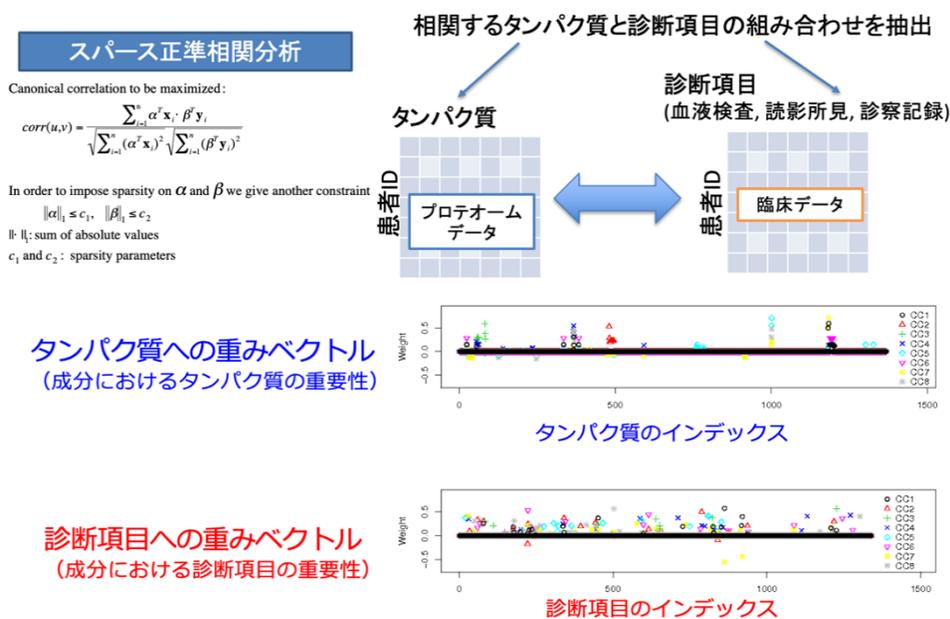
（倫理面への配慮）

本研究は、医薬基盤・健康・栄養研究所において倫理審査、承認を得た後、ヒトゲノム・遺伝子解析研究に関する倫理指針及び、人を対象とする医学系研究に関する倫理指針に従って遂行した。

