大規模データ解析と人工知能技術による
がんの起源と多様性の解明

文部科学省「富岳」成果創出加速プログラム

Newsletter

No. 2

2021年6月

Unravelling origin of cancer and diversity by large-scale data analysis and artificial intelligence technology

「富岳」成果創出加速プログラム

大規模データ解析と人工知能技術による がんの起源と多様性の解明

説明可能なAIによる 大規模遺伝子ネットワーク解析

課題参加者　東京医科歯科大学M&Dデータ科学センター 教授　Heewon Park
連携研究者　富士通研究所　プロジェクトマネージャ　丸橋弘治

疾患のシステム異常の解明を目指して大規模遺伝子制御ネットワーク解析やその方法論に関する研究を行っています。

遺伝子ネットワークは遺伝子発現の因果関係や相互作用をネットワークとしてモデル化したグラフ構造のデータです。複雑なメカニズムを持つ疾患のシステム異常を理解するためには、遺伝子ネットワークの解析に基づく分子異常メカニズムの解明は大変重要です。 しかしながら、従来のネットワーク解析研究では、全てのサンプル（細胞・患者集団）のデータから、単一のネットワークを推定し、それに基づく解析・解釈を行ったため、患者個々の疾患のシステム異常の解明はできませんでした。その問題を解決するため、「次世代スパコンプロジェクト」時代からNetworkProfilerを始め様々なネットワーク解析技術「SiGNシリーズ」を開発してきました。NetworkProfilerは患者（サンプル・細胞）のあるバイオメディカルな特徴を基準にして、その類似度に基づいて解析の対象であるターゲット患者と似た特徴を持つ患者集団の情報を利用し、患者個々人の特有な遺伝子ネットワークを構築する技術です（図1）。 この技術の活用し、細胞の薬剤感受性による遺伝子ネットワーク（***PLoS ONE***. 2014; 9(10): e108990）やがんの悪性化に関わる重要なメカニズムである、細胞形質変化を表す上皮間葉転換(Epithelial-Mesenchymal Transition: EMT)の度合による遺伝子ネットワークの推定や解析に関する研究をやって来ました（***PLoS ONE***. 6(6): e20804, 2011）。

しかしながら、約2万個の遺伝子から構成されている数百個の遺伝子ネットワークをどういうふうに解釈すればよりかがいつも大きな壁でした（図2）。人間の目で見ても解釈不可能であり、従来の研究では疾患の重要なマーカーとして知られている遺伝子に着目し、その周りを見るような局所的な解析しかできませんでした。しかし、複雑なメカニズムを持つがんなどの疾患の生体システム異常を理解するためには、複数のネットワークの総合的な解析・解釈が必須・不可欠です。 残念ながら、このような研究は未だ行われておらず、それを可能にする計算科学のインフラもまだできていない状況でした。

