別紙３

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に関する研究
研究分担者名：国立大学法人　京都大学　奥野恭史

研究要旨
創薬標的を探索する AI 開発
薬剤反応性の識別、バイオマーカー推定、分子メカニズムの解明と創薬標的探索を行う AI の開発
医薬品開発において、近年国内外を問わず創薬ターゲットの枯渇が問題となっている。現在残されているのは高難易度の創薬ターゲットのみであるがために、新薬の研究開発には多額の費用が必要となっており、これが高薬価、ひいては医療費の高騰の要因となっている。
更に、臨床試験段階で期待していた薬効が得られず開発が中断する例が増えていることも問題点として挙げられる。特に医薬品開発の 70～80%が Phase2 で中止となっており、この約 60%が、薬効が得られなかったことが原因との報告がある。つまり、「動物では効くが、ヒトでは効かなかった」という事案が多発している。これは現在の創薬研究開発スキームの限界であると考えられる。
このような現状を打開する解決策として、人工知能（AI; Artificial intelligence）が注目されている。AI のパフォーマンスと可能性に創薬・医療・ヘルスケア分野が大きな期待を寄せており、今後国際競争が激化することが必至である。
これらの現状を背景に、本事業では、「創薬ターゲットの枯渇問題」を克服すべく、動物からではなくヒトの情報から創薬ターゲット分子を探索する AI の開発実装を目的とする。つまり、臨床情報（=電子カルテを始めとする診療情報＋オミックスデータ）を収集・利用して創薬標的を探索する AI 手法の開発をおこなう。本事業では、対象疾患として難病指定の IPF（特発性肺線維症）を含む間質性肺炎および部位別がん死亡者数１位である肺がんを選択し、これらの臨床情報収集とそれを支援する基盤構築、異種かつ大量のデータを統合して創薬標的候補となる生体分子群を自動的に抽出する AI 手法の開発、創薬標的候補の実験的検証に有益となる基盤構築を包括的に遂行する。当該年度は、i)データ収集： IPF を含む間質性肺炎で 250 検体及び肺がんデータ収集の完結、ii)データ解析：これまでに開発した患者層別化 AI の改良および創薬標的探索に有用な AI の多角的開発、iii)結果解釈・仮説創出：データウェアハウス TargetMine への公共・市販データベース追加と基礎科学系文献からの自動知識抽出に必要な言語リソースおよび AI 手法構築の継続、iv）エクソソーム糖鎖データベース構築に向けた基盤技術の確立、v)肺がんを併発した IPF における一細胞解析を介した対象疾患の分子基盤分析を目標とする。

Ａ．研究目的

患者個別の疾患タイプ、薬剤反応性等について、識別・分類、分子メカニズムの解明、バイオマーカー推定と創薬標的分子探索を可能にする AI 技術を開発する。特に、生命科学や医療における様々な情報を創薬に応用する上で問題となる以下の課題の克服を目指す。

「克服すべき技術的な課題」
- 全ゲノム配列（超次元データ）、マルチオミクスのデータ処理
- 薬剤反応性などの患者層別化・個別化
- 多種多様情報（実験データ、臨床データ、文献データ等）の統合
- 分子-パスウェイ-細胞-臓器-動物-ヒトのマルチスケールモデル
- AI モデルへの時間的・空間的・定量的概念の導入

Ｂ．研究方法

① DNN や機械学習による薬剤反応性の識別とバイオマーカー推定

限られたデータを有効活用し、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。また、これらのデータを組み合わせて利用する方法として、オミクスデータ、化合物、臨床情報などの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薬剤反応性を同時に扱うマルチタスクアプローチが知られている。今年度では、前年度で開発した実装を元に、実際に公共で利用可能なデータを用いてモデル構築を行い、これら二つのアプローチを薬剤反応性識別やバイオマーカーの推定に利用する場合にどういった長所や短所があるかといったことを明らかにする。