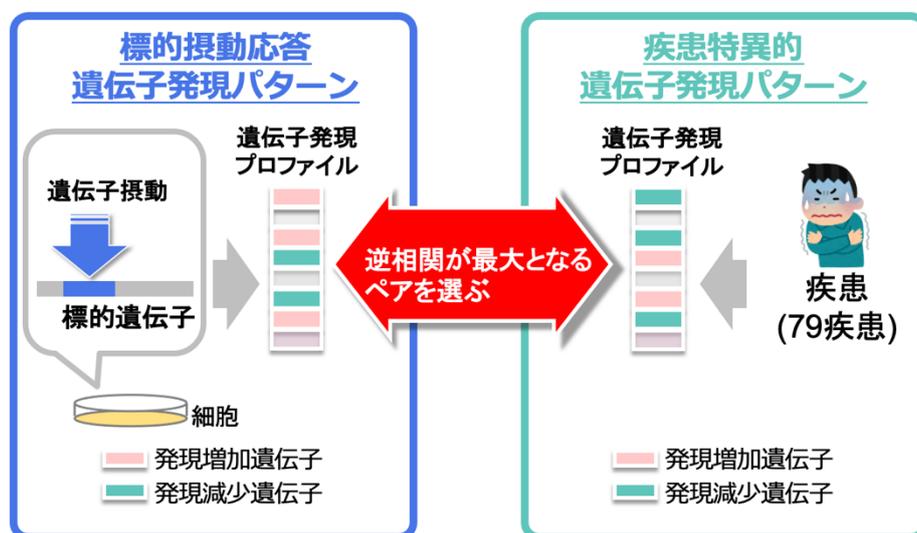
C. 研究結果**研究結果の概要：**

- ① 臨床情報と多階層オミックス情報を統合解析するための機械学習手法を開発：

臨床情報と多階層オミックス情報を統合解析するための機械学習手法の開発を開始した。基盤研の夏目先生のグループから阪大コホートのデータの一部をいただき、データ解析を開始した。生データには、欠損値（データが空白）の処理方法、コントロールデータ、性別情報、読影所見の項目名の記号などで不明な部分があったが、基盤研の研究者と逐次確認を行ない対応することができた。プロテオームデータと臨床データのデータ間の相関に寄与する特徴を抽出するため、スパース正準相関分析、種類の行列積におけるモジュール検出、スパース性を考慮した非負値行列分解などのアルゴリズムを実装して、解析を行なった。しかしながら、モジュール検出と非負値行列分解は、解釈しやすい結果が得られなかった。現在、スパース正準相関分析において同一成分で高い重みで抽出されたタンパク質と検査項目のセットを検証し、IPF との関連について生物学的な解釈を行なっている。さらに、基盤研の伊藤先生に各検査項目の IPF との関連度の情報を付加していただいた。IPF と関連度の強い検査項目に対して、正例と負例を分類する L1 正則化ロジスティック回帰モデルを適用し、IPF への分類に寄与するタンパク質を抽出した。様々な検査項目に対して抽出されたタンパク質の重みを統合し、最終的なタンパク質の IPF との関連度を評価して、IPF 特異的なタンパク質を予測した。現在、予測したタンパク質と IPF との関連について生物学的な解釈を行なっている。



ヒト細胞における創薬標的分子の摂動応答プロファイルと疾患特異的な遺伝子発現プロファイルを用いた創薬標的分子の予測手法の研究開発を行なった。ドラッグリポジショニングからターゲットリポジショニングへのパラダイムシフトを目指している。創薬標的分子をコードする遺伝子のノックダウン (KD) プロファイルが標的分子を阻害した薬の機能を反映し、同様に遺伝子の過剰発現 (OE) プロファイルが標的分子を活性化した薬の機能を反映していると仮定して、阻害標的分子と活性化標的分子を分けて予測した。疾患特異的な遺伝子発現パターンと、ヒト細胞における遺伝子摂動応答の遺伝子発現パターンとの間の逆相関を計算する方法は、精度はあまり高くなかった。部分的に既知の創薬標的分子の情報を学習し、疾患間の類似性を考慮して、創薬ターゲット分子の予測を行うマルチタスク学習アルゴリズムは、疾患の既知の阻害標的分子や活性化標的分子を上手く予測することができた。



② 上記の手法を本事業で収集した臨床情報と多階層オミックス情報に適用：

プロテオームデータと臨床データのデータ間の相関に寄与する特徴を抽出するため、マルチタスク学習を導入したスペース正準相関分析や、IPF と関連度の強い検査項目に対して患者を分類する L1 正則化ロジスティック回帰モデルを適用し、IPF への分類に寄与するタンパク質を抽出した。様々な検査項目に対して抽出されたタンパク質の重みを統合し、最終的なタンパク質の IPF との関連度を評価して、IPF 特異的なタンパク質を予測した。重みが大きいタンパク質は IPF と関連するタンパク質と予測される。今回予測されたタンパク質には、肺の線維化との関係性が示唆されているものが含まれていた。以上から、L1 正則化ロジスティック回帰モデルを用いることで、IPF に特異的なタンパク質を選択できることが示唆された。その他にも IPF との関連が報告されていないタンパク質が多数選択されており、これらのタンパク質には IPF の治療標的や

