厚生労働行政推進調査事業費（化学物質リスク研究事業）
トキシコゲノミクスとシステムバイオロジーとの融合による
新型化学物質有害性評価系の実装研究
（２１ＫＤ２００１）

令和 3 年度　分担研究報告書

システムバイオロジーによる毒性解析の AI 化

研究分担者　　北野 宏明
特定非営利活動法人 システム・バイオロジー研究機構
会長

研究要旨

システム毒性では、一連の解析手順の高度な連携と同時に、大規模データベースから多くの情報を抽出し、それを解析へと結びつける必要がある。本分担研究では、深層学習(Deep Learning)を用いて膨大な遺伝子変動データから有意に変動した遺伝子を高精度で自動同定させる技術ならびに解析パイプラインの連動強化を行った。

研究協力者

長谷 武志　　特定非営利活動法人
システム･バイオロジー研究機構

Natalia Polouliakh　株式会社ソニーコンピュータ
サイエンス研究所

Ａ．研究目的

システム・レベルで毒性を理解するには、膨大な実験データを格納したデータベース、文献、数値モデルなどを統合的に解析する必要があり、大規模かつ複雑なデータを意味のある形で解析するには、深層学習やテキストマイニングなどを含めた一連の人工知能（AI）アルゴリズム群の連携が有効である。さらに、複数の解析ツールをスムーズに連動させる必要がある。本分担研究では、一連の解析過程の AI 化を実施し、ツール間連動を強化することで、高度な AI 駆動型システム毒性学基盤の構築を推進する。

Ｂ．研究方法

システム・レベルで毒性を理解するには、膨大な実験データを格納したデータベース、文献、数値モデルなどを統合的に解析する必要があり、大規模かつ複雑なデータを意味のある形で解析するには、深層学習やテキストマイニングなどを含めた一連の人工知能（AI）アルゴリズム群の連携が有効である。本分担研究では、一連の解析過程の AI 化を実施する。

●深層学習を用いた大規模遺伝子発現データベースからの重要遺伝子群の判別

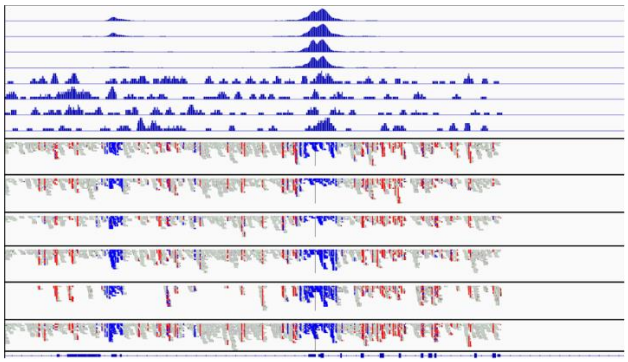
先行研究で開発した、深層学習を用いた３次元グラフの画像解析システム DTOX について改良、即ち特異パターンの追加学習と GUI 実装の改良を進めた。

追加学習用の画像セットは、遺伝子発現を用量×時間×発現量（Percellome 法により細胞 1 個あたりの mRNA コピー数に換算したデータ）の 3 次元グラフに描画したものを用いた。また GUI 実装は、python

の代表的な GUI 作成用のモジュールである、PYQT5 と、Qtdesigner を用いて改良を進めた。

●深層学習を用いたエピゲノム解析データからの有意なエピゲノム修飾の判別

エピゲノム解析では、曝露下の遺伝子のエピゲノム修飾（ヒストン修飾およびゲノム DNA メチル化の状態）を、下図に示すような解析画像として表示し、研究者の判断を助けている。



図．エピゲノム解析画像の例
上から８行で示される分布がクロマチン修飾の状況を表しており、下から８行のプロットがメチル化の状況を表している。横軸は配列の位置を表している。メチル化については、青いプロットがメチル化されていない状態を表している。

有意なエピゲノム修飾を同定するために、長年の経験を積んできた研究者が、それぞれの遺伝子に対するエピゲノム解析画像を検討し、分類を行ってきた。しかしながら、化学物質数×遺伝子数の解析画像が存在し、網羅的に有意なエピゲノム修飾を同定するには、多大な時間と労力が必要となっている。これを解決すべく、エピゲノムデータ（解析画像）から、効率良く有意なエピゲノム修飾を判別する手法の構築を目的として、深層学習モデルの構築を行った。

深層学習モデルの訓練には、専門家により分類されたエピゲノム解析画像を訓練データとして用いた。このエピゲノム画像データは、四塩化炭素、バルプロ酸ナトリウム、クロフィブラートの新型反復曝露と単回曝露において得られたものである。各遺伝子に対するエピゲノム解析画像を、専門家が視覚的に検証して、以下の３群に分類している。