本年度では、公開ベンチマークデータを用いて、上記の既存手法に関する調査および実装を行い、これらのアプローチの優位性について評価を行った。

② GCN によるネットワーク構造を考慮した薬剤反応性の識別、バイオマーカー推定、創薬ターゲット分子探索

限られたコストで効率的に実験を行い、得られた実測データを効果的に活用する方法の一つは、文献や既存の公共データベースの知識を利用することである。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の方法を開発することで、少ない実測データからの効率的な予測を実現する。これらの異なる性質を持つ情報を柔軟に表現する方法として、本研究課題ではグラフ表現（ネットワーク）を利用する。

今年度ではプロトタイプモデルを元に、より実装の強化を行い、翌年度以降の実データを用いた予測モデルの構築に向けたベースとなる手法開発を行った。

③ ベイジアンネットワークによる分子メカニズムの解明、バイオマーカー推定、創薬ターゲット分子探索

ベイジアンネットワークを用いた研究のうち、今年度は薬剤感受性に基づく患者層別化が可能な分子メカニズム解明・解析手法の確立を目指す（全体計画 3-1）。前年度においてトランスクリプトーム・データを用いたサンプル単位でのベイジアンネットワーク推定法は確立している。これを応用し、新規に取得するオミクス・データと薬剤感受性試験データを用いて、患者ごとの薬剤への反応の違いを、ベイジアンネットワークを用いてモデル化する。得られたネットワークモデルを用いてゲノムデータ及び薬剤感受性との相関を解析する。以上により、各薬剤において感受性の違いを説明するトランスクリプトーム・ネットワークを抽出する方法を確立し、薬剤感受性に違いが発生するメカニズムの、トランスクリプトーム・ネットワークレベルでの解明及び薬剤感受性に基づく患者の層別化を目指す。

（倫理面への配慮）

本研究は、医薬基盤・健康・栄養研究所ならびに本学において倫理審査、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

① DNN や機械学習による薬剤反応性の識別とバイオマーカー推定

Omics ベースのアプローチ(Ding, Michael Q., et al. ″Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics.″ *Molecular Cancer Research* 16.2 (2018): 269-278.)と薬剤ベースのアプローチ (Chang, Yoosup, et al. ″Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature.″ *Scientific reports* 8.1 (2018): 1-11.) の二つのアプローチに関する二つの既存手法の評価を行った。

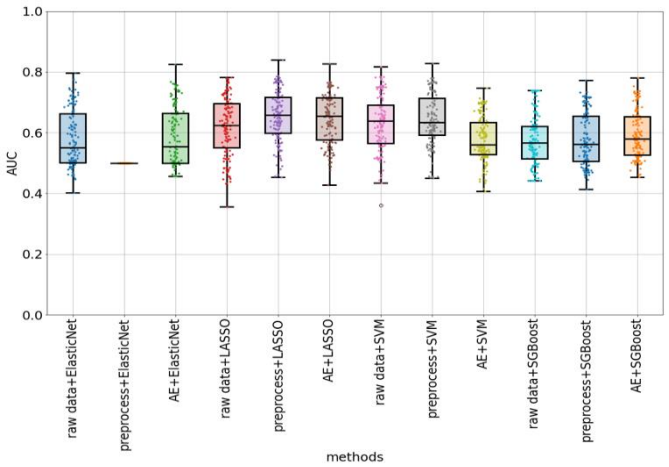

図 1：Omics ベースのアプローチ：従来法における薬剤活性予測の正答率比較

Omics ベースのアプローチでは、手法が有効な薬剤とそうでない薬剤が存在していることがわかり、薬剤によって手法を使い分ける必要性などが示唆された（図 1)。また、薬剤ベースのアプローチでは、従来モデルに加えいくつかの新規モデルを用いて評価を行った（図 2)。データとしては、現状公開データＧＤＳＣ・ＣＣＬＥを用いたものである。評価結果としてはデータのバランスを調整するなどのデータに合わせた処理がいくつか効果的であることが判明した。

