【報告様式Ａ】

作成／更新日：令和４年５月30日
課題管理番号：21am0101067j0005

# 日本医療研究開発機構　創薬等ライフサイエンス研究支援基盤事業
# 事後評価報告書

公開

## Ⅰ 基本情報

補助事業課題名：　（日本語）創薬等ライフサイエンス研究支援基盤事業
（プログラム名）　（英　語）Platform Project for Supporting Drug Discovery and Life Science Research

実施期間：平成29年４月１日～令和４年３月31日

補助事業担当者　氏名：（日本語）木下　賢吾
（英　語）Kengo Kinoshita

補助事業担当者　所属機関・部署・役職：
（日本語）国立大学法人東北大学・大学院情報科学研究科・教授
（英　語）Professor・Graduate School of Information Sciences・Tohoku University

## Ⅱ 補助事業の概要

近年、個人のゲノム解析が数千～数万人規模で行われるようになり、タンパク質の配列・構造・機能に影響し得るゲノム上の一塩基変異・挿入・欠失が多数報告されている。臨床現場においても精神神経疾患や自己免疫性疾患などの診断と治療のため患者個人の全ゲノム・エクソームシークエンスが普及し、がんゲノム医療が保険収載されるなど、ゲノム情報に基づく治療法の選択に貢献することが期待される時代となっている。そこで本事業では、ゲノム情報、発現量情報及びタンパク質情報をつなぎ、日本における個別化創薬の推進を目指した支援と高度化を行った。

(1)支援

DBの継承と運用　①ゲノムとタンパク質関連DB

本機関で開発を行った天然リガンドデータベース(NLDB)及び早稲田大学などと連携して開発したVaProSを核として、データベースの運用を行った。また、関連するデータベースの調査を行い、有用なデータベースがあれば、リンクを張るなど関連性が分かるように整理すると共に、講習会等を通じてデータベースの利用・運用の支援を行った。さらに、これまで開発を行ってきたデータベースを利用して「データベースがカバーしきれていない部分の可視化」を行い実験系の研究者に提示するために、KEGG Reactionを元に構築されているNLDBにて「KEGG代謝パスウェイに対するNLDBでのタンパク質―天然リガンド構造カバー率」を新たに算出した。今後、新規構造の増加とともにカバー率の変化を観察することで、構造生物学へ新しい示唆を与えられると共に、創薬ターゲットの選定にも寄与することができたと考えている。

DBの継承と運用　②タンパク質発現関連DB

当機関で開発を行ってきた、遺伝子共発現データベースとしてCOXPRESdbはVaProSの要素データベースとしても提供してきたが、開発から時間も経過していることもあり、サーバが老朽化している。そこで、安定したデータベー

ス提供のために、サーバの更新とサーバのプログラム群の整理を行った。また、利便性向上のために、データ提供のための各種 API を整理し、適宜講習会等で支援のニーズを探りつつ運用を行った。また、本課題に関連して 2 件の個別支援を受けたので、対応を進めながら支援側のニーズを探りつつ、DB の改良によるサービスの向上へとつなげた。

COXPRESdb に関しては、システムの再構築も行い、最新のサーバにてサービスを提供することで、安定した高速利用環境を提供した。さらに、データ利用のためのツールの高度化や、API、RDF によるデータ連携を大幅に強化し、遺伝子共発現データベースとしての利用価値を飛躍的に向上させた。2018 年 9 月に新システム上で稼働する COXPRESdb version 7.0 をリリースし、2021 年度には version 8.0 をリリースした。最新のリリースでは、提供する遺伝子共発現データについて、11 生物種 34 プラットフォームのすべての共発現データを更新した。

③講習会・個別支援

個別支援として8件の支援を実施した。主な支援は以下の通りである。

課題名「カーボンナノチューブ受容体の動的構造解析」については、カーボンナノチューブ(CNT)の毒性分子機構解明のため、CNT とその受容体であるマクロファージレセプターTim4 との結合様式を、分子動力学(MD)シミュレーションを用いたドッキングシミュレーションにより予測した。予測された箇所に変異を導入した Tim4 による生化学実験を行ったところ、結合活性が有意に下がることを確認できた。この変異型 Tim4 を導入したマウス培養細胞を作成したところ、野生型 Tim4 を導入した細胞に比べて CNT との結合活性が低下することも観察された。この支援課題を通じて、信頼に足る複合体予測構造から CNT の毒性分子機構が明らかになることによって、今後の CNT による公害の予防や炎症の治療法の開発につながることが期待される。(Omori et al, Cell Rep, 2021)

課題名「血清総 IgE 値の GWAS により新規発見された SNP の影響予測について」については、血清総 IgE 値についてのゲノムワイド関連解析(GWAS)の結果、関連が見られたアミノ酸変異についてタンパク質立体構造を用いた解釈を行った。今回の解析で IL-4Rα の日本人で比較的頻度が高い Ala82Thr 変異が IL4 濃度を下げることが示された。この GWAS では約1万人のゲノムデータを用いた解析で見つかった変異の有意性がゲノムワイドな有意水準にギリギリ達する状態であったが、構造をベースとした解釈を行うことで、偽陽性である可能性を排除しつつ関連変異として結論をつけることができた点で支援の貢献が果たした役割は大きかった(Shido et al, J. Invest Derm. 2021)。

