提出日：2020 年５月１９日

2019 年度研究実施報告書

(対象期間 2020 年 1 月 1 日～2020 年 3 月 31 日)

担当課題　　：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

研究機関名：国立大学法人　京都大学
研究責任者：教授　　奥野　恭史

【１．実施内容】

実施項目１

１－１　研究目的

既存公開症例データベースを人工知能技術に適用可能な状態に整備し、そのデータを用いて主に深層学習技術の一つである Graph Convolutional Network (GCN) 及び説明可能人工知能技術の一つであるベイジアンネットワーク (BN) などを利用・拡張・改良することにより創薬ターゲット探索・推定が行えるアルゴリズムを開発する。

１－２　研究概要・要旨

既存公開がん症例データベースである The Cancer Genome Atlas (TCGA)、Genomcs of Drug Sensitivity in Cancer (GDSC)、及び Cancer Cell Line Encyclopedia (CCLE) に登録されている各種がん症例及びがん細胞株薬剤応答データを、人工知能・機械学習アルゴリズムに適応可能な状態な状態に整備する。

１－３　実施内容

まず既存公開がん症例データベースである The Cancer Genome Atlas (TCGA)、Genomcs of Drug Sensitivity in Cancer (GDSC)、及び Cancer Cell Line Encyclopedia (CCLE) に登録されている各種がん症例及びがん細胞株薬剤応答データを、人工知能・機械学習アルゴリズムに適応可能な状態な状態に整備する。

次に、そのデータを用い、主に GCN 及び BN を用いて創薬ターゲット因子の予測のためのアルゴリズムの改良を行う。GCN は大量のデータからグラフ（ネットワーク）で表現される関係性の予測に使用できるもので、すでに遺伝子ごとのゲノム変異情報、薬剤、及びタンパク間相互作用情報から構築したネットワークを用いた関係性予測で一定の予測能力を確認している。これを創薬ターゲットとなる因子の予測に改良を行う。BN は因子間の因果関係をグラフとして明に表す教師なしの機械学習アルゴリズムである。そのため疾患や薬剤の作用機序解明に有用である。データから推定されたグラフ構造から創薬ターゲットとなりうる因子の特徴を見出すことにより、創薬ターゲット予測へと応用可能であると考え、そのグラフ上の特徴量の創出を行う。

１－４　結果、成果等

効率的に薬剤反応性の予測やバイオマーカーの推定を行うためには、ゲノムやオミクスデータ、薬剤（化合物）、臨床情報の情報を組み合わせた超高次元のデータを扱う必要がある。そのため、適切に feature engineering、次元圧縮、特徴選択等の技術を用いて、本質的なデータを抽出することが重要である。本年度は、データを組み合わせる方法として、複数種類のデータを入力するマルチモーダルアプローチと複数の出力を同時に扱うマルチタスクアプローチの二つのプロトタイプ実装を行い、公共利用可能なデータ（CCLE/GDSC 等）を用いて評価することのできる環境を整えた。

上述データを用いた GCN を用いて創薬ターゲット因子の予測のためのアルゴリズムの改良として、本年度は、文献や既存の公共データベースにある網羅的な情報（文献空間の情報）を実験によって得られた実測空間の情報を統合した機械学習手法のプロトタイプ開発を行った。また、次年度以降にこれらを容易に利用できるようオープンソースの GCN フレームワークである kGCN（https://github.com/clinfo/kGCN）の基盤開発を行った。

BN を用いた創薬ターゲット因子の特徴量創出については、これまで肺がん培養細胞株を用いたトランスクリプトーム・ネットワーク解析により TGFβ投与によって引き起こされる上皮間葉移行(Epithelial-Mesenchymal Transition: EMT)の重要サブネットワークの抽出に成功している。この EMT サブネットワークを TCGA 肺がん患者に適応することにより、患者予後の層別化が可能なことまではわかっている（Tanaka et al., 2020)。この BN による EMT サブネットワークを用いて、今回新たに患者毎のサブネットワークの可視化を試み、それに成功した（図１)。前述の論文(Tanaka et al., 2020）では EMT サブネットワークより算出した患者サンプルの枝貢献量行列をクラスタリングすることにより患者の層別化を行ったが、この図はこのクラスタリング結果から抽出した各クラスタの代表的な４人の患者の EMT サブネットワークを新しく開発した方法で可視化したものである。この図で示した患者毎のネットワークの違いは TGFβ投与に反応するサブネ

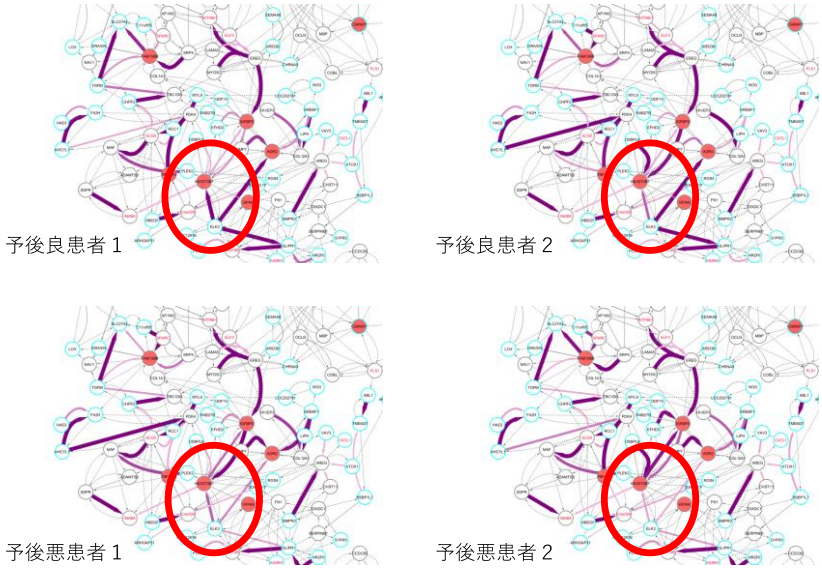

図１．患者別 EMT サブネットワークの部分拡大図。赤丸で示した周辺が特に変化が大きいことがわかる。

ットワークにおける各患者の違いを示しており、特定のハブ遺伝子周辺に変化が大きいことが観察される。これは推定したネットワーク構造の評価だけでは抽出不可能で、したがって、これらの特徴的な遺伝子が創薬ターゲットになりうると考えられる。またこれによって可視化されたサブネットワークの違いに基づき患者毎に異なる創薬ターゲットを考えることや薬剤感受性の違いの説明も可能になると考えられる。今後はこれらの特徴量の薬剤ターゲットとしての可能性を解析評価予定である。

【２．外部発表、論文投稿等】

Tanaka, Y., Tamada, Y., Ikeguchi, M., Yamashita, F., Okuno, Y., System-based differential gene network analysis for characterizing a sample-specific subnetwork, Biomolecules, 10(2), 306, 2020.

【３．知財化について】

奥野恭史、玉田嘉紀．特願 2020-002923「特徴ネットワーク抽出装置、コンピュータプログラム、特徴ネットワーク抽出方法及びベイジアンネットワーク分析方法」（出願日：2020/01/10）