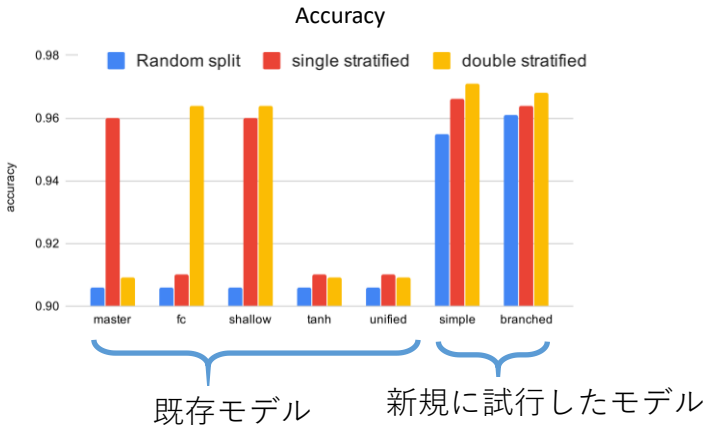

図 2：薬剤ベースのアプローチ：ゲノム変異と薬剤(抗がん剤)の情報を用いたマルチモーダル薬剤応答予測モデルの比較

② GCN によるネットワーク構造を考慮した薬剤反応性の識別、バイオマーカー推定、創薬ターゲット分子探索

遺伝子・薬剤・疾患の文献等から構築したグラフから新たな関係を予測する手法と予測結果を説明する部分を可視化する手法を開発し、GCN を用いた手法が他の手法よりも精度よく予測を行うことができることを示した（図 3-1)。また、この時の予測結果を説明する部分を可視化する手法を開発した（図 3-2)。

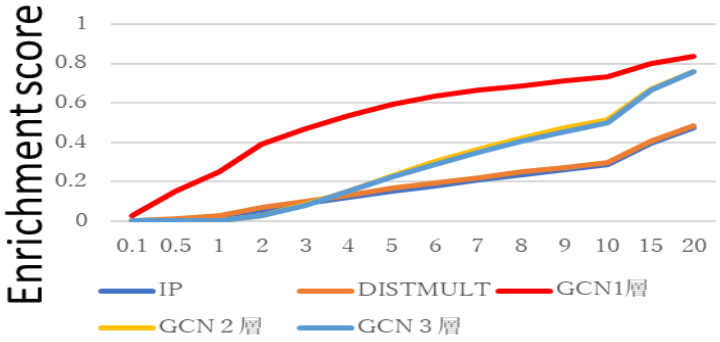

図 3-1：遺伝子・薬剤・疾患グラフの補間問題における手法間比較

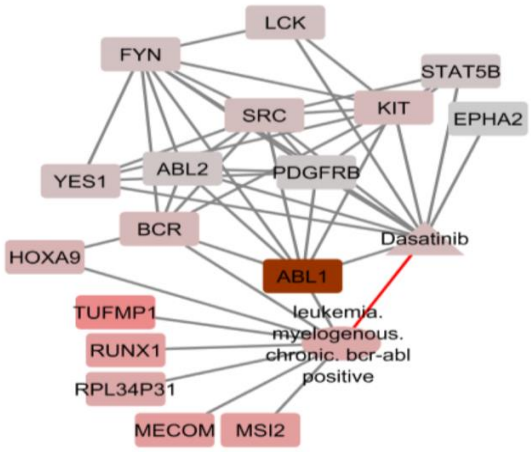

図 3-2：遺伝子・薬剤・疾患グラフの予測結果の可視化

また、このグラフ補間問題をさらに、予測問題へと適用する手法開発を進めており、文献情報と実験情報を合わせたモデルの開発が完了しており、詳細な評価および次年度に向けた実データへの適用可能性について検討を進めている。