(2)高度化: ① ゲノム変異のタンパク質アノテーションツールの開発

タンパク質立体構造情報を活用した変異情報の解釈を進めるための手法として、ゲノム解析結果の VCF ファイルへのアノテーション手法を開発した。具体的には、構造情報の有無、天然変性領域の実験・予測による情報、二次構造、埋もれ度、タンパク質・リガンドとの相互作用面、リン酸化などの修飾情報など、タンパク質科学の分野では重要ながら、ゲノム情報とリンクしてあまり利用されてこなかった情報をゲノム解析研究者が利用しやすい形で、最新の PDB 情報に基づいて利用できる環境の基盤を構築した。これにより、既存の変異データにおいても構造上重要な変異を新しく発見することができると期待される。この手法の応用として東北メディカル・メガバンク機構で行われた日本人約 14000 人の全ゲノム解析の結果について構造アノテーションを行い、この情報は現在同機構のサイト jMorp (https://jmorp.megabank.tohoku.ac.jp/) で公開している。さらに、これらのバリアント解析を行えるウェブツールも作成し公開も行った(https://wupsivs.sb.ecei.tohoku.ac.jp/ )。これは、利用者がゲノムバリアントのリストをテキストボックスに入力・あるいはファイルとしてアップロードすることで、タンパク質構造に基づくアノテーションを取得できる。結果ページはバリアントごとに、対応する PDB のアミノ酸の構造的特徴のサマリを提供し、各バリアントページへのリンクを提供している。各バリアントのページはそれに対応する全ての PDB のアミノ酸残基の情報を表示し、また MolMil 構造ビューワによって残基の位置や構造を確認することができる。これらのツールを応用し、疾患関連バリアントの機能評価を効率化するために ClinVar、COSMIC 等の疾患関連データベースのバリアントやゲノム・エクソーム解析で同定された疾患患者のバリアントについても速やかに情報を更新している。

②遺伝子強共発現データベース:COXPRESdb では大量の共発現情報を提供しているが、オミクス解析の下流解析など、多くの遺伝子に着目する状況において、研究に重要な関係を選択するのは容易なことではない。

COXPRESdb version 8.0 では、提示されている共発現関係に関する先行研究を提示する機能 CoexPub を実装した。この機能はヒトとマウスに関する文献のフルテキストデータから、遺伝子ペアが共起している文章を一文単位で検出・抽出し解析することで、文献による遺伝子共発現の支持度合いを計算するものである。COXPRESdb version 8.0 の共発現遺伝子リストにおいて、CoexPub によって検出された関連論文数を表示するとともに、CoexPub の詳細ページでは、共発現関係の吟味に重要な文章を優先順位付きで表示するシステムとなっている。これにより、網羅性が高いが不確実である遺伝子共発現情報と、網羅性は低いが確実な知見が組み合わさり、知識抽出の効率が大きく向上した。

(3)連携

ユニット間連携として、最適化ユニットの各担当者が手分けして他のユニットの情報収集と整理を実施した。また、他ユニットのシンポジウムなどで、プラットフォーム機能・最適化ユニットの研究支援一括窓口の紹介をするとともに、ウェット系研究者によるデータサイエンス研究の成果の利用を促進するため、我々が開発を行ってきた天然リガンドデータベース(NLDB)や遺伝子共発現 DB(COXPRESdb)や jMorp 等を含めたデータベース利用の実践的な紹介を行った。

コロナによる学会がオンライン中心になる以前は、学会やシンポジウムでの連携加速のための試みとしてブースでの説明サポートや発表を行った。例えば、BINDS 公開シンポジウムでは「ゲノム変異・遺伝子発現量・蛋白質情報をつなぐ試み」として、ConBio では「低頻度一塩基多型のタンパク質立体構造を利用した機能アノテーションに向けて」と題して我々の試みを発表した。これらの講演では、我々がこれまで行ってきたヒトゲノム中一塩基変異のタンパク質構造からの解釈、特にタンパク質間相互作用部位やリガンド結合部位などの機能部位と変異の位置やアリル頻度の関係についての解析結果を、具体的な事例も交えて紹介するとともに、集団ごとのアレル頻度に着目した場合や、異なる集団間でのアレル頻度の差についての解析結果も紹介を行った。学会がオンライン中心になってからは講演の中で BINDS の紹介や DB の紹介を入れるなど連携を促す工夫を行ってきた。最終年度には、BINDS の化合物ライブラリーの全体像を明らかに効率的なスクリーニング推進に向けて、筑波大学とも連携し化合物ライブラリーの統合 DB の構築も行った。

