

## 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



### 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を促進する研究支援とデータサイエンス  
（プログラム名）（英語）Support for BINDS program of life sciences research with drug discovery  
and promotion of data 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）田之倉 優  
（英語）Masaru Tanokura

補助事業担当者 所属機関・部署・役職：  
（日本語）東京大学・大学院農学生命科学研究科・特任教授  
（英語）Project Professor, Graduate School of Agricultural and Life Sciences,  
The University of Tokyo

### II 補助事業の概要

国立研究開発法人日本医療研究開発機構（AMED）では、平成29年度から5か年計画の創薬等ライフサイエンス研究支援基盤事業「創薬等先端技術支援基盤プラットフォーム」（BINDS）を起ち上げた。東京大学の田之倉は平成14年度から始まった「タンパク3000プロジェクト」において文部科学省の委託で事務局を引き受けて以来、その時々での事業の主旨や内容の変化に柔軟に対応しつつ事業支援に徹し、アカデミアを中心とする公的プロジェクト運営に寄与し、内外から高い評価を得てきた。豊富な知識・経験・ノウハウ・人材等を活かして同プロジェクトの下で「創薬等ライフサイエンス研究を促進する研究支援とデータサイエンス」に取り組むこととなった。

本課題は、以下のとおり、同プロジェクトの円滑な運営と事業の発展を目的とした実務的な支援業務とデータサイエンスに関する研究の2つのプログラムに大別される。

実務的な支援業務では次のことを行った。

#### ① 運営会議への連携協力

支援業務を担う「支援オフィス」はBINDSの運営会議/事務局との間に緊密な協力関係を構築し、運営会議等の意向や指針に沿って主にプラットフォームにおけるワンストップサービス事業、広報・イベント活動、連携事業の支援を展開してきた。一方現場で得られる最新の分析資料や情報の提供、広報やイベント企画の提案を行い、PDCAサイクルの確立、BINDS事業の円滑な運営と事業の発展に寄与した。

## ② プラットフォームへの実務支援

Webを利用してコンサルティング・支援申請を一体化して受け付けるワンストップサービスの窓口業務を正確かつ迅速に一連の処理を滞りなく実施した。広報活動の強化を反映して、事業最終年度である年度末現在、支援承認総数は2742件に達した。初期段階の問い合わせ及び申請受付だけでなく、事前相談から論文の掲載等のフォローアップに至るまで一貫したサービス全体の向上に努めた。

放射光施設のビームタイム配分調整業務の一翼を担い、利用申請者に対するIDの発行、マシンタイムの通知等を行うことにより構造解析の研究推進に資した。

## ③ 広報活動支援と事業成果の社会還元

事業成果等の社会還元、ワンストップサービス事業の推進のため、さまざまな広報資料の発行・配布やホームページの運用、各種イベント（シンポジウム、セミナー、ワークショップ、講習会等）の広報と開催の企画・支援を行った。ホームページによる広報については、積極的かつ迅速なニュース配信等によりその充実を図ったほか、公的な機関や学会等の外部のホームページと連携し事業のPRも行った。また、外部に対して事業の普及・成果紹介を目的として開催されているBINDS公開シンポジウム、国内主要学会や大学等でのセミナー等の多くの研究イベント開催について協力・支援を実施した。また事業成果集の発行・全国への配布に協力した。

## ④ 事業内の領域の連携強化

ユニット/領域間内の相互理解の醸成、連携の強化と研究交流の促進のために毎年度開催されるBINDS（成果）報告会に対し協力・支援を行ったほか、東京大学など大学が主催する講習会に対して企画、ノウハウの提供、広報・参加登録等積極的に協力・支援を実施した。

Our program has two missions: practical support for the smooth operation and development of the project, and research on data science.

