戦略的創造研究推進事業　AIP 加速研究
（AIP 加速 PRISM 研究）
研究課題「人工知能技術を活用した
革新的ながん創薬システムの開発」

研究終了報告書

研究期間　2018年8月～2021年3月

研究代表者：浜本　隆二
（国立がん研究センター研究所、分野長）

§1 研究実施の概要

（1）実施概要

次世代シークエンサーや遺伝子編集技術に象徴されるバイオテクノロジーの加速的な進化の一方で、医薬品開発を巡る近年の研究開発の生産性は急激に低下し、「イノベーションの欠乏」が叫ばれて久しい。10 年前後の期間にわたって開発される新薬にかかる費用はますます増大し、一つの薬を上市するために必要な開発コストは 1000 億円以上に上る。また、候補化合物から医薬品として世に出されるものは 2 万から 3 万に一つであり、その成功確率は極めて低い。特に、新薬候補物質の効果を始めて患者で示す”Phase II 試験“での開発中止（Phase II attrition）が顕著であり、開発した薬剤の多くがマウスでは効くものの、人間ではその有効性が証明されないという、マウス実験を基盤とする現状の医薬品開発パイプラインの構造的な課題がある。本研究課題申請者らが対象疾患とする肺がんにおいても、患者の遺伝子を調べることで最適な治療薬を選択する「がんゲノム医療」が本邦でも開始されたものの、遺伝子の異常が検出されながらも適合する薬剤が存在しないケースが多数出現することが判明しており、社会問題となっている。以上の事から、新薬開発プロセスにおける構造的な課題を克服し、疾患治療で現状打破を得るためのイノベーションが現在期待されている。そのための鍵となるものが、ヒトの薬剤投与時や疾患罹患時における網羅的分子プロファイルに現れる生体変化の統合的理解と、実際の医療から得られるリアルワールド・データの解析に基づいた薬剤のヒトにおける有効性や毒性予測、更にこうした解析を可能にする最新の計算機アルゴリズムであると考えられる。

近年、「50 年来の技術的ブレークスルー」とも言われる深層学習と最新の機械学習手法の組み合わせが、生命情報処理に対するアルゴリズム革命をもたらすものと期待されている。特に、生物の脳神経系にヒントを得た情報処理メカニズムである人工ニューラルネットワークを多層化した深層学習は、社会や産業の形を変え得る画期的な情報技術として大きな注目を集めている。深層学習を既存の機械学習手法と比して特徴付けるものとしては、多種のデータを入力として取り扱えるマルチモーダル学習、複数の異なるタスクをモデルの一部として共有できるマルチタスク学習、少数の教師ありデータから汎化性能を得ることのできる半教師あり学習や教師無し学習、階層的な特徴を自動的に獲得できる表現学習などが挙げられる。近年のコンピュータ処理能力の飛躍的向上と加速的に蓄積されるビッグデータを基盤として、深層学習は画像や音声といった比較的高度な知的タスクで、既に人間を上回る性能を示すに至り、生命情報処理ならびに創薬の分野においてもその応用に関する世界的な研究開発競争が激化している。そこで、本邦における分子標的治療薬の創生及び生命情報処理に関わる第一線の研究者が一つのチームとして協力し合うことで、がんの生体時空間にわたるシステム的統合理解に基づいた創薬候補因子の探索と、がん治療におけるあるべき不均一性を包含した、臨床のリアルワールド・データに基づいたヒトにおける有効性や毒性の予測を可能にすると捉え、医薬品開発における「イノベーションの欠乏」を解決するための新しい技術開発に取り組んだ。その結果 AI を活用した複数の解析プラットフォームの開発に成功した。また AI 解析において最重要な事項の一つとして、質の高いデータベースの構築が挙げられるが、我々は臨床データを効率的に収集するプラットフォームの構築に成功した。その結果 AI 解析を志向した世界最大規模の肺がん統合データベースを、本研究課題で構築することに成功した。本データベースは質の高いデータが保存されている事から、本邦の財産と考えられる貴重な研究成果であると判断している。

(2)顕著な成果
＜優れた基礎研究としての成果＞
1.
概要:AI の学術研究を行う上で最重要事項の一つとして、質の高いデータベースを構築することが挙げられる。我々は本研究課題で肺がんのオミックスデータベース構築に取り組み、全エクソン解析及び診療情報においては、1500 例以上という米国の TCGA(The Cancer Genome Atlas)を超える、世界最大規模の肺がん統合データベースの構築に成功した。本データベースは本邦の財産ともいえる、貴重な研究成果であると判断している。

2.
概要: AI 研究により優れた研究成果を出していく為には、質の高い構造化データを準備する事は非常に重要である。しかし、多くの医療データは不均質で構造化されておらず、価値創造の機会を逸しているのが現状である。我々は AI 技術と最先端の ICT 技術を組み合わせる事で、効率的にデータを構造化するプラットフォームの構築に成功した。本プラットフォームはメディカル AI 研究を発展させる上で、重要なイノベーションであると判断している。

3.
概要:がんは複雑な疾患である為、その本態解明を行い創薬に発展させる為には、臨床検体から得られた膨大なオミックスデータを、詳細な診療情報と共に効率的に解析する技術を開発することが必須であると考えている。そこで、我々は機械学習・深層学習技術を活用して、がんに関する医療ビッグデータを解析する手法の開発に取り組んだ。その結果肺がんに関する、3 種類の新規プラットフォームの開発に成功し、2 つの手法は既に国際学術誌に掲載されており、残りの1つは現在論文投稿準備中である。

