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

統合ツール“Percellome Integrator”の開発

研究分担者　　相﨑 健一
国立医薬品食品衛生研究所
安全性生物試験研究センター　毒性部
第一室　室長

研究要旨

本研究は、毒性の分子機序に基づいて、現行の不確実係数（安全係数）を利用する有害性評価手法を補強し、より迅速で、高精度且つ省動物を具現化した新たな有害性評価系の開発を目標として、マイクロアレイ（GeneChip)と次世代シーケンサを用いて基盤となる遺伝子発現及びエピゲノムの網羅的データを得つつ、独自開発のソフトウェア群による化学物質の生体影響の網羅的分析法の体系化を行い、これに、毒性学・分子生物学に精通したデータサイエンス専門家を擁して、システムバイオロジー及び人工知能（AI）技術を融合した新たな有害性評価系の開発を進める。

特に先行研究において、Percellome 法*を基盤とする「新型」反復曝露実験**の蓄積によりプロトタイプを構築した、化学物質の反復曝露による生体影響のデータベースについては、溶媒の反復曝露影響や、反復曝露影響の可逆性・非可逆性を遺伝子単位で取得、反映することにより、解析精度を向上させる。単回曝露のデータベースと共にこれを利用することで、現在は長い時間と多額の費用を要している長期反復曝露の毒性評価の期間短縮・効率化を検討する。

令和 4 年度は、新型化学物質有害性評価システムの解析コアの主要ユニットとしての、遺伝子発現とエピゲノムのデータ解析ユニットの開発を進めた。これに組み込むライブラリの選定を進めつつ、本ユニットの基本情報となる GeneChip プローブセット 4 万 5 千件の最新マウスゲノム(mm10)上へのマッピングを実施した。

(*)　mRNA発現値を細胞１個当たりのコピー数として絶対定量する方法。

(**) 全動物に同量の検体を反復曝露し、遺伝子発現測定直前の曝露時に、溶媒群、低用量群、中用量群、高用量群に分けて最終曝露を一回行う。実験の反復曝露と単回曝露の回数をもとに[14+1]、[4+1]、[0+1]等と表記することとした。

Ａ．研究目的

本研究は、独自構築したトキシコゲノミクス・データベース（DB）にインフォマティクス、及び、人工知能（AI）を拡大適用し、化学物質が実験動物に惹起する遺伝子発現変動等の分子毒性学情報から、科学的根拠に基づく有害性予測評価手法を確立する。これにより「安全係数」を用いる従来の有害性評価手法を補強するとともに、迅速、高精度、省動物を具現化する新たな評価システムを構築することを目的とする。

即ち、先行研究にて構築済みの延べ 8 億 5 千万遺伝子発現情報からなる高精度トキシコゲノミクスデータベースと単回曝露及び反復曝露の毒性ネットワーク解析技術を基盤に、これらを維持・拡充しつつ、さらに臓器別のゲノム DNA メチル化及び代表的物質の反復曝露によるヒストン修飾情報を加えて、毒性ネットワーク解析による、短期間試験での反復曝露毒性の予測評価技術を開発する。この際、インフォマティクス専門家によりシステムトキシコロジーや人工知能の技術を融合し、反復曝露にも対応する新型化学物質有害性評価系の実装を進める。

Ｂ．研究方法

ソフトウェアの in house 開発に際しては、先行研究で開発したソフトウェアの改良の際は開発効率と生成する実行バイナリの実行速度を重視して、Win32/64 開発及び Web アプリケーション開発は RAD（Rapid Application Development）対応の Delphi（Object Pascal 言語、USA, Embarcadero Technologies, Inc.）を用いた。データベースエンジンには組込型の DBISAM（USA, Elevate Software, Inc.）を、一般的なグラフ描画には TeeChart（Spain, Steema Software SL）を利用した。新たに開発するソフトウェアについては、ライブラリが充実している Python (ver.3.6.9)を用いた。主な解析ライブラリとしては numpy (ver.1.19.4)、pandas (ver.1.1.5)、scikit-learn (ver.0.22.2.psot1)を用いた。