③　ベイジアンネットワークによる分子メカニズムの解明、バイオマーカー推定、創薬ターゲット分子探索

（１）がん患者層別化手法の確立

昨年度開発した遺伝子ネットワークからの重要サブネットワーク抽出技術を改良し、分子メカニズム解明に有効ながんサブタイプ分類法を開発した。TCGA 胃がんデータセット（The Cancer Genome Atlas Research Network, *Nature*, 2014）では、そのオミクスデータによる４つの新規サブタイプが提案されているが、患者の予後との関連は見出されていなかった。ベイジアンネットワークを用いた重要サブネットワーク抽出技術では、患者サンプル毎に、ネットワークの枝（エッジ）毎の重要度を定量化することが可能であるが、この定量化方法によるデータ行列（＝患者毎の遺伝子ネットワークのプロファイル）をクラスタリングすることにより、生存時間に統計的な差がある３つのサブタイプを同定することができた（図４）。同定されたサブタイプ毎に重要なサブネットワークは異なっていることもわかり、これらの患者グループ毎に分子メカニズムが異なることが示唆された。本研究は BioRxiv に投稿済みで（Nakazawa et al. *BioRxiv*, 2021.03.24.436731）、学術雑誌への投稿も目前である。

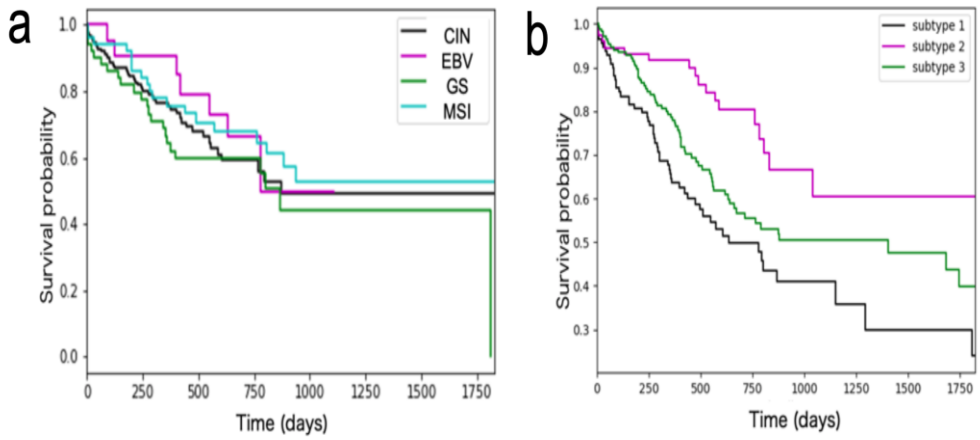

図４．(a) 既存４サブタイプの生存時間解析。(b) 新規開発手法により同定した３サブタイプの生存時間解析（任意の２グループ間の p-value ＜ 0.05）。

（２）遺伝子ネットワーク・サブネットワーク抽出法の検証

次に、開発した手法のさらなる検証を目的として、最近公開された SARS-CoV-2/COVID-19 データを用いて遺伝子ネットワーク解析を行い、創薬ターゲット探索への適用可能性を検証した。結果は論文にまとめ arXiv に登録後、すでに学術雑誌に投稿済みである（Tanaka et al., 2020, arXiv:2008.09261）。抽出した遺伝子

ネットワークには既存薬ターゲット遺伝子が多く含まれていることが確認でき、既存知識も矛盾のない結果が得られた（図５）。このサブネットワークに含まれる遺伝子は基本的に全て新規の創薬ターゲットになりうると言えるが、特にモジュールとして得られた境界の遺伝子や、子供遺伝子の多いハブ遺伝子、既存標的遺伝子の近傍の遺伝子は特に有力な新規候補遺伝子と言える。