① suppression 群：反復曝露によるエピゲノム修飾により、遺伝子発現が抑制されたもの（5,937 画像）
② induction 群：反復曝露によるエピゲノム修飾により、遺伝子発現が誘導されたもの（457 画像）
③ non significant 群：反復曝露によるエピゲノム修飾により、遺伝子発現が有意な変動をしめさなかったもの（2,349 画像）

エピゲノム画像データの 80%をトレイニングデータとして用いて深層学習モデルを構築し、残りの 20%のデータをテストデータとして用いて構築したモデルの分類精度の検証を行った。

●転写領域解析ソフトウェア SHOE の改良

SHOE の開発は、Java 言語（USA, Oracle Inc.）で行った。Garuda Platform 用ソフトウェア（Garuda ガジェット）の開発や他の Garuda ガジェットとの連動については、GarudaDevPack を使用した。性能評価や試験運用には、Percellome データベースより実際の化学物質曝露による遺伝子発現時系列データを用いた。

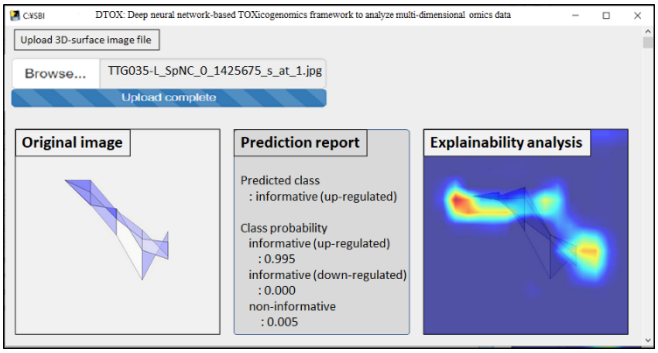
C．研究結果

●深層学習を用いた大規模遺伝子発現データベースからの重要遺伝子群の判別

トレイニング等に使用していないデータセットに対し、先行研究で開発した深層学習モデル DTOX を適用して予測を実行し、その結果を専門家の判定結果と比較して、予測が食い違った遺伝子の 3 次元グラフ画像を集め、追加学習を試みた。例数が少なかったためか、その効果は大きなものでは無かったが、過学習等の悪影響も起こらず、深層学習モデルによる DTOX の予測精度は、一般的なバイオインフォマティクス解析パイプラインによる予測精度を大きく上回った。

また研究者等のユーザーが使いやすいグラフィカルユーザーインターフェイス（GUI）の実装及び改

良を進めた。



このインターフェースデザインでは、まず Browse ボタンをクリックして、遺伝子の三次元画像ファイルを選択し読み込む。読み込んだ画像は、Original image のサブウインドウ(左側)に表示され、同時に画像解析システム中の深層学習モデルで解析され、予測が行われる。予測結果は、Prediction report のサブウインド(中央)に表示される。この画像の例では、この遺伝子は化学物質により up-regulation されている確率が 0.995 であると予測されている。また、深層学習モデルの判断根拠に関する解析も行われ、その結果は、Explainability analysis サブウインドウ(右側)に表示される。判断根拠とされた画像領域は赤色でハイライトされる。

現在、連続処理機能やその際のレポート作成機能の実装を検討しているほか、DTOX に関する成果の原著論文化を進めている。

（参考文献）

1．David Silver et al (2006) Mastering the game of Go with deep neural networks and tree search. Nature 529:484 – 489.
2． Andre Esteva et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 52:115-118.
3． Yoshimasa Sakai, Satoko Takemoto, Keisuke Hori, Masaomi Nishimura, Hiroaki Ikematsu, Tomonori Yano and Hideo Yokota (2018) "Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network", 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
4．Kaming He et al (2015) Deep Residual Learning for Image Recognition. arXiv:1512.03385.

●深層学習を用いたエピゲノム解析データからの有意なエピゲノム修飾の判別

エピゲノム解析画像を①suppression 群、②induction 群、③non significant 群の三群に分類することを目的として、多様な深層学習アーキテクチャに基づく分類モデルを構築した。使用した深層学習アーキテクチャは 8 種類であり、それらの内最も複雑なものは 121 層のレイヤーで構築されている（下表）。