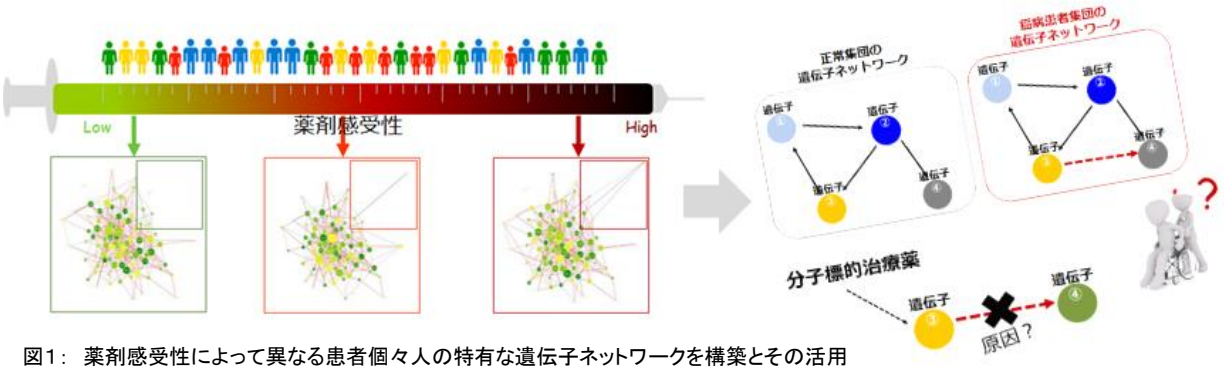

図1： 薬剤感受性によって異なる患者個々人の特有な遺伝子ネットワークを構築とその活用

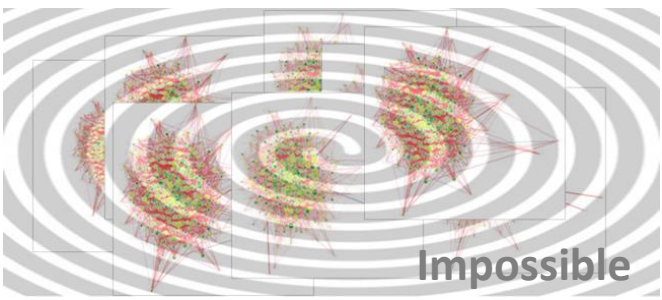

図2： 約2万個の遺伝子から構成されている数百個の膨大・大量の遺伝子ネットワーク解釈の壁

その壁を超えるため、最近様々の分野で注目されている人工知能（AI）技術の活用を考慮しましたが、従来のAI技術は人間を超える高精度の判別・予測を実現できますが、データから何を学習して、どんな根拠でそのような結果が導出されたのかの解釈性問題（中身がわからないBlack Box問題）があり、予測精度だけではなく膨大な遺伝子ネットワークの解釈から疾患のメカニズムを解明を目指す我々の研究では不十分な術でした。すなわち、専門家の判断に説明責任が問われる医療分野などへのAIの適用にはまだ限界ある状況です。

本研究プロジェクトでは、数理的な高精度だけではなく高解釈性を持つ説明可能AI技術DeepTensor（***AAAI 2018***. 3770-3777, 2018)の改善版であるTRIP（***arXiv: 12007.03912***, 2020）を富士通研究所の人工知能研究所・愛知県がんセンター研究所システム解析学分野との共同研究で開発しました。

DeepTensorは、遺伝子ネットワークのような頂点間の関係の強さを要素とする行列、あるいは行列を拡張したテンソルの分解を応用し、多数の頂点の線形結合で表現される少数の成分を抽出し、それらの成分の間の関係によって元のグラフ構造データを近似します。このとき、予測に必要とされる特徴が極力保持されるように抽出される成分を学習し、元の頂点の線形結合として表される抽出成分の間の関係が予測値の変動に与える影響を分析することにより、グラフ構造データ全体の構造の観点から予測理由提示する説明可能AI技術です（図３）。本研究では、遺伝子ネットワーク解析への適用を通じて、Deep Tensorの説明可能性を高めたTRIP技術を開発しました。また、Deep Tensor/TRIPの「富岳」上での適用のため、コード最適化を実施し、最適化実施前に比べ約４倍の高速化に「富岳」上で成功しました（図４）。

開発されたTRIPを用いて上皮系細胞から間葉系細胞への転換(EMT)を決定する遺伝子ネットワークの解析し、EMTメカニズム解明を行いました（***PLoS ONE***. 15(11); e0241508)。 EMTを決定する遺伝子ネットワークは英国サンガーセンターが公開しているCell Line Projectの762種のがん細胞における約13,000個の遺伝子発現データに基づいて、上皮・間葉転換の度合を定量化した変量を細胞の特徴として使用し構築された762個の遺伝子ネットワークです（***PLoS ONE***. 6(6): e20804, 2011)。

TRIPを使ってその膨大・大量の遺伝子ネットワークを表す重要な成分を抽出し、その成分の解析に基づいて、EMTを決定する762個の遺伝子ネットワーク全体構造の総合的な解析を行いました。

