

厚生労働行政推進調査事業費補助金（化学物質リスク研究事業）
新型毒性試験法とシステムバイオロジーとの融合による有害性予測体系の構築
(H30-化学-指定-001)

平成 31/令和元年度 分担研究報告書

分担研究課題：システム毒性解析の人工知能化

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）

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、ならびに転写制御領域の解析ソフトウェア SHOE と GarudaPlatform の連携を中心に解析パイプラインの機能強化を行なった。

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

●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

先行研究において開発した、多層（畳み込み層やプーリング層で構成される 18 層）からなる resnet モデル (Resnet は、画像分類で高い精度を示すことが多く、ImageNet2015 (<http://image-net.org/challenges/LSVRC/2015/>) における画像分類に関するコンペティションで、最も高い精度を示したモデルである (Kaming He et al (2015) Deep Residual Learning for Image Recognition. arXiv:1512.03385.)) を利用した深層学習システムをベースに、予測精度を向上させるためにトレーニング及び検証用の画像データ（遺伝子発現変動を示す 3 次元グラフ）セットを新たに作成した。この際、H30 年度は①ノイズの原因となる数値軸や文字を除去し、②3 次元グラフの描画角度

を最適化、さらに③情報量を増やすために、グレーからフルカラーに変更し且つ画素数を大幅に増加させた。H31/R01年度はさらに④10種の描画角度から3次元グラフを作成し、トレーニング画像数を10倍に増加させることで、深層学習による分類器の一層の精度向上を試みた。また構築した深層学習モデルが、画像のどの部分に着目して判別を行っているかを調べるため、最新の 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)を実装してモデルの解析を行った。

なお深層学習システムのトレーニングとチューニングした予測器の予測精度検証には、熟練した研究者が逐一検証・分類 (Positive, Negative, Rough の3分類) した3次元グラフ画像データセットを重複しないよう二分して用いた。

●Garuda Platform 上での Percellome と SHOE の連動強化

先行研究で開発を進めていた転写領域解析ソフトウェア SHOE の機能追加や改良として、特に将来的な機能拡張 (AI など) を想定した API の定義と実装を進めた。また Percellome ガジェットを含む他の Garuda Platform 用ソフトウェア (Garuda ガジェット) との連動については、実データを利用した試験運用により、プログラム改良による効果を評価した。

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

近年、大規模かつ多次元的な遺伝子発現データが蓄積したデータベースが構築されている。特に、Percellome 等の化合物毒性に関わる発現データベースは、多様な要素が絡み合った、化合物毒性のメカニズムを調査するのに有用である。しかしながら、このような複雑性の高いデータから有用な情報を引き出すことは、従来の研究で用いられてきた統計解析法等では極めて困難である。そこで、我々は、機械学習アルゴリズム、特に深層学習の手法を利用して、Percellome データベースから、有用な特徴量を認識して、有意な遺伝子を抽出するという問題を設置し、その実現可能性について検討を重ねてきた。

Percellome データベースでは、化学物質投与下の遺伝子の発現情報を画像で表す。これまで、Percellome データベースの解析では、化学物質投与下における有意に変動する遺伝子を同定するために、長年経験とトレーニングを積んできた研究者が、それぞれの遺伝子の発現パターンの画像を目で見て検討し、分類を行ってきた (図1)。

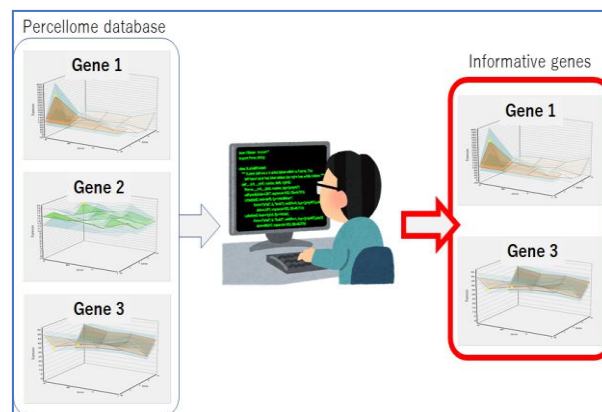


図1：専門家の Visual inspection による、有意な遺伝子の同定

しかしながら、Percellome データベースは、100 以上の化学物質それぞれに対し、40,000 以上のプローブの発現情報が含まれており、400 万以上のプローブの発現情報に対応した 400 万以上の画像データで構成されている。この様な、大量の画像の全てを目で見て、有意な遺伝子を分類するためには、多大な労力と時間が必要となってしまう。