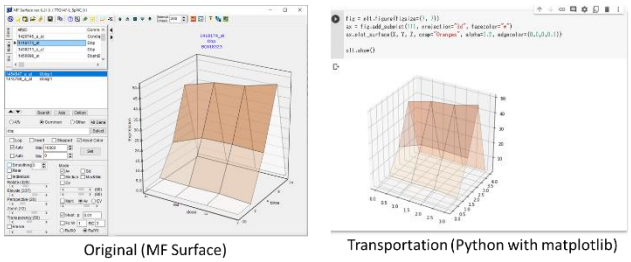
また GeneChip Mouse Genome 430 2.0 のプローブセットのマウスゲノム上へのマッピングには BWA(Li H. and Durbin R. (2009) Bioinformatics, 25:1754-60.) ver 0.7.12-r1039 を使用した。

Ｃ．研究結果

最終目標である新型化学物質有害性評価システムの解析コアの主要ユニットとしての、遺伝子発現とエピゲノムのデータ解析ユニットの開発を進めた。これに組み込むライブラリの選定を進めつつ、データ解析ユニット運用の基盤となるデータ群の結合に必要な情報(基盤情報)の整備を進めた。

遺伝子発現とエピゲノムのデータ解析ユニットの開発に際しては、Percellome 法などの独自技術の実装部分を除き、実績のある公開ライブラリの利用が望ましい。有用なライブラリは多数リリースされているが、今回の調査では、バイオデータ全般の処理が可能な biopython (https://biopython.org/)、有力なゲノムブラウザである IGV (https://software.broadinstitute.org/software/igv/)の開発グループがリリースしている組み込み用 JavaScript ライブラリ igv.js (https://github.com/igvteam/igv.js)とその Jupyter 用ラッパー ipyigv (https://github.com/QuantStack/ipyigv)を中心に、その他のバイオ系データに対応可能な可視化ライブラリとして、HiGlass (https://higlass.io/) 、 Plotly (https://plotly.com/graphing-libraries/)などの機能や動作性能(処理速度、安定性)、ライセンス、依存関係、コンフリクト状況の確認を行なった。

例えば、ヒューマンキュレーションで多用している「Surface グラフ」については、先行研究で in house 開発した MF Surface.exe の描画と同等の 3D グラフを代表的なグラフ描画ライブラリ matplotlib で高い再現性で同等機能を移植可能であることを確認した。



Original (MF Surface)　　Transportation (Python with matplotlib)

基盤情報の整備では、最重要と考えられる項目として、主要データの大半を占める遺伝子発現データを生成したマイクロアレイ GeneChip Mouse Genome 430 2.0 のプローブ 50 万件の最新のマウスリファレンスゲノム(mm10)上での座標の確定であるが、作業効率の観点から、11～25 本のプローブ群からなるプローブセット 4.5 万件の座標確定作業を進めた。各プローブセットのターゲット配列はメーカーが提供する fasta フォーマットの配列データ Mouse430_2_target より取得し、python スクリプトにより fastq 形式に変換した。mm10 へのマッピングは、BWA の mem オプションにて実施した。この結果、45101 件の target 配列のうち、複数領域にマッピングされたものが 785 件、マッピングされなかったものが 457 件あった。またマッピングされたものでも、メーカーが提供している Mouse430_2.mm10.bed の情報、具体的には染色体番号がマッピング結果と異なるものが 1320 件あった。Mouse430_2.mm10.bed で複数の領域（2～12 箇所）にマッピングされていたプローブセット 785 件と合わせ、少なくとも 2105 件（全プローブセットの 4.7%）については、個別の手動調整が必要であることが明らかになった。

個別に調査したところ、BWA によりマッピングされた染色体と、メーカー情報での染色体が異なるプローブセットの多くは、複数のエクソンを含む領域に設計されていることが確認された。

ただし、一部のプローブセットでは、元来の遺伝子座と異なる染色体上の領域をターゲットに設計されていることも判明しており、解析精度を高めるためには丁寧な確認作業が必要であった。

