

説明可能なAIによるゲノム医療の新世界

Heewon Park¹, 丸橋弘治², 山口類^{3,4}, 宮野悟⁵

1. 東京医科歯科大学 M&D データ科学センター AI 技術開発分野
2. 株式会社富士通研究所人工知能研究所トラステッドAIプロジェクト
3. 愛知県がんセンターシステム解析学分野
4. 名古屋大学大学院医学系研究科がんシステム情報学分野
5. 東京医科歯科大学 M&Dデータ科学センター

Abstract

患者個々はそれぞれ違うゲノムの特性を持ていて、疾患の治療の成功率を向上させるためにはそれぞれのゲノム特性に合わせた個別化治療が必要である。また、取られてSample-specific特徴を総合的に解析することから疾患のシステム異常の解明と、それに基づく治療法の改善は不可欠である。がんなどの複雑なメカニズムをもつ疾患の治療における計算科学の役割はデータから治療のエビデンスを取り出すことであり、そのため色んな技術が開発・活用されている。しかしながら、高度情報処理と計測技術の進歩により、データの規模は超巨大になって来て富岳などの高精度のスーパーコンピュータは必修・不可欠になって来た。本研究では、富士通人工知能研究所との共同研究で、**説明可能なAI技術「DeepTensor」の改善版を開発し、それに基づいて超大規模かつ大量の遺伝子データを「富岳」上で計算することによって、細胞の上皮系細胞から間葉系細胞への転換の予測とその予測根拠を人間が理解できる形で取り出すことを成功した。**

Keywords:

- ・ 遺伝子ネットワーク
- ・ スーパーコンピューティング
- ・ 説明可能なAI
- ・ Sample-specific analysis
- ・ 上皮間葉転換関連ネットワーク
- ・ 薬剤耐性ネットワーク

説明可能なAIによるによる**超**大規模遺伝子ネットワーク ワーク解析

ロネットワーク推定

従来のネットワーク推定方法：すべてサンプル（データを構成しているすべての細胞）に対して単一のネットワークを計算するため、個別化医療のエビデンスである患者個々人の特徴を取り出すことが出来なかった。

→ その限界を超えるため、次のVarying coefficient modelを考える。

$$Y_i = \sum_{j=0}^q \beta_{ji}(m_\alpha) X_j + \varepsilon_i \quad (1)$$

ただし、 $X_1 \dots X_q$ は q 個のregulator遺伝子、 Y_i はtarget遺伝子の発現量であり、 m_α は α 番目の細胞の特徴を表す変数（細胞の薬剤耐性など）である。 $\beta_{ji}(m_\alpha)$ は α 番目の細胞における X_j の Y_i への因果関係の強さを表す回帰係数（varying coefficient）である。

→ **細胞の特徴 m_α の値によって異なる遺伝子間の因果関係推定が可能。**

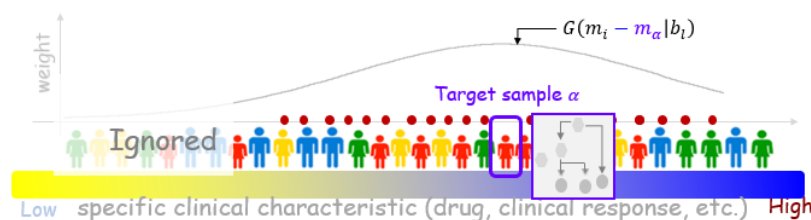
Varying coefficientの推定：NetworkProfiler (Kernel-based L1-type regularization method; Shimamura et al., 2011)

$$\operatorname{argmin}_{\beta_{i\alpha}} \left\{ \frac{1}{2} \sum_{i=1}^n \left(y_{ii} - \sum_{j=1}^q \beta_{ji\alpha} X_{ij} \right)^2 G(m_\alpha - m_i | b_i) + P(\beta_{i\alpha}) \right\} \quad (2)$$

ただし、 $P(\beta_{i\alpha})$ はelastic netタイプのペナルティ(Zou et al., 2005)であり、

$$G(m_\alpha - m_i | b_i) = \exp\left(-\frac{(m_\alpha - m_i)^2}{b_i}\right). \quad (3)$$

→ **細胞の特徴を表すmodulatorの類似度に基づいて、似たようなclinical特徴を持つ細胞の情報のみを使ってsample-specificネットワーク推定**