ところで、近年、機械学習の手法の発展は目覚ましく、機械学習を用いた画像解析の手法も数多く開発されつつある。特に、深層学習の手法の発展はめざましく、分野によっては、人間の能力を超えつつある(参考文献 1)。医療分野で最も成功している例としては、深層学習の技術を使用した医療画像データ解析であり、高精度に皮膚がんの分類が出来る深層学習モデルが、発表されている(参考文献 2)。また、最近では、発見が困難であり、早期胃がんの自動検出に対しても、深層学習モデルが応用されている(参考文献 3)。

そこで、昨年度 (H30 年度)、我々は、Percellome データベースの画像データから、効率良く有意な遺伝子を判別する手法の構築を目的として、最新の多数の層で構成される深層学習の手法を用いた、画像解析システムの実装を行った。実装した深層学習に基づく分類器の予測精度の検証を、熟練した研究者が目で見えて作成した画像データセットを用いて行った。検証の結果、我々が構築した深層学習モデルは、高精度で有意な遺伝子を分類できることが分った。

しかしながら、新規データに対して適用した場合、有意であると予測された遺伝子数が非常に多く、十分な絞り込みが出来ないという問題があった。また、従来のモデルは、一つの角度からみた画像のみ

を用いていたため (図 2 A)、専門家と比較して、十分な情報を捉えていない可能性があった。

そこで、今年度は、多様な角度からみた画像 (図 2 B) を用いて、多数の深層学習モデルを作成した。

さらに、作成した多数の深層学習モデルの予測結果を統合し、有意な遺伝子を絞り込むための、computational pipeline を構築した。また、今年度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判別を行っているかについて調べるために、最新の解析手法を実装してモデルの解析を行った。

画像データセットとしては今回、10 種の角度からみた 10 種の画像データセット (図 2 B) を生成し、それぞれのデータセットを用いて、10 個の深層学習に基づく分類器を構築した。

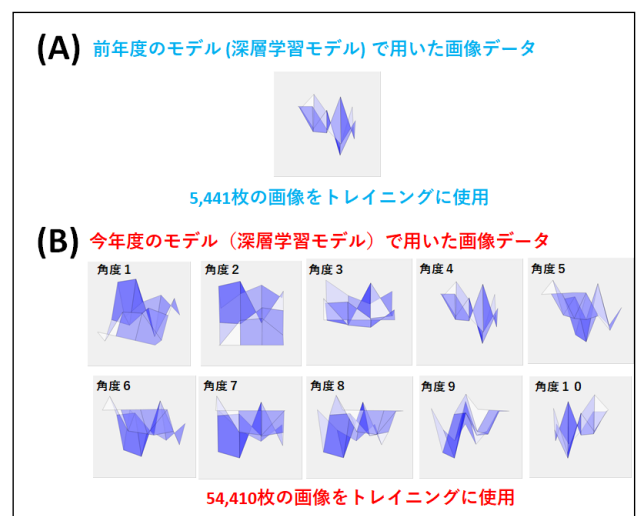


図 2. 深層学習のトレーニングに用いた、画像データ
(A) 前年度の深層学習モデルのトレーニング画像データ。500×500 ピクセルのサイズの、RGB カラー画像。positive、negative、other クラスのデータとして、4,591、377、496 プローブに対応した、5,441 枚の画像データを作成した。(B) 今年度の深層学習モデルのトレーニングデータ。500×500 ピクセルのサイズの、RGB カラー画像。それぞれの角度について、positive、negative、other クラスのデータとして、4,591、377、496 プローブに対応した、5,441 枚の画像データ (合計 54,410 個) を作成した。