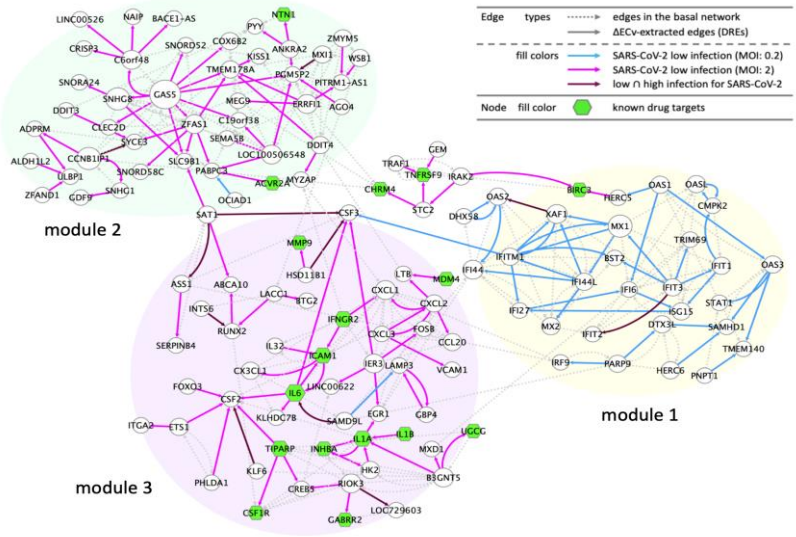

図５．SARS-CoV-2 データの遺伝子ネットワーク解析結果。ウィルス量の違いによるサンプル間の比較によりサブネットワークを抽出した。３つのモジュールが観測され、これらの境界に存在する遺伝子が、新規創薬ターゲットとしては有力である。緑色のノードが既存薬の標的遺伝子である。

（３）薬剤感受性に基づくネットワーク抽出技術の開発

１０患者からの複数薬剤毎の体積変化（腫瘍増殖率）とトランスクリプトームデータ３２サンプルを入手することができたので、その解析を行なった。目的は遺伝子ネットワーク解析により薬剤感受性を予測・説明できる遺伝子間の関係性（＝遺伝子ネットワーク中のエッジ）の抽出が可能かどうか検証することである。これまで開発した部分ネットワーク抽出法は、２群間、あるいは２サンプル間での推定されたシステム上での振る舞いの違いにより、部分ネットワークを抽出する方法だったが、今回新たな方法として、薬剤ごとの腫瘍増殖率と強く相関するエッジを抽出する方法を考案し、試行した。この方法は、コントロールサンプルなどの比較対象や薬剤感受性・耐性のラベル付けが不要であるため、応用範囲が広いと思われる。偽相関により単一の非連結エッジが多数抽出される可能性も考えていたが、薬剤に関連する連結部分ネットワークがモジュールとして抽出可能であることが示された（図６）。次に抽出した遺伝子ネットワークモジュールによって、腫瘍増殖率が予測できるかどうかの検証を行なった。予測がある程度可能であれば、抽出した遺伝子ネットワークモジュールに薬剤応答性を説明できる情報が十分あることが言える。図７は抽出した連結成分ごとに、leave-one-out 交差検証法による予測能力の結果である。腫瘍増殖率のデータをクラスタリングすることにより、サンプルを薬剤応答群と非応答群に二分し、予測問題の正解ラベルとした。上段は連結成分の枝の ECv を特徴量として用いた場合の予測率、下段は連結成分を構成する遺伝子の発現量を特徴量として用いた場合の予測率を示している。サンプル数が１０と非常に少ないが、連結成分によっては十分な予測精度を示しており、抽出したサブネットワークのサンプルごとに使われ方の違いにより薬剤の効果を予測できるという結果になった。以上より、抽出したサブネットワークが薬剤反応機序を説明できている可能性が確かめられた。今後はより大規模なデータセットでの検証を引き続き進める予定である。

以上のほか、理化学研究所のスーパーコンピュータ「富岳」を一般利用開始前に利用できる早期利用課題に採択されたため、遺伝子ネットワーク推定ソフトウェアの「富岳」への対応を行い、大規模ベイジアンネットワーク推定アルゴリズムである NNSR アルゴリズムおよびサンプル毎の枝貢献量計算の富岳対応を完了している。