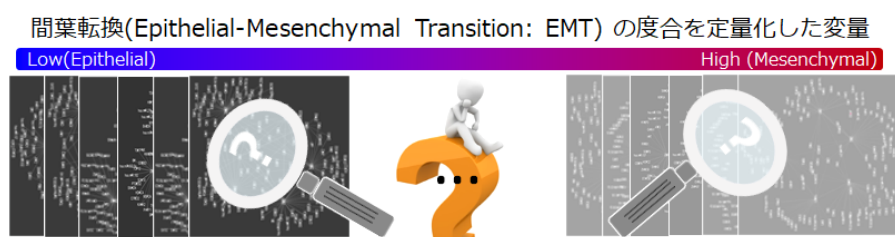


ロ NetworkProfilerによる上皮間葉転換関連ネットワーク

- ・ Sanger Cell Line Projectの762のがん細胞 (Shimamura et al., 2011)
- ・ 細胞形質変化を表す上皮・間葉転換 (Epithelial-Mesenchymal Transition: EMT) の度合を定量化した変量を定義 (EMT modulator)
- ・ NetworkProfilerを使って、各細胞のEMT modulatorの値に対する762個のネットワークを推定

→ **個別化医療をはじめ、network biology研究の大きく貢献**

✓ 構築された大量・大規模ネットワークの局所的な解析の限界



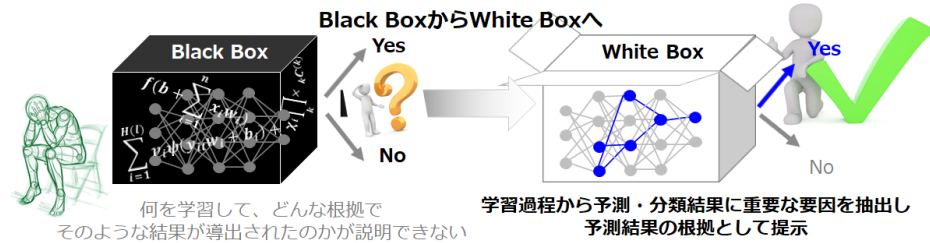
約13,000個の遺伝子からなる762個の遺伝子ネットワーク

- ・ **そこからどうする？** という壁が...
- ・ がんなどの複雑なメカニズムを持つ疾患のメカニズム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
- ・ **このような研究はまだ.. それを可能にする計算科学インフラもまだ...**

ロ 説明可能なAI(XAI)によるネットワーク解析の技術開発

従来の(AI)技術 (Black Box²)

1. 高精度予測能力
2. **どんな根拠でそのような結果が導出されたのかが説明不可**
3. 判断に説明責任が問われる医学・生物学分野への適用は困難



→ **高精度予測能力+解釈性を持つAI技術**

：説明可能AI (Explainable AI: XAI)技術の開発

：Tensor Reconstruction based Interpretable Prediction (TRIP)

[Stage 1] Sample-specific 遺伝子ネットワーク推定

：NetworkProfiler(2)を用いて、細胞のclinical特徴 m_α によって異なる遺伝子ネットワーク推定

[Stage 2]説明可能AIによる情報抽出

a.人間が理解可能な低次元部分空間を推定し、stage1で推定された大量・大規模ネットワークデータをその部分空間に射影する。

$$\bar{X}_i = X_i \prod_k \times_k C^{(k)}$$

ただし、 X はサイズ $I_1 \times \dots \times I_K$ の K -modeテンソルであり、 $C^{(k)} \in \mathbb{R}^{I_k \times K}$ は射影行列である。

b. 射影されたネットワークテンソル $\bar{X}_i$ に基づいて、目的変数を予測する。

$$\hat{y}_i = f(\bar{X}_i, \theta) = f^*(r_i, \theta')$$

ただし、 $r_i = M_i^T \text{vec}(\bar{X}_i)$, M_i は重みテンソル、 θ' は M_i 以外のパラメータである。

TRIPは次の目的関数を最小化することによって、ネットワークの射影と予測を行う。

$$O_T = \frac{1}{n} \sum_{i=1}^n \{ \mathcal{L}(y_i, \hat{y}_i) + \gamma \| X_i - \bar{X}_i \prod_k \times_k C^{(k)T} \|^2 \}, \quad \text{s.t.} \quad C^{(k)T} C^{(k)} = I$$

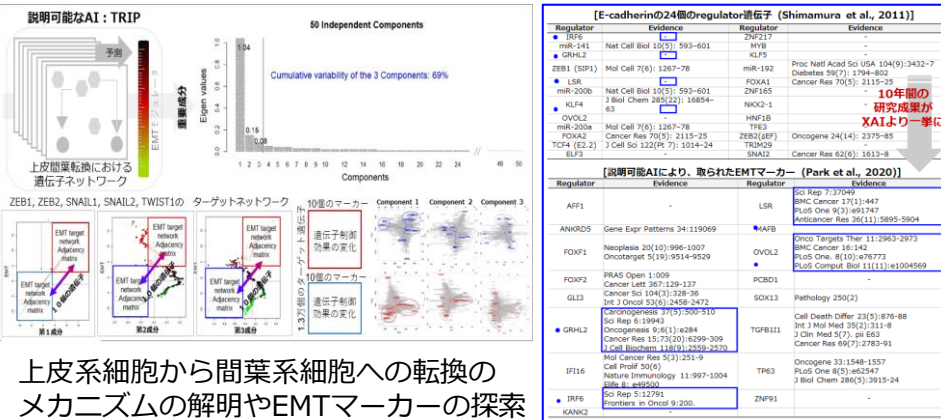
パラメータ推定の細かいことは、Maruhashi et al. (2020)を参考せよ。