深層学習に基づく、画像データから有意な遺伝子を予測する分類器

今回、深層学習モデルとしては、18層（畳み込み層やプーリング層）のレイヤーで構成される、resnetモデルを用いた(参考文献4)。Resnetは、画像分類で高い精度を示すことが報告されている(参考文献4)。今回、我々が作成した分類器は、Resnetの多数の畳み込み層を用いて、画像データから、特徴量を自動で抽出する。その後、最後の層において、抽出した特徴量を組み合わせてスコアを算出し、それぞれの遺伝子を、3つのクラスに分類する。3つのクラスは、positive class（化学物質により発現が up regulate されるクラス）、negative class（化学物質により発現が抑制されるクラス）、rough class（化学物質により up regulate も抑制もされないクラス）、である。Positive と negative class に分類された遺伝子が、有意な遺伝子と予測されるものである。

今回、上記の10種の角度からみた画像データに対する、10種の深層学習モデルを構築した。構築したモデルに対して、5 fold cross validation を実施し、予測精度の検証を行った。図3に、10種の深層学習モデルの予測精度の検証結果を示す。

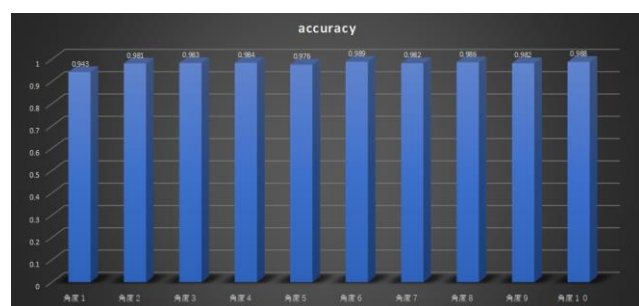


図3. 予測精度：各分類器の予測精度の比較を行った。それぞれの予測モデルの予測精度、正答率、をバーで表した。作成した10種のモデル全てが、高い予測精度（94%以上の正解率）を示した。

上図のように、作成した10種のモデル全てが、高い予測精度（94%以上の正解率）を示した。この結果から、これらのモデルを適切に組み合わせることで、高い精度を保ったまま有意な遺伝子を絞り込むことができる可能性が示唆される。

そこで、今年度、複数のモデルの予測で共通する遺伝子群に着目して遺伝子を絞り込む、computational pipelineの構築を行った(図4)。図4では、3種の角度からみた画像でトレーニングした3種の深層学習モデルの予測結果を組み合わせた、有意な遺伝子の絞り込みについて示している。それぞれの円が、それぞれのモデルで抽出された有意な遺伝子群を表しており、これらの円の重なる部分に着目することで（赤い枠線で囲まれた領域）、有意な遺伝子群を絞り込むことが出来る。現在、深層学習モデルの組み合わせのパターンや閾値など、遺伝子を絞り込むために適した条件の探索を進めている。

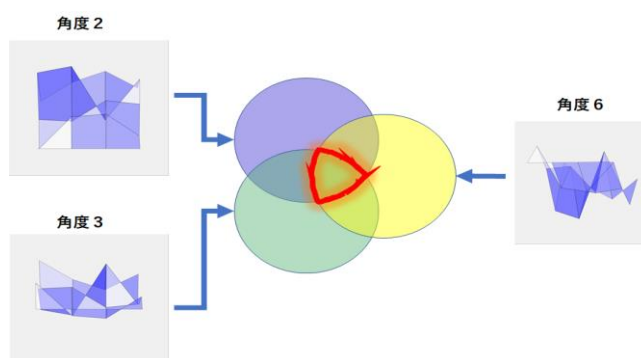


図4：複数の深層学習を組み合わせ有意な遺伝子を絞り込むための computational pipeline

ところで、最近、機械学習の予測モデルの判断基準を調べるための、explainability modelに関する研究が、注目を浴びている(参考文献5)。そこで、我々は、今年度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判別を行っているかを調べる

ため、最新の explainability model (grad cam : 参考文献6)を実装してモデルの解析を行った。その結果、図5に示すように、専門家も着目しているピークの立ち上がりの部分に着目して（黒丸内）、判別を行っていることが明らかになった。

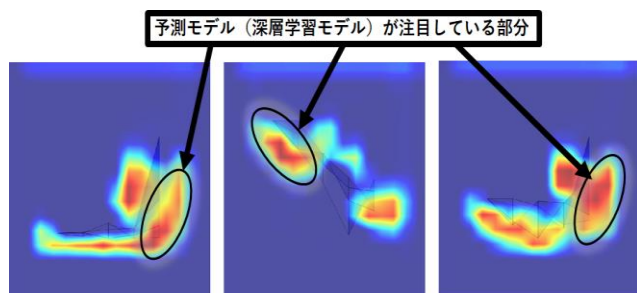


図5. Explanatory model (Grad cam) による深層学習モデルの解析の結果：深層学習モデルは、赤く強調されている所に着目して、判別を行っている。

今後は、未知のデータセットに対し、今回構築した複数の深層学習モデルを統合した computational pipeline を適用し予測を行い、予測結果を、共同研究者による visual inspection を実施し、比較する予定である。Visual inspection の結果を基に、さらに、より精緻に computational pipeline における絞り込みの条件の調整を進める。

(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.