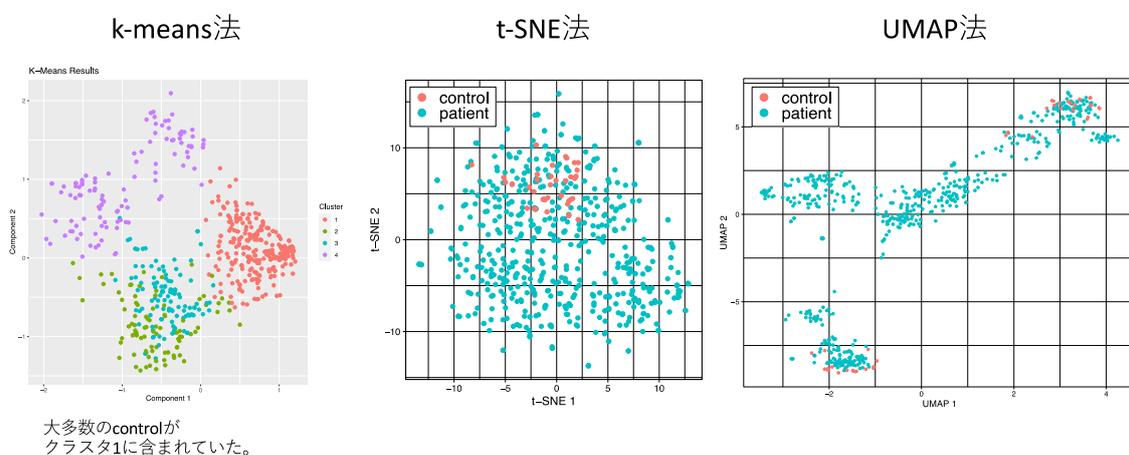
バイオマーカーとなるタンパク質が含まれる可能性がある。

次に、スパース正準相関解析を用いた、プロテオームデータと臨床データのデータ間の相関に寄与する特徴の抽出を引き続き行った。スパース正準相関解析には、マルチタスクスパース正準相関解析法を用いた。解析結果から算出した患者毎の正準スコアの分布を描き、コントロール群と患者群を異なる集団として分類されることを確認した。また、患者群において複数の集団が形成されていた。その分布を以下に示す。

IPF 患者では IPF に強く関連する読影所見の正例数の合計が大きくなると想定し、患者ごとの読影所見の正例合計数を正準スコアの分布に反映させた。その結果、強く関連する読影所見の正例数が多い患者が複数の集団を形成することが確認できた。

以上の結果を踏まえ、患者毎に正例である読影所見を調査し、IPF 特徴を持つ患者がグループとなっていないかを検討中である。また、提供いただいた臨床データには、ばち指や捻髪音といった IPF に関連する診療記録の情報が含まれていた。このような情報を読影所見のデータと組み合わせて解析することで、より精度良く IPF 患者を判定できると考え、再度スパース正準相関解析を行っている。さらにこれらの臨床データには、IPF 以外の特発性間質性肺炎患者の特徴である読影所見の項目も含まれている。臨床データを用いたクラスタリングにより IPF と他の特発性間質性肺炎患者を分類可能か検討中である。

次に、読影所見や診療記録のデータから、IPF の患者を分類可能であるか検討した。具体的には、k-means 法、UMAP 法及び t-SNE 法を用いたクラスタリングを行い、形成されたクラスターから IPF 患者を選択できるのか検討した。以下にクラスタリングの結果を示す。



3~5 つ程のクラスターが形成されることと、コントロール患者は個別のクラスターではなく、IPF の疑いのある患者も含まれるクラスターを形成することが確認できた。以上から、IPF の疑いのある患者にはコントロール患者と近い特徴を持つものが含まれることが予想された。上述したが、この結果はバイナリである特徴データのみを使用したものである。連続値であるプロテオームデータも IPF 患者の分類を行う上で重要な情報となりうる。そこで、プロテオームデータをバイナリデータに変換し、特徴データと結合した新たなバイナリデータを作成した。この新規バイナリデータを用いて k-means 法、UMAP 法及び t-SNE 法を用いてクラスタリングを行い、IPF 患者とそうでない患者の分類が可能であるか検討中である。

③ 前年度までに開発したツール化合物探索 AI を創薬標的探索に応用：

これまでに開発してきた臨床投薬データ、薬剤応答トランスクリプトームデータ、グラフ畳み込みニューラルネットワーク、ランダムフォレスト回帰モデルに基づくインシリコ手法による IPF の創薬標的予測結果を統合して、予測スコアが最も高かった生体分子を IPF の創薬標的候補として提示した。その実験検証を加藤先生のグループと共同研究で行ってきた。提案手法の妥当性を調べるため、IPF 以外の疾患でも同様の傾向があるか検証した。確実な創薬ターゲットが既知である疾患に適用し、既知の創薬ターゲットのタンパク質が正しく予測されるかどうか検証した。その結果、既知の創薬ターゲットのタンパク質は提案手法によって上位に予測され、フィッシャーの正確確率検定でも統計的に有意であり、提案手法の妥当性を確認できた。