The practical support is as follows;

- 1) We built a close cooperative relationship with the BINDS Steering Committee/Secretariat, support the following activities, and provide useful analytical information to the headquarter. we contributed to the smooth operation and the development of the BINDS project by establishing the PDCA cycle.
- 2) We carried out the work of one-stop consulting service using the Web accurately and quickly without delay. Eventually, the total number of support approvals reached about 2,700. We also endeavored to consistently improve the overall service, from prior consultation to follow-up such as publication of papers.
- 3) In order to contribute to the academic society by returning the research results obtained from the project and to promote participation in the project, we publish and distribute various public relations materials, operate the website, plan and publicize the holding of research events such as the BINDS symposium and seminars at major academic societies and universities.
- 4) We cooperated with the BINDS Project Research Results Report Meeting held every year to foster mutual understanding among BINDS participants, strengthen cooperation between each area, and promote research exchanges. We actively provided support for the workshops and course held in each unit of the project and many universities.

データサイエンスに関する研究では次のことを行った。

文献データベース PubMed に登録されているすべての文献のアブストラクトより生命現象や生体分子に関する用語間の関連性を数値化（距離化）し、それらをつなぎあわせることにより新たに関連性を推定するシステムを開発した。それを Relation Inferring System by Text Mining (RISTM、リステム) と命名した。本システムは、計算アルゴリズムの検討を含め、手法的、手順的な検討を多数行いながら開発したものである。その最終形は以下の通り。

- ① 生命現象の百科事典ともいえる KEGG データベースより、すべてのエントリーを取得し NAME フィールド等から生命現象・生体分子に関する検索語を作成した。
- ② これらの検索語から KEGG コードへの逆弾き辞書を作成した。これにより、約 10 万語の KEGG コードに紐づけされた約 23 万語の検索語が作成された。
- ③ すべての PubMed アブストラクト（PubMed に登録されているエントリーのうちアブストラクトが収録されているもの、約 2 千 200 万件）にこれらの検索語が含まれるかを調べた。これはアブストラクト本文ならびに上記検索語に対し独自開発のステマーを通してから行なった。
- ④ すべての検索語の組合せ（二項関係）について PubMed アブストラクト内での共起関係を導き出し、それらを集計し、検索語間の関連性の大小を数値化した。これを KEGG コード間の関連性に変換した。
- ⑤ この関連性をつなぎ合わせるにより新規の関連性が推定できるようにした。
- ⑥ PubMed、KEGG 両データベースのアップデートに応じてシステムが自動的にアップデートされるような仕組みを作成し、組込んだ（PubMed については毎日、KEGG については毎週更新）。

この最終的なシステム構造を導き出すために、検索語の抽出方法の開発と検討、このシステムに最適な独自のステマーの開発と検討、関連性の数値化（距離化）に最適な計算式の検討、関連性をつなぎ合わせの方法とその計算アルゴリズムの開発・検討、アップデート手法の検討など、すべて一つ一つ計算結果を見ながら検討し、組み立てた。

さらに、このシステムを用いて、疾病からその関連薬を距離の近い順にリストアップするプログラムを開発、それを COVID-19 に応用し、RISTM の原理とともに論文として発表した。また、このシステムを使って「支援」も開始した。

As the research on data science, we have developed a system to infer the relationship, by text mining, between terms of biological phenomena/biomolecules (Relation Inferring System by Text Mining; RISTM). The RISTM system is based on the co-occurrence of two terms, for each biological phenomenon or biomolecule, within the same abstract. Each relationship between terms is summarized for all PubMed abstracts and converted into relationship between KEGG (Kyoto Encyclopedia of Genes and Genomes) codes to which each term is linked. By connecting these relationships, a new relationship can be inferred.

To build up this system, we devised a method to get search terms from all KEGG entries, created a special stemmer for this system, devised a calculation formula to quantify the relationship as a distance from co-occurrence statistics, and devised a method to connect relationships, and made update programs of the distance data.

Then we identified candidates COVID-19 drugs by this system (MURAMATSU, T. and TANOKURA M., *Bioinform. Adv.* **1**, vbab013, doi:10.1093/bioadv/vbaab013).