5. Leilani H Gilpin et al (2018) Explaining Explanations: An Overview of Interpretability of Machine Learning.

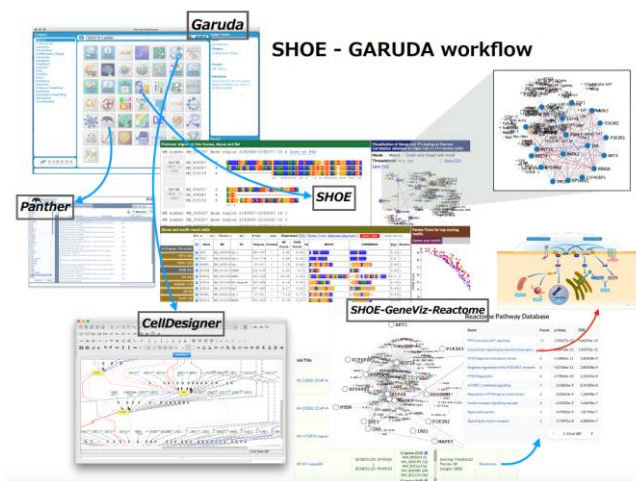
arXiv:1806.00069 [cs.AI]

6. 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

●Garuda Platform 上での Percellome と SHOE の連動強化

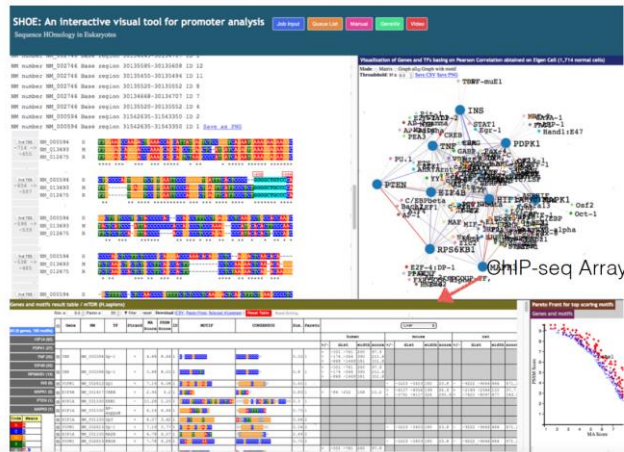
Garuda Platform 上で、Percellome DB と SHOE の連動をより密にした。

下に連動の概要を、そのあとに、一連の画面を提示する。



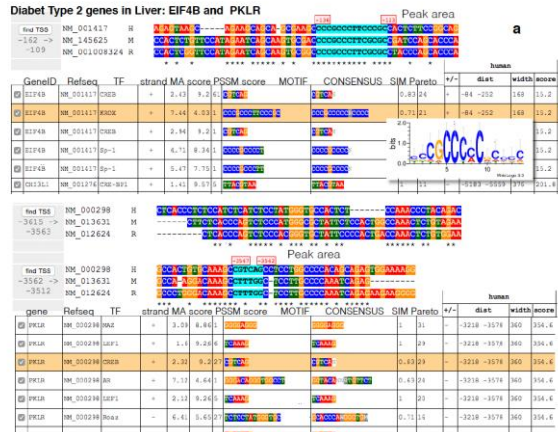
SHOE では、複数の解析ソフトウェアと連動し、効率的な解析フローを実現している。