抽出された重要な成分中、Regulator側（影響を与える遺伝子）の第1・2・3成分に着目して、各成分の領域（図5）におけるEMT転写因子（ZEB1, ZEB2, SNAIL1, SNAIL2, TWIST1）を中心にネットワークを解析し、EMTメカニズムの解明やEMTメカニズムに関わっている重要なマーカー（遺伝子）探索を行いました。

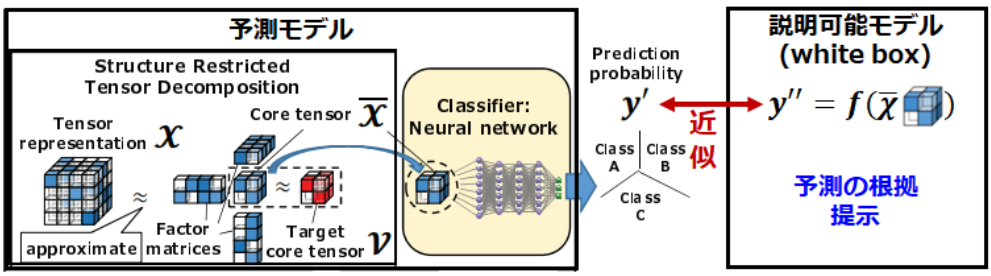

図３：DeepTensorの解析プロセス － 頂点間の関係の強さを要素とする行列・テンソル形式のデータから予測に必要とされる特徴が極力保持されるように重要成分と推定し、その成分を解釈することによって予測理由の説明を可能にするAI技術 （AAAI 2018. 3770−3077, 2018）

	最適化実施前	最適化実施後
Intel (1 socket)	783 秒	278 秒
GPU (V100)	762 秒	358 秒
A64FX 搭載 FX700 (1 CMG)	1,678 秒	443 秒

図４：Deep Tensorの100 epochあたりの学習時間：Deep Tensor/TRIPの「富岳」上でのコード最適化を実施し、最適化実施前に比べ約4倍の高速化に成功

取られたマーカーには、先行研究で発見されたEMTマーカーも含まれていていましたが、その時点では立証できなかったEMT関連メカニズムが、10年経った今、ほとんど解明できていることが確認できました。すなわち、本プロジェクトで開発された説明可能AIにより、10年前のデータから10年間のEMTメカニズム関連研究成果を一挙に取りまとめることができました。

この研究は、説明可能AI技術による大規模遺伝子ネットワーク解析の最初の研究であり、今後、データから得られた解析結果に関する解釈性を持つ説明可能AI技術を発展させることにより、判断に説明責任が問われる医学分野を含め様々な分野に大きく貢献できると期待されます。

また、本研究で行った説明可能AI技術による遺伝子ネットワーク解析研究は、まだ端緒についたばかりであり、改善すべき点が多いと考えられます。富岳の計算パワーを活用し、それらの問題を改善・発展させ、高精度だけではなく、高解釈性を持つ、医療現場で実用化可能な説明可能AI技術に関する研究を続けてやっていきます。