ロ 説明可能なAIによるEMT関連ネットワーク解析

：ネットワークの**局所的な解析の限界を超える！** (Park et al., 2020)

：TRIPによる762個のEMT関連ネットワークの**総合的に解析**

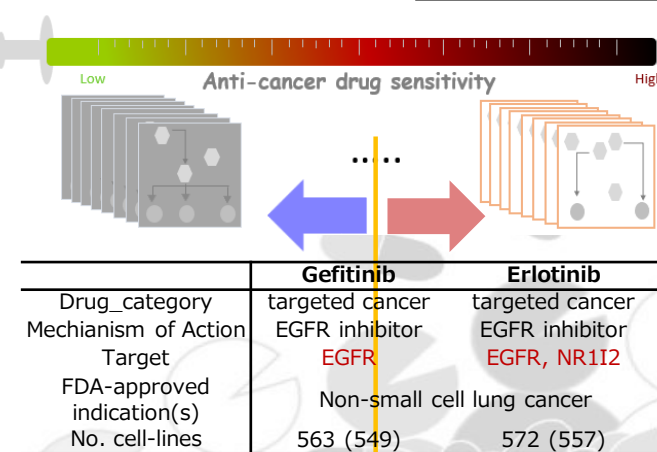
：皮系細胞から間葉系細胞への転換の予測



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が一挙に取りまとめられた！

ロ 説明可能なAIによる薬剤耐性ネットワーク解析



薬剤感受性・耐性を支配している遺伝子ネットワーク推定

Motivation :

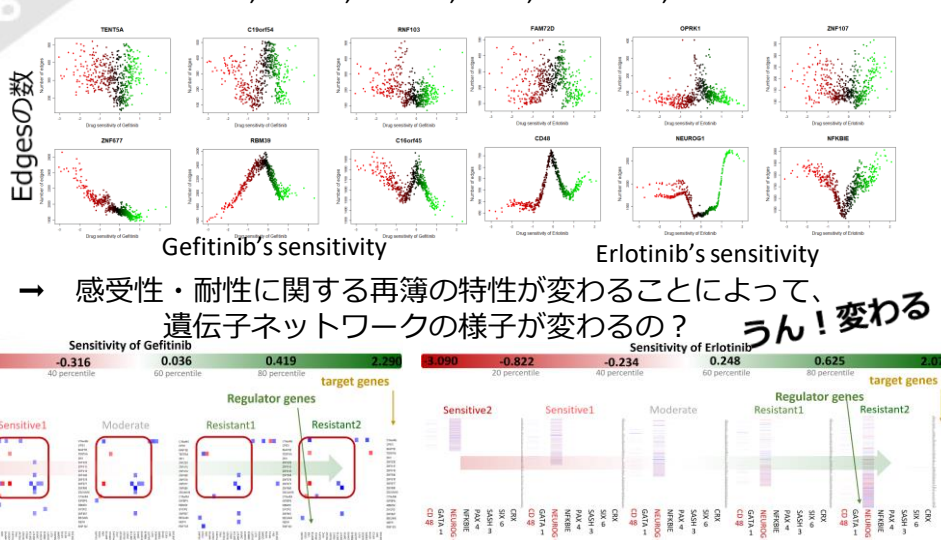
がん細胞が薬剤耐性を獲得しないようにするには？

説明可能なAIによる薬剤耐性ネットワーク解析

ロ 薬剤感受性・耐性を支配する遺伝子間の因果関係探索

Gefitinib: TENT5A, C19orf54, RNF103, ZNF677, RBM39, C16orf45

Erlotinib: FAM72D, OPRK1, ZNF107, CD48, NEUROG1, NFKBIE



どの遺伝子間の因果関係が薬剤耐性・感受性を支配しているか…？

Reference

- [1] Maruhashi K et al. arXiv:2007.03912 2020.
- [2] Park H et al. PLoS One 2014; 9(10): e108990.
- [3] Shimamura T et al. PLoS ONE 2011;6(6): e20804.
- [4] Zou H and Hastie T, J Royal Stat Sci B. 2005; 67(2): 301-320.

終わりに 本研究で行った説明可能AI技術による大規模ネットワーク解析技術は世界初の研究であり、まだ端緒についたばかりである。**富岳の計算力**を基²づいて、**本研究を改善・発展させ、高精度・高解釈性をもつ、医療現場で実用化できるように発展し続けて、社会に貢献できる形で研究を進めて行きたいと思う。**