2. Integration of ChIP-seq Array data into SHOE



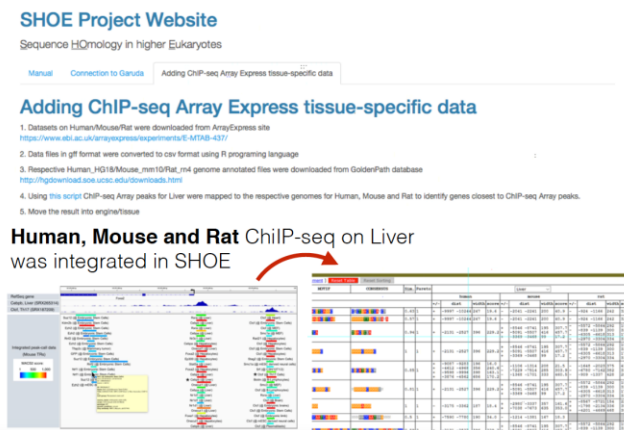
ChIP-Seq で得られたデータを SHOE に送り、マウスのゲノムでの応答を解析できるようになった。

Example 1 of ChIP-seq peak in Liver predicted by SHOE

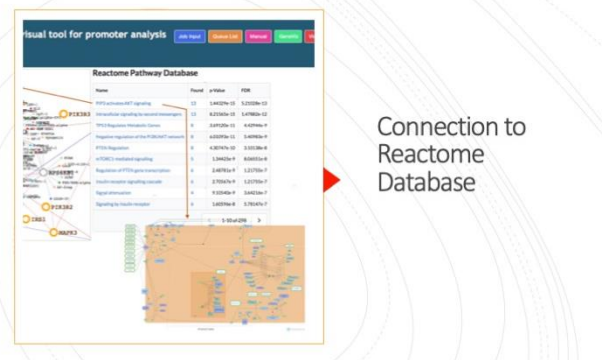


本図は、肝臓における ChIP-seq の解析を行っている例である。

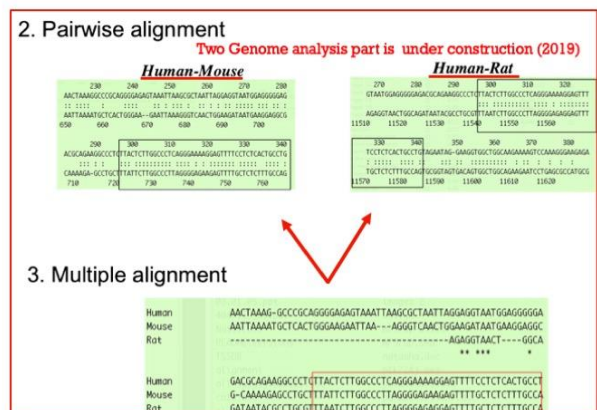
3. Integration of ChIP-seq Array data into SHOE



ヒト、マウス、ラットの ChIP-Seq データを SHOE で解析する様子である。3種のデータを照合することで正確性を強化している。



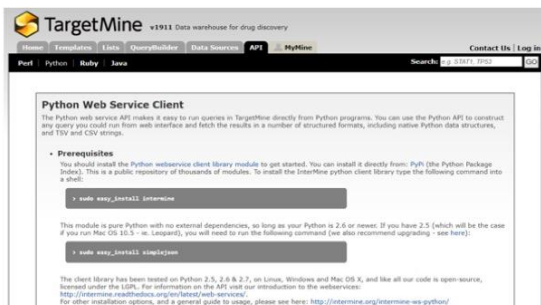
さらに、REATOM DB との照合し、解析を精度の向上を図った。



ヒト、マウス、ラットの3種間保存では条件が厳しすぎる場合に対応するための追加機能も検討した。

How to connect SHOE to the TargetMine database?

Using the TargetMine provided API Using Python Under construction (2019)



さらに、TargetMine への連動性を確保した。TargetMine は、医薬基盤研の開発するソフトウェアであり、創薬に関連する情報を提供し有用である。

What will it provide?
Gene Enrichment Analysis
Gene Pathway Heat Map
Network Interaction



その他、Gene Enrichment 解析、Gene-Pathway Heat Map などのソフトウェアとの連動も可能である。

D. 考察

深層学習は、大規模データの自動分類に有効であることがわかった。データの形状を複数の角度から学習データとして利用することで、より精緻に分類器のチューニングを行うことが可能であることが分かった。

E. 結論

今回の成果で、人手ではできなかったスケールの遺伝子発現と毒性への連動解析が可能となる。また、SHOE と Percellome の連動も密になり、よりスムーズな解析が可能となった。