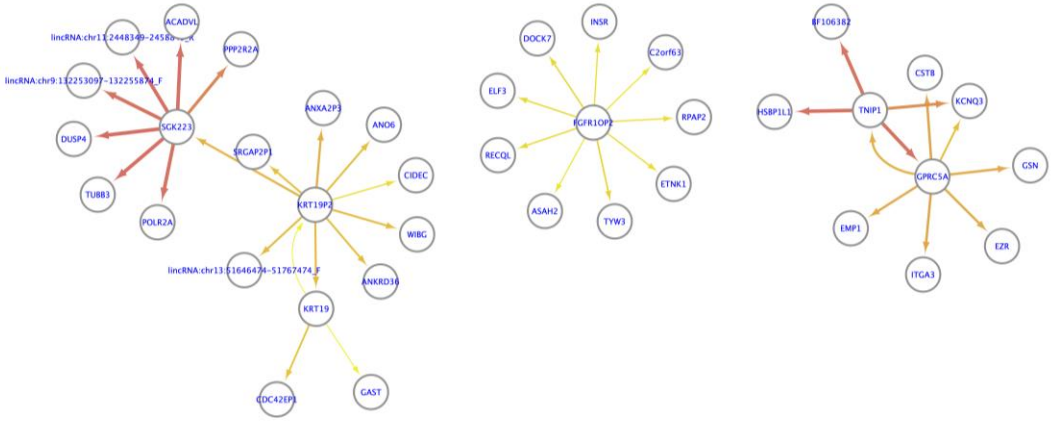

図６　セツキシマブ投与に対する腫瘍体積変化率と相関する遺伝子ネットワークエッジ抽出結果例（preliminary results)。エッジ単体ではなくハブ的な遺伝子を中心とした部分ネットワークがモジュールとして抽出できていることがわかる。

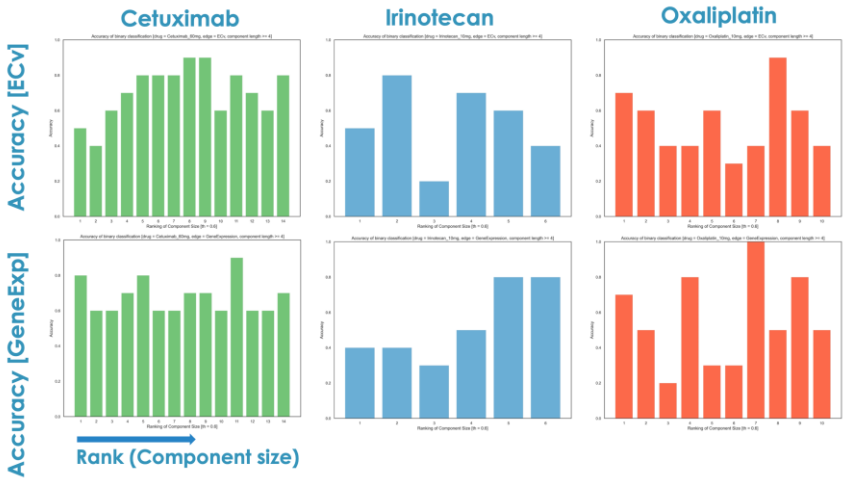

図７．抽出したサブネットワークの連結成分ごとの腫瘍増殖率予測結果。上段は枝貢献量による予測。下段は連結成分に含まれる遺伝子遺伝子発現量による予測結果。サブネットワークにより８～９割程度の予測能力があることがわかった。

D．考察

①　DNN や機械学習による薬剤反応性の識別とバイオマーカー推定

Omics ベースのアプローチおよび薬剤ベースのアプローチの結果より、Omics ベースのアプローチでは、手法が有効な薬剤とそうでない薬剤が存在していることがわかり、薬剤による影響を強く受けてしまうことが分かった。また、薬剤ベースのアプローチでは、データの偏りや前処理にモデルの性能が影響を受けるため、これらを考慮した手法の性能が高いことが分かったため、より精度の高いモデルを構築するためには、データの前処理等を含めたシステム全体の設計を検討する必要があることが分かった。

②　GCN によるネットワーク構造を考慮した薬剤反応性の識別、バイオマーカー推定、創薬ターゲット分子探索

ネットワーク予測に関しては、遺伝子・薬剤・疾患グラフの補間問題におけるベンチマークでの本アプローチの有効性を示した。今後、ベンチマークデータではなくよりリアルなデータを用いての評価を行い、有用性を評価する必要がある。