D. 考察

① 臨床情報と多階層オミックス情報を統合解析するための機械学習手法を開発：

スパースモデリングにより、臨床データとオミックスデータから IPF 特異的なタンパク質を抽出できたので、現在、予測したタンパク質と IPF との関連について生物学的な解釈を行なっている。ターゲットリポジョ

ニングの情報技術を開発したが、疾患と生体分子の遺伝子発現パターンの逆相関に基づく方法は、精度はあまり高くなかった。マルチタスク学習アルゴリズムは、疾患の既知の阻害標的分子や活性化標的分子を区別して高精度で予測できたので、有用性が期待できる。

②上記の手法を本事業で収集した臨床情報と多階層オミックス情報に適用：

IPF の疑いのある患者に IPF かどうかのラベリングが現時点ではされていないが、患者が IPF かどうか判定される予定である。IPF 患者が判明すれば、その他の特発性間質性肺炎患者やコントロール患者と比較することで、IPF の新規バイオマーカーや治療標的となるタンパク質をより正確に判定できると考えられる。また、プロテオームと特徴のバイナリデータを用いてクラスタリングを行った結果から、IPF 患者のクラスターが形成されていたかどうか判定する予定である。

③前年度までに開発したツール化合物探索 AI を創薬標的探索に応用：

確実な創薬標的が分かっている検証可能な IPF 以外の疾患でも提案手法の妥当性が確認できたため、IPF に対する予測結果の妥当性が示唆される。

E. 結論

当初の研究計画の達成状況は概ね良好である。公共データの臨床投薬データとオミックス情報を融合解析して創薬ターゲットを予測するアルゴリズムのプロトタイプを開発できた。阪大病院の臨床データとプロテオームデータを統合解析し、バイオマーカー候補の分子を予測できた。期待するフェノタイプを持つツール化合物の結合タンパク質を探索することで IPF の創薬標的候補を見出した。

F. 研究発表

1. 論文発表

1) 該当なし

2. 学会発表

- 1) 難波里子、岩田通夫、山西芳裕, "From drug repositioning to target repositioning: omics-based prediction of therapeutic targets for a variety of diseases", 情報計算法学化学会 (CBI 学会) 2021 年大会, 口頭発表セッション, オンライン, 10/26-10/28 (発表日 10/27), 2021.
- 2) 岡本紗枝、澤田隆介, 山西芳裕, "Estimation of disease preventive drugs and therapeutic targets using clinical big data", 情報計算法学化学会 (CBI 学会) 2021 年大会, 口頭発表セッション, オンライン, 10/26-10/28 (発表日 10/27), 2021.
- 3) 難波里子、岩田通夫、山西芳裕, "ターゲットリポジショニング：遺伝子摂動応答トランスクリプトームを用いた創薬標的予測", ポスターセッション, 第 10 回生命医薬情報学連合大会, オンライン, 9/27-9/29, 2021.
- 4) 岡本紗枝、澤田隆介, 山西芳裕, "臨床ビッグデータからの疾患予防薬の探索と治療標的の推定", ポスターセッション, 第 10 回生命医薬情報学連合大会, オンライン, 9/27-9/29, 2021.

G. 知的財産権の出願・登録状況 (予定を含む。)

1. 特許取得

2021 年 10 月 27 日、特許出願を完了した。特願 2021-175674

【発明の名称】抗特発性肺線維症剤

【発明者】加藤明良・石崎 敏理・濡木真一・平山 文博・山西 芳裕・澤田 隆介

【特許出願人】国立大学法人 大分大学・国立大学法人九州工業大学

2. 実用新案登録

なし

3. その他

なし