深層学習アーキテクチャ	層の数	参考文献
resnet18	18	https://arxiv.org/abs/1512.03385
resnet34	34	https://arxiv.org/abs/1512.03385
resnet50	50	https://arxiv.org/abs/1512.03385
alexnet	8	https://dl.acm.org/doi/10.1145/3065386
densnet121	121	https://arxiv.org/abs/1608.06993
squeeznet_0	15	https://arxiv.org/pdf/1602.07360v3.pdf
vgg16	16	https://arxiv.org/abs/1409.1556
vgg19	19	https://arxiv.org/abs/1409.1556

これらの深層学習アーキテクチャは、1000 カテゴリに分類できる 120 万枚の画像で構成されるデータ(ImageNet dataset)を用いてプレトレイニングされたものである。このプレトレイニングされた深層学習アーキテクチャを、上記のエピゲノム解析画像に対して転移学習を行うことで、分類モデルを構築した。

エピゲノム解析画像データの 80%をトレイニングデータとして用いて深層学習モデルを構築し、残りの 20%のデータをテストデータとして用いて構築したモデルの分類精度の検証を行った結果、下図に示すように、構築した 8 種の分類モデル全てで、non-significant 群と、それ以外の 2 群を正確に分類することが出来た。

予測

正解

Resnet18

	induction	ns	supression
induction	0	3	89
ns	0	466	4
supression	0	4	1184

Resnet34

	induction	ns	supression
induction	0	2	90
ns	0	468	2
supression	0	3	1185

Resnet50

	induction	ns	supression
induction	0	2	90
ns	0	465	5
supression	0	3	1185

Alexnet

	induction	ns	supression
induction	0	2	90
ns	0	461	9
supression	0	4	1184

Densnet121

	induction	ns	supression
induction	0	3	89
ns	0	466	4
supression	0	2	1186

Sqeeznet1.0

	induction	ns	supression
induction	0	1	91
ns	0	465	5
supression	0	4	1184

vgg16

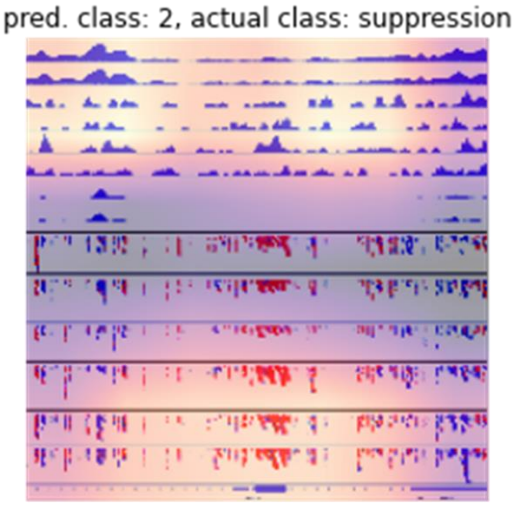
	induction	ns	supression
induction	0	3	89
ns	0	467	3
supression	0	3	1185

vgg19

	induction	ns	supression
induction	0	1	91
ns	3	464	3
supression	0	3	1185

このことは、構築した深層学習モデルは、有意なエピゲノム解析画像を正確に抽出できる、つまりは、反復暴露により生じるエピゲノム修飾の中でも遺伝子変動に影響を与えるものを見分けることが出来る可能性を示唆している。また、suppression 群も、8 種の深層学習モデル全てで、正確に分類することが出来た。しかしながら、構築した 8 種類のモデル全てで、induction 群を分類することが出来なかった。

induction 群の分類不調の原因を調べるために、構築した深層学習モデルが画像のどの部分に着目して分類を行っているのかを示すべく、代表的な explainability model (grad cam : R. R. Selvaraju, M. et al (2019) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis doi:10.1007/s11263-019-01228-7) を用いて深層学習モデルの解析を行った。解析結果は、下図の様に示され、深層学習モデルが図中の明るい部分に着目して分類を行っていることが示された。



現在、専門家と協力し、着目している部分の生物学的な解釈を進めている。

(参考文献)

1. Andre Esteva et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 52:115-118.
2. R. R. Selvaraju, M. et al (2019) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis doi:10.1007/s11263-019-01228-7
3. Ian J. Goodfellow et al. (2014) Generative Adversarial Networks. https://arxiv.org/abs/1406.2661

●転写領域解析ソフトウェア SHOE の改良

今年度は、GARUDA プラットフォーム上で動作する SHOE ガジェットの開発や、SHOE 本体の機能追加などを進めた。また一部の環境でインストールできないなどの事例があったため、原因調査と対応策の検討を進めた。

D．考察

本分担研究においては、独自開発した深層学習遺伝子発現グラフ画像解析システム DTOX の予測精度をより一層高め、一般的なバイオインフォマティクス解析パイプラインの精度を大きく上回る性能を示した。これは、DTOX の深層学習モデルが、研究者によるヒューマンキュレーションの際に評価している領域を着目して判定していることから、トレイニングにより