③ ベイジアンネットワークによる分子メカニズムの解明、バイオマーカー推定、創薬ターゲット分子探索

今年度は、昨年度開発したサンプルごとに遺伝子ネットワークを推定可能な方法を応用し、患者サブタイプ層別化、薬剤耐性層別化、薬剤標的遺伝子探索への応用の方法論の開発と検証を行なった。開発した方法論により、これまでできなかった遺伝子ネットワークを用いたサブタイプや薬剤耐性などが可能であることが確認でき、また抽出したサブネットワークは既存薬剤標的遺伝子を多く含むことがわかった。今後は計算機による結果を実データで検証する必要があり、それを通して手法の更なるブラッシュアップを図る必要がある。

Ｅ．結論

① DNN や機械学習による薬剤反応性の識別とバイオマーカー推定

薬剤ベースのアプローチをベースにしつつ、適宜 Omics 情報を入れていくアプローチが有用なのではないかと推察され、今後、この薬剤ベースのモデルに②で検討中の文献情報やより精緻な特徴量を用いてさらに精度の高いモデルを目指す。

② GCN によるネットワーク構造を考慮した薬剤反応性の識別、バイオマーカー推定、創薬ターゲット分子探索

ネットワーク構造を考慮する手法である GCN の有効性を確かめることができたため、これを用いて、①とも合わせて、薬剤予測問題へと適用する手法開発を進めて、文献情報と実験情報を合わせたモデルの構築を行う。さらに今後、この手法を強化し、バイオマーカー推定、創薬ターゲット分子探索へも利用していく方針である。また、実データへの適用に向けて、ツールとしてのさらなる整備を進めている。

③ ベイジアンネットワークによる分子メカニズムの解明、バイオマーカー推定、創薬ターゲット分子探索

トランスクリプトームデータを用いて、メカニズム解明、バイオマーカー推定、創薬ターゲット探索がこれまで以上に可能になることがわかった。今後は、プロテオームデータやゲノムデータなどの統合解析が鍵であり、計算による仮説の実験による検証を行いつつ、データ統合のための方法論の開発を進めたい。

Ｇ．研究発表

１．論文発表

1) Matsumoto S, Ishida S, Araki M, Kato T, Terayama K, Okuno Y ″ Extraction of protein dynamics information from cryo-EM maps using deep learning ″ Nat Mach Intell 3, 153-160 (2021).
2) Ryosuke Shibukawa; Shoichi Ishida; Kazuki Yoshizoe; Kunihiro Wasa; Kiyosei Takasu; Yasushi Okuno; Kei Terayama; Koji Tsuda,″CompRet: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration,″J Cheminform 12, 52 (2020)
3) Masuda, N., Murakami, K., Kita, Y., Hamada, A., Kamada, M., Teramoto, Y., Matsumoto, K., Sano, T., Saito, R., Okuno, Y., Ogawa, O., Kobayashi,T. Trp53 mutation in Krt5-expressing basal cells facilitates the development of basal squamous-like invasive bladder cancer in the chemical carcinogenesis of mouse bladder, Am. J. Pathol., 190(8), 1752-1762 (2020)
4) Saito, Y., Koya, J., Araki, M., Kogure, Y., Shingaki, S., Tabata, M., McClure, M., Yoshifuji, K., Matsumoto, S., Isaka, Y., Tanaka, H., Kanai, T., Miyano, S., Shiraishi, Y., Okuno, Y., Kataoka, K. Landscape and function of multiple mutations within individual oncogenes, Nature, 582, 95-99 (2020)
5) Ono, F., Chiba, S., Isaka, Y., Matsumoto, S., Ma,B., Katayama, R., Araki, M., Okuno, Y. Improvement in predicting drug sensitivity changes associated with protein mutations using a molecular dynamics based alchemical mutation method, Sci. Rep., 2020, 10(1):2161.
6) Tanaka, Y., Tamada, Y., Ikeguchi, M., Yamashita, F., Okuno, Y. System-based differential gene network analysis for characterizing a sample-specific subnetwork, Biomolecules, 2020, 10(2):306.
7) Hatae, R., Chamoto, K., Young, Hak, Kim., Sonomura, K.,Taneishi,K., Kawaguchi, S., Yoshida, H., Ozasa, H., Sakamori, Y. ,Akrami, M., Fagarasan, S., Masuda, I., Okuno, Y., Matsuda, F., Hirai, T. ,Honjo, T. Combination of host immune metabolic biomarkers for the PD-1 blockade cancer immunotherapy, JCI Insight, 2020, 5(2), 133501.
8) Iwata, H., Kojima, R., Okuno, Y. An in Silico Approach for Integrating Phenotypic and Target-

Based Approaches in Drug Discovery, Mol. Inform., 2020, 39(1-2):1900096.