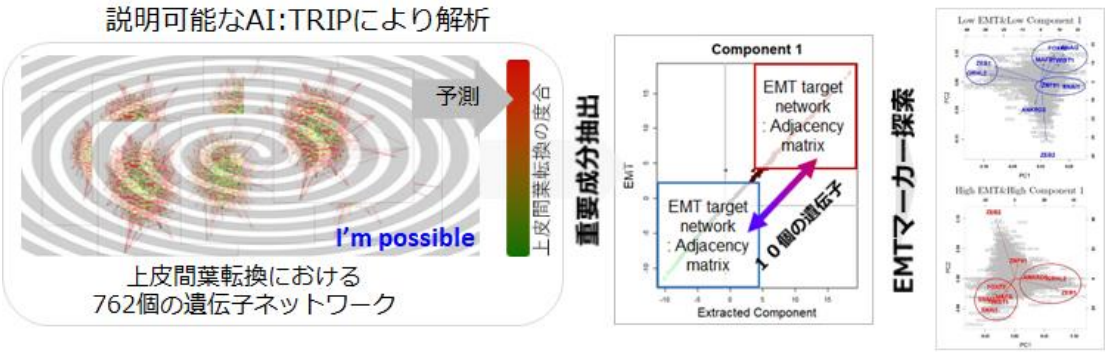

図5: TRIPによる上皮系細胞から間葉系細胞への転換を決定している762個の遺伝子ネットワーク解析とEMTマーカー探索のイメージ図（医学のあゆみ. 276(9): 21674-21678）

1. 代表機関

国立大学法人東京医科歯科大学

研究課題開発責任者

M&Dﾃﾞｰﾀ科学ｾﾝﾀｰ　特任教授　宮野悟

2。協力機関

国立大学法人京都大学

協力機関代表者

大学院医学研究科　教授　小川誠司

愛知県がんセンター

協力機関代表者

研究所ｼｽﾃﾑ解析分野　分野長　山口類

3. 連携機関

（株）富士通研究所

連携機関連絡担当者

富士通株式会社 富士通研究所 フェロー

岡本青史

国立大学法人東京大学

連携機関連絡担当者

医科学研究所　准教授　片山琴絵

Karolinska Institutet

連携機関連絡担当者

Department of Medicine/Center for Hematology and Regenerative Medicine Visiting Professor 小川誠司

4. 事業参加者

事業参加者（代表機関）

東京医科歯科大学M&Dﾃﾞｰﾀ科学ｾﾝﾀｰ

統合解析分野　特任教授　宮野悟

講師　長谷川嵩矩

助教　伊東聰

特任助教 角田将典

研究支援者 田中洋子

AI技術開発分野　教授　Heewon Park

事業協力者（協力機関）

京都大学大学院医学研究科

腫瘍生物学講座　教授　小川誠司

特定教授　南谷泰仁

助教　越智陽太郎

愛知県がんセンター研究所

ｼｽﾃﾑ解析分野　分野長　山口類

連携参加者（連携機関）

富士通株式会社 富士通研究所 フェロー

岡本青史

富士通研究所　プロジェクトマネージャ

丸橋弘治

東京大学医科学研究所

ﾋﾄｹﾞﾉﾑ解析ｾﾝﾀｰ　准教授　片山琴絵

先端医療研究ｾﾝﾀｰ　教授　南谷泰仁

（2021年6月07日時点）

スーパーコンピュータ「富岳」成果創出加速プログラムについて（理化学研究所計算科学研究センターより抜粋）

https://www.r-ccs.riken.jp/jp/fugaku/promoting-researches

スーパーコンピュータ「富岳」成果創出加速プログラムは、「富岳」を用いた成果を早期に創出することを目的として文部科学省が設置しました。①人類の普遍的課題への挑戦と未来開拓、②国民の生命・財産を守る取組の強化、③産業競争力の強化、④研究基盤の4領域について課題の選定が行われ、19課題が選定されています。選定された課題は、「富岳」の計算資源を優先的に無償で使用し、速やかな成果創出を目指します（2020～2022年度）。

Information

TOKYO MEDICAL AND DENTAL UNIVERSITY
M&D Data Science Center

文部科学省「富岳」成果創出加速プログラム
課題名：大規模データ解析と人工知能技術によるがんの起源と多様性の解明
ニュースレター No.2
発行日★2021年6月7日
課題代表者★宮野 悟

- 東京医科歯科大学M&Dデータ科学センター　統合解析分野
- 郵便物宛先：　〒113-8510　東京都文京区湯島1-5-45
- 宅配便宛先：　〒101-0062　東京都千代田区神田駿河台2-3-10
  駿河台キャンパス22号館5階
- TEL: 03-5280-8620 FAX: 03-5280-8632
- E-mail: mdsc.dsc@tmd.ac.jp
- HP: https://odcla.mddsc.jp