D．考察

開発準備はほぼ計画通りに進行している。新型化学物質有害性評価システムの解析コア開発のための開発用ライブラリ探索は順調で、必要な機能の大半は発見済みである。検討課題としては、データの可視化に際して、従来の様式より効果的な表現の検討が挙げられる。特にグラフや解析図等の可視化手法については、分担研究「システムバイオロジーによる毒性解析の AI 化」での学習用画像としての利用を前提として、特徴を捉えやすい表現形式となるよう考慮している。

GeneChip Mouse Genome 430 2.0 のプローブセットのマウスゲノム mm10 上へのマッピングについては、メーカーから提供されている座標情報との乖離が少数ながら確認され、逐一、整合処理を行った。このマッピング情報はデータ群の結合に必要であるだけでなく、GeneChip の遺伝子発現データと次世代シーケンサを用いた RNA-Seq の遺伝子発現データとの相互データ変換の精度向上にも寄与する重要な基盤情報となるため、整合処理は慎重かつ正確に進めた。

E．結論

本分担研究は、ほぼ計画通りに進捗した。

必要な機能に対応した主なライブラリについての情報収集を終えた。また基盤となる情報、とりわけ GeneChip Mouse Genome 430 2.0 のプローブセットの target 座標情報について整備した。来年度からの本格的な開発への前処理は整ったと考えられる。

F. 研究発表

１．論文発表

(1) Yuhji Taquahashi, Shuji Tsuruoka, Koichi Morita, Masaki Tsuji, Kousuke Suga, Ken-ich Aisaki, Satoshi Kitajima, A novel high-purity carbon-nanotube yarn electrode used to obtain biopotential measurements in small animals: flexible, wearable, less invasive, and gel-free operation. Fundam. Toxicol. Sci. 2022; 9: 17-21.[doi.org/10.2131/fts.9.17]

2．学会発表

(1) J. Kanno, K.-I. Aisaki, R. Ono, S. Kitajima、Analysis of Murine Liver mRNA Expression, DNA Methylation, And Histone After Repeated Exposure To Chemicals. EUROTOX 2021 virtual congress 、(2021.9.29)、Oral

(2) 菅野純、北嶋聡、相﨑健一、齊藤洋克、種村健太郎、肺の遺伝子発現応答と毒性機序予測解析. 第48回日本毒性学会学術年会、(2021.7.9)、神戸国際会議場、シンポジウム、口演

(3) 菅野純、高木篤也、相﨑健一、北嶋聡、異物発癌に関わるトランスクリプトミクス特性. 第48回日本毒性学会学術年会、(2021.7.8)、神戸国際会議場、シンポジウム、口演

(4) 相﨑健一、小野竜一、菅野純、北嶋聡、トランスクリプトミクスから見た発癌物質の特性. 第48回日本毒性学会学術年会、(2021.7.8)、神戸国際会議場、シンポジウム、口演

(5) 菅野純、相﨑健一、小野竜一、北嶋聡、毒性OmicsとAIによる慢性毒性予測. 第48回日本毒性学会学術年会、(2021.7.7)、神戸国際会議場、シンポジウム、口演

(6) 夏目やよい、相﨑健一、北嶋聡、Samik GHOSH、北野宏明、水口賢司、菅野純：PPARαリガンドの比較毒性オミクス. 第48回日本毒性学会学術年会、(2021.7.7)、神戸国際会議場、シンポジウム、口演

(7) J. KANNO, K. AISAKI, R. ONO, S. KITAJIMA , Comprehensive Histone, DNA Methylation and mRNA Expression Analysis of Murine Liver Repeatedly Exposed to Chemicals. CTDC11, (2021.6.15), Virtual, Oral

(8) 菅野純、外来性化学物質(xenobiotics)により誘発される生体反応の分子機構解析と創薬加速、第3回医薬品毒性機序研究会、(2021.1.15)、online meeting、口演

(9) Jun Kanno, Ken-ichi Aisaki, Ryuichi Ono and Satoshi Kitajima 、 Application of PERCELLOME database as a part of big data to toxicological research: The 36th Annual Meeting of KSOT/KEMS, Special lecture, Web, Oral presentation.

G．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