今後は、この成果をさらに解析プロセスに展開を目指す。

F. 研究発表

1. 論文発表

- ① N Polouliakh. In Silico Transcription Factor Discovery via Bioinformatics Approach: Application on iPSC Reprogramming Resistant Genes. Leveraging Biomedical and Healthcare Data, 2019 – Elsevier
- ② V Gupta, A Crudu, Y Matsuoka, S Ghosh, R Rozot, X Marat, S Jäger, H Kitano, L Breton Multi-dimensional computational pipeline for large-scale deep screening of compound effect assessment: an in silico case study on ageing-related compounds. npj Syst Biol Appl 5, 42 (2019).
- ③ S Uchida, Y Asai, Y Kariya, K Tsumoto, H Hibino, M Honma, T Abe, F Nin, Y Kurata, K Furutani, H Suzuki, H Kitano, R Inoue, Y Kurachi Integrative and theoretical research on the architecture of a biological system and its disorder J Physiol Sci 69 433 2019
- ④ Y Kato, K Tabata, T Kimura, A Yachie-Kinoshita, Y Ozawa, K Yamada, J Ito, S Tachino, Y Hori, M

Matsuki, Y Matsuo, S Ghosh, H Kitano, Y Funahashi Lenvatinib plus anti-PD-1 antibody combination treatment activates CD8+ T cells through reduction of tumor-associated macrophage and activation of the interferon pathway PLOS ONE doi.org/10.1371/journal.pone.0212513

- ⑤ Polouliakh N, Horton P, Shibana K, Takata K, Ludwig V, Ghosh S, Kitano H. Sequence homology in eukaryotes (SHOE): interactive visual tool for promoter analysis. BMC Genomics. 2018 Sep 27;19(1):715. doi: 10.1186/s12864-018-5101-3.

2. 学会発表

- ① Natalia Polouliakh. Garuda/Shoe and Percellome analytic workflow. OPENTOX ASIA 2018, Asahi Seimei Otemachi Building, Tokyo, May 24, 2018.
- ② Takeshi Hase, Samik Ghosh, Ken-ichi Aisaki, Satoshi Kitajima, Jun Kanno, and Hiroaki Kitano. DTOX: Deep neural network-based computational framework to analyze omics data in Toxicology. OPENTOX ASIA 2018, Asahi Seimei Otemachi Building, Tokyo, May 25, 2018.
- ③ Ayako Yachie (2018) PREDICTIVE SYSTEMS TOXICOLOGY: CHALLENGES IN MECHANISTIC MODELING AND STRUCTURE-BASED ASSESSMENT. OPENTOX ASIA 2018 2018/5/25
- ④ Vipul Gupta (2018) GARUDA: AN INTEGRATED ANALYTICS PLATFORM TOWARDS TOXICITY ASSESSMENT AND BEYOND. OPENTOX ASIA

2018 2018/5/25

- ⑤ Samik Ghosh (2018) NEW HORIZONS IN COMPUTATIONAL TOXICOLOGY. OPENTOX ASIA 2018 2018/5/25
- ⑥ Kitano, H. AI-driven systems toxicology. Toxicogenomics for Accelerated and Refined Hazard Identification of Chemicals (Joint Symposium between IUTOX and IUPHAR), 18th World Congress of Basic and Clinical Pharmacology, Kyoto International Conference Center, Kyoto, July 6, 2018. (invited)
- ⑦ 北野宏明. AI 駆動型システム医科学の展望. 山口大学第2回シンポジウム: 人工知能・システム医学による難治性疾患への新たな挑戦, KKR 山口あさくら, 山口, Aug. 4, 2018. (invited)
- ⑧ Kitano, H. AI Grand Challenges. 2018 年日台科学技術フォーラム, リージェント台北, 台湾, Sep. 10, 2018. (invited)
- ⑨ 北野宏明. 人工知能駆動型システム毒性学の展望. CBI 学会 2018 年大会, タワーホール船堀, 東京, Oct. 9, 2018. (invited)
- ⑩ 北野宏明. AI 戦略とムーンショット. データサイエンティスト協会シンポジウム, JP タワーホール&カンファレンス東京, 東京, Oct. 19, 2018. (invited, keynote)

G. 知的財産権の出願・登録状況

なし