＜科学技術イノベーションに大きく寄与する成果＞
1.
概要:肺がん統合データベース構築においては、ゲノム解析のみならず、エピゲノム解析も施行し、特にヒストン修飾解析に関しては、FFPE(ホルマリン固定パラフィン包埋)サンプルからも効率的かつ精度の高い ChIP-seq 解析を行う事が可能な、新しい ChIP-seq 解析手法(RCRA ChIP-seq 法)を開発した。FFPE サンプルを用いた ChIP-seq 解析は世界中で高い需要があり、本手法は科学技術イノベーションに大きく貢献する成果であると判断している。また、世界最大規模の肺がん統合データベースも、本邦にとって貴重な財産で、科学技術イノベーションに大きく寄与する可能性があると判断している。

2.
概要:上述のようにメディカル AI 研究を行う上で、質の高い構造化された医療データを蓄積していくプラットフォームを構築することは重要である。我々は AI を用いて、自動的に放射線画像にアノテーション付けを行うプラットフォームを開発し特許を取得後、既に大手医療機器メーカーに導出している。我々が開発した技術が社会実装される事により、AI 技術を活用した放射線画像解析に関して、研究及び臨床双方に大きく貢献すると考えられ、科学技術イノベーションに大きく寄与する成果であると判断している。

3.
概要:我々はがんに関する大規模オミックスデータを、効率的かつ高精度に解析する AI 技術を用いたプラットフォームを複数種類開発した。医療データは、サンプル数に比してパラメータ数が多いという新 NP 問題と呼ばれる課題など、AI 技術を有効に活用する上で解析しないといけないいくつかの問題がある。実際ゲノム・エピゲノム解析においては、医用画像解

析と比して未だ AI 技術の導入は進んでいないという現実もある。一方、今回我々が開発した手法は、大規模ながんオミックスデータを診療情報と合わせて効率的かつ精度高く解析することが可能であり、科学技術イノベーションに大きく寄与する成果であると判断している。

＜代表的な論文＞

1.

概要：Nature Communications 10, 3925 (2019)；我々は、細胞の老化過程で蓄積したDNA損傷を引き金として、変異のリスクが上がるゲノム不安定性の状態になり、またこれに伴う変異導入に起因して、がん抑制遺伝子の変異したクローン進化が誘導されることを *in vitro* のモデル解析系を用いて示すことに成功した。本研究により、ゲノム不安定性のリスク要因、ゲノム不安定性リスクの上昇要因への研究展開が期待され、その解析からは、外的なリスク要因等（放射線や紫外線などを含む）による発がんへの影響とそのリスク要因が明確になることも期待される。

2.

概要：Biomolecules, 10, 524 (2020)；我々は TCGA lung adenocarcinoma データ（mRNA 及び miRNA）を解析する手法として、深層学習技術を活用した新手法の開発に取り組んだ。その結果、autoencoder を活用したマルチオミックス解析により、肺がん予後良好群と予後不良群を分類することに成功した。本研究により、肺がん予後を規定する遺伝子を同定することに成功したが、これらの遺伝子は新規肺がん創薬標的となる可能性が示唆された。

3.

概要：Biomolecules, 10, 1249 (2020)；我々は深層学習を中核とした機械学習技術を用いることで、がんの表現型を特徴付ける新規因子の探索を行うことを目指し、改変型 Diet Networks を用いた解析プラットフォームを構築した。その結果、80%という高い精度をもって、肺がん病理型（lung adenocarcinoma 及び lung squamous cell carcinoma）の分類を行うことに成功した。隣接行列からグラフ構造を描画し、各因子の近接性を可視化して評価できる形式に変換したところ、最終的に大きな二つのクラスタが形成され、これが lung squamous cell carcinoma あるいは lung adenocarcinoma それぞれの出力に対する各因子の寄与率に一致しているだろうことが考察された。

§2 研究実施体制

(1)研究チームの体制について

① 浜本グループ
研究代表者:浜本 隆二(国立がん研究センター研究所がん分子修飾制御学分野 分野長)
研究項目:研究計画全体の統括、肺がん統合データベース及び主要解析パイプラインの構築

② 岡野原グループ
研究代表者:岡野原 大輔(Preferred Networks 取締役副社長)
研究項目:肺がん統合データベースの構築及び深層学習技術を用いたデータ解析

③ 瀬々グループ
研究代表者:瀬々 潤(産業技術総合研究所人工知能研究センター 招聘研究員)
研究項目:肺がん統合データベースの構築及び機械学習技術全般を用いたデータ解析

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について
共同研究先:国立がん研究センター研究所
*がんゲノム生物学研究分野
*細胞情報学分野

共同研究先:国立がん研究センター中央病院
*医療情報部
*放射線診断科
*放射線治療科
*呼吸器内科
*呼吸器外科
*病理・臨床検査科

共同研究先:理化学研究センター革新知能統合研究センター
*上田修功・副センター長

共同研究先:医薬基盤・健康・栄養研究所・AI健康・医薬研究センター
*夏目やよい・サブプロジェクトリーダー

共同研究先:京都大学大学院情報学研究科・知能情報学専攻
*黒橋禎夫・教授

共同研究先:奈良先端科学技術大学院大学先端科学技術研究科・情報科学領域
*荒牧英治・教授

共同研究先:富士フイルム株式会社
*メディカルシステム開発センターIT開発グループ