２．学会発表

1) 日本学術会議　未来からの問い―日本学術会議100年を構想する（2020発行）
第4章　医療の未来社会、4-3　医療におけるビッグデータ・AI、（２）ビッグデータ・AIが拓く医療・創薬の未来
2) 「スーパーコンピュータ「富岳」によるCOVID-19治療薬探索」 ″Drug Screening for COVID-19 using Supercomputer “Fugaku” 第94回日本薬理学会年会　特別講演　AI創薬の現状と未来 Today and future of AI-based drug development シンポジウム（2021.03.08）
3) 「AI・シミュレーションによる新型コロナウイルス治療法開発への挑戦」　特別講演AIネットワーク社会フォーラム ～AI-Ready社会の実現に向けて～（2021.03.01）
4) 「データ駆動型　医療・医薬品・ヘルスケア産業」IQVIAセミナー（2021.02.15）
5) 「スーパーコンピュータ「富岳」・AIによる新型コロナウイルス治療法開発への挑戦」 第38回コロイド界面技術シンポジウム 「みんなを元気にするすごい技術　アフターコロナの研究開発　～動向/指針/変化する研究」（2021.02.04）
6) 「AIが拓く創薬イノベーション」第16回がんトランスレーショナルリサーチ（TR）ワークショップ AIが創る医薬品開発のカッティング・エッジ ）（2021.01.19）
7) 「AI・シミュレーションが拓く創薬・医療の未来」けいはんな R&D イノベーションフォーラム（2020.12.15.）
8) 「「富岳」がもたらす創薬・医療へのインパクト」第13回 スーパーコンピューティング技術 産業応用シンポジウム （2020.12.10.）
9) 「スーパーコンピュータ・AIが拓く創薬の未来」Life Science Startup Ecosystem （2020.12.10.）
10) 「スーパーコンピュータ・AIが拓く創薬の未来」第3回産学官連携情報交流セミナー（2020.12.03.）
11) 20年度精神・神経疾患研究開発費班会議（2020.12.01.）
12) 「AI・シミュレーションが拓く創薬・医療の未来」CHUGAI DIGITAL DAY～ヘルスケア×デジタルの2030未来予想図（2020.11.27）
13) 「スーパーコンピュータ「富岳」・AIによる新型コロナウイルス治療法開発への挑戦」 総務省　情報通信政策研究所主催「ＡＩネットワーク社会推進会議」（2020.11.12.）
14) 創薬におけるAIの現状と未来　奥野恭史 第21回日本毒性病理学会・教育セミナー 2020.11.7
15) スパコン「富岳」・AIによる新型コロナウイルス治療法開発への挑戦　奥野恭史　数理AIデータサイエンス市民講座 （2020.10.23）
16) スーパーコンピュータ・AI が拓くがん分子標的治療戦略　奥野恭史　第24回日本がん分子標的治療学会学術集会 （2020.10.8）
17) スーパーコンピュータ「富岳」vs 新型コロナウイル　奥野恭史　STS フォーラム公開シンポジウム（2020.10.3）
18) スーパーコンピュータ・AI が拓く創薬・医療の未来　奥野恭史　大阪府工業協会・役員会講演（2020.9.28）
19) AI・シミュレーションによる薬効・毒性予測　奥野恭史　第47回日本毒性学会学術年会シンポジウム（2020.6.29）

Ｆ．知的財産権の出願・登録状況（予定を含む。）

１．特許取得
なし

２．実用新案登録
なし

３．その他
1) Life Intelligence Consortium (LINC) 代表