人の判断のパターンを上手く捉えることで、高い精度を実現している可能性を示唆している。

エピゲノム解析においても、深層学習モデルは、反復曝露によるエピゲノム修飾の内、遺伝子発現に影響がある修飾を見分けることが出来る可能性があることが示唆された。今回、suppression 群を正確に分類することが出来たが induction 群を見分けることはできなかった。これは、induction 群の訓練用画像の枚数が少なく、induction 群に関する情報が十分に学習できていないことが原因であると考えられる。現在、この問題に対処するために、画像の生成モデル（Generative adversarial network;　Ian J. Goodfellow et al. (2014) Generative Adversarial Networks. https://arxiv.org/abs/1406.2661）を活用して、induction 群の訓練用画像の増幅を進めている。

E．結論

本分担研究についてはほぼ計画通り推移した。先行研究により開発した解析用ソフトウェアは実用の段階に進みつつある。また新たに開始したエピゲノム解析の AI 自動化についても、充分な訓練用画像があれば、より正確に反復曝露により遺伝子発現に影響を与えうるエピゲノム修飾を予測できることが示唆されている。現在不足しているカテゴリ（反復曝露によるエピゲノム修飾により、遺伝子発現が誘導されるもの）の遺伝子は実データにおいて他のカテゴリより少ないため、今後は不足している訓練用画像を AI で増幅する手法（Generative adversarial network）の活用などにより深層学習モデルの訓練状況を改善し、実用レベルの予測性能を目指す。

G．研究発表

１．論文発表

(1)　Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. npj Syst Biol Appl. 7, 29 (2021).[DOI: 10.1038/s41540-021-00189-3]

(2)　Ostaszewski M, Niarakis A, Mazein A, Kuperstein I, Phair R, Orta-Resendiz A, Singh V, Aghamiri SS, Acencio ML, Glaab E, Ruepp A, Fobo G, Montrone C, Brauner B, Frishman G, Monraz Gómez LC, Somers J, Hoch M, Kumar Gupta S, Scheel J, Borlinghaus H, Czauderna T, Schreiber F, Montagud A, Ponce de Leon M, Funahashi A, Hiki Y, Hiroi N, Yamada TG, Dräger A, Renz A, Naveez M, Bocskei Z, Messina F, Börnigen D, Fergusson L, Conti M, Rameil M, Nakonecnij V, Vanhoefer J, Schmiester L, Wang M, Ackerman EE, Shoemaker JE, Zucker J, Oxford K, Teuton J, Kocakaya E, Summak GY, Hanspers K, Kutmon M, Coort S, Eijssen L, Ehrhart F, Rex DAB, Slenter D, Martens M, Pham N, Haw R, Jassal B, Matthews L, Orlic-Milacic M, Senff Ribeiro A, Rothfels K, Shamovsky V, Stephan R, Sevilla C, Varusai T, Ravel JM, Fraser R, Ortseifen V, Marchesi S, Gawron P, Smula E, Heirendt L, Satagopam V, Wu G, Riutta A, Golebiewski M, Owen S, Goble C, Hu X, Overall RW, Maier D, Bauch A, Gyori BM, Bachman JA, Vega C, Grouès V, Vazquez M, Porras P, Licata L, Iannuccelli M, Sacco F, Nesterova A, Yuryev A, de Waard A, Turei D, Luna A, Babur O, Soliman S, Valdeolivas A, Esteban-Medina M, Peña-Chilet M, Rian K, Helikar T, Puniya BL, Modos D, Treveil A, Olbei M, De Meulder B, Ballereau S, Dugourd A, Naldi A, Noël V, Calzone L, Sander C, Demir E, Korcsmaros T, Freeman TC, Augé F, Beckmann JS, Hasenauer J, Wolkenhauer O, Wilighagen EL, Pico AR, Evelo CT, Gillespie ME, Stein LD, Hermjakob H, D'Eustachio P, Saez-Rodriguez J, Dopazo

J, Valencia A, Kitano H, Barillot E, Auffray C, Balling R, Schneider R; COVID-19 Disease Map Community. COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. Mol Syst Biol. 2021 Oct;17(10):e10387. PMID: 34664389. [DOI: 10.15252/msb.202110387]

2. 学会発表

(1) 北野宏明, 「AI で変わる毒性学、変わらない毒性学」, 第 48 回 日本毒性学会学術年会, 日本毒性学会, 神戸(オンライン), July 7, 2021(invited)

H. 知的所有権の取得状況

１．特許取得

なし

２．実用新案登録

なし

３．その他

なし