他のユニットの高度化にも寄与する連携も進めてきた。例えば、産業技術総合研究所人工知能研究センターによって運用されている、アミノ酸配列の類似性検索システム、FORTE に利用可能な形式に調整し、公開サーバー上での環境の整備を行った。これは、昨今の利用可能な配列情報の増加によりアミノ酸配列の類似性検索にかかる時間的コストが問題となっているが、それを解決するために重要な貢献であると考えている。また、ケミカルシード・リード探索ユニットの山本雅之と連携して、クライオ電顕の東北大への導入も実現した。この電顕は AMED のスパコンとも連携することで高スループットでの解析を実現できる環境を整える予定であり、データの保存にも十分な領域を確保しており、今後の創薬研究の核となることを目指している。事業間連携としては、jMorp に関しては、AMED の事業である東北メディカル・メガバンク計画との連携として実施しており、日本人におけるゲノムデータの創薬事業への活用として今後も力を入れていきたい。最終年度には、14000 人の全ゲノムデータに対して構造情報の付加を行うことで貢献した。

In recent years, analysis of individual genomes has been carried out on a scale of several thousand to several tens of thousands of individuals, and many single nucleotide mutations, insertions, and deletions in the genome that can affect the sequence, structure, and function of proteins have been reported. In the clinical field, the whole genome exome sequence of the individual patient is spread for diagnosis and treatment of neuropsychiatric diseases and autoimmune diseases, and the cancer genome medical treatment is registered in the insurance, and it is expected to contribute to the selection of the therapy based on the genome information. In this project, genome information, expression level information, and protein information were connected, and supports and developments were carried out with the aim of promoting personalized drug discovery in Japan.

As a support, the natural ligand database (NLDB) developed by this institute and VaProS developed in cooperation with Waseda University were used as a core to operate the database. Relevant databases were also investigated, and if there are any useful databases, links were created to identify their relevance. At the same time, support for the use and operation of databases was provided through seminars, etc. Furthermore, to visualize the areas not covered by the database using the database developed so far and present it to researchers in the experimental system, we newly calculated the "protein-natural ligand structure coverage in NLDB for the KEGG metabolic pathway" in the NLDB constructed based on the KEGG Reaction. In the future, by observing changes in the coverage ratio as new structures increase, we were able to provide new implications for structural biology and contribute to the selection of drug discovery targets.

In addition, we supported 8 research activities, some of which are described below.

To elucidate the molecular mechanism of toxicity of carbon nanotube (CNT), the binding mode of CNT and its receptor, macrophage receptor Tim4, was predicted by docking simulation using molecular dynamics (MD) simulation. Biochemical experiments with Tim4, in which a mutation was introduced at the predicted position, confirmed that the binding activity significantly decreased. When cultured mouse cells transfected with this mutant form of Tim4 were produced, the binding activity to CNT was also observed to be lower than that in cells transfected with wild-type Tim4. Through this support project, it is expected that the molecular mechanism of toxicity of CNT will be clarified from the reliable complex prediction structure, which will lead to the prevention of pollution by CNT and the development of the therapy of inflammation in the future. (Omori et al., Cell Rep. 2021). The subject "Prediction of the effect of SNP newly discovered by GWAS on serum total IgE levels" was interpreted using protein conformation for amino acid mutations associated with serum total IgE levels as a result of genome-wide association analysis (GWAS). In the present analysis, it was shown that Ala 82 Thr mutation with comparatively high frequency in Japanese of IL -4 Rα lowered the IL 4 concentration. In this GWAS, the significance of the mutation found in the analysis of the genome data of about 10,000 people barely reached the level of genome-wide significance, but the contribution of the support was significant in that it was able to exclude the possibility of false-positive by the interpretation based on the structure and conclude that it was related mutation (Shido et al., J. Invest Derm. 2021).

As a development, we have constructed an environment that enables genome analysis researchers to use information that has not been widely used in the field of protein science, although it is important in the field of protein science, such as the presence or absence of structural information, information from experiments and predictions of natural denatured regions, secondary structure, degree of burial, interaction with proteins and ligands, and modification information such as phosphorylation, based on the latest PDB information. This is expected to allow the discovery of new structurally important mutations in existing mutation data. As an application of this method, structural annotation was performed on the results of whole genome analysis of about 14000 Japanese conducted by Tohoku Medical Megabank Organization, and this information is now available at jMorp (https://jmorp.megabank.tohoku.ac.jp/). He also created and published a web tool to perform these variant analyses (https://wupsivs.sb.ecei.tohoku.ac.jp/). This allows users to get protein structure-based annotations by typing a list of genome variants into a text box or by uploading the list as a file. For each variant, the results page provides a summary of the structural features of the corresponding PDB amino acids and provides links to each variant page. The page for each variant displays information on the amino acid residues of all corresponding PDBs, and the MolMil Structure Viewer allows you to view the positions and structures of the residues. These tools are being applied to rapidly update information on variants in disease-related databases, such as ClinVar and COSMIC, as well as variants in patients identified by genomic exome analysis, in order to streamline functional assessment of disease-related variants.