【報告様式Ａ】

作成／更新日：令和４年５月31日
課題管理番号：21am0101111j0005

# 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）生薬データベースの高度化と構造創薬への応用
（プログラム名）　（英　語）Sophistication of crude drug database and application to structural drug discovery

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）金谷重彦
（英　語）Shigehiko Kanaya

補助事業担当者　所属機関・部署・役職：
（日本語）奈良先端科学技術大学院大学・先端科学技術研究科・教授
（英　語）Nara Institute of Science and Technology, Graduate School of Science and Technology, Professor

## Ⅱ 補助事業の概要

和文

本課題では，申請者の開発した世界唯一の生薬データベースである KNApSAcK を高度化し，創薬応用，特に構造ベース創薬（SBDD）に活用するための情報基盤の構築を行う。具体的には，ヒト代謝・遺伝子制御マップを中心に据えて、i)生薬データベースの拡充・高度化、ii)生薬・承認薬・代謝物の立体構造の網羅的比較・分類、iii)生薬のヒト代謝・遺伝子制御マップへの関連付けと新規医薬品骨格の提案、iv)天然変性タンパク質データベース IDEAL へのヒト代謝・遺伝子制御アノテーション、v)生薬-タンパク質複合体モデル構築と新規創薬ターゲットの提案の５項目を柱に高度化研究を行い、これら情報資源を活用して外部創薬研究者の支援を行うことを目的とする.

### 1）天然化合物データの拡充、新規創薬ターゲットの提案、リガンド結合予測

新規ドラッグ骨格の提案やヒト代謝物と天然物および承認薬の比較による共通ドラッグデザインの解析に利用する目的で，KNApSAcK 収録化合物(生理活性天然物分子，3,055 分子)， 承認薬(DrugBank，1,985 分子)，ヒト代謝物(KEGG,1,396 分子)， 毒物(517 分子)立体構造データベースを構築し,COMPLIG を用いたグラフマッチによる化合物の立体構造クラスタリングによる構造分類を行った．これらの分子データは KNApSAcK を中心に，ユニット間連携により大村ライブラリー， DTX(Drug Target Excavator)から収録され， 2,017 の構造クラスターに分

類された．引き続きこのデータについて 医薬情報のアノーテーションを実施した．KNApSAcK で付与された生理活性（Activity term）と国際疾病分類第 11 版（ICD-11）および解剖治療化学分類（ATC）との単語比較によるマッチングを行った．結果として，生薬成分の分子 3,209 個に対し 2,242 個を ICD-11 大分類に， 3,414 個を ATC に対応づけることができた．KNApSAcK の 156 個の Activity terms のうち，44 が ICD-11（大分類）に，63 が ATC コードに関連付けられ， 生薬成分 3,209 分子に対し 2,242 個を ICD-11 大分類に，3,414 個を ATC に対応づけすることができた．

また、KNApSAcK 収録天然物分子，承認薬，ヒト代謝物のクラスタリングの解析を進め，天然物と承認薬の構造比較による新規ドラッグ骨格の抽出，ヒト代謝物と天然物および承認薬の比較による共通ドラッグデザインの具体的な方法論を確立し，アプリケーションやデータベースとして実装した．

本DBでは、57, 906種の天然有機化合物を文献情報から入力した. また，これらの天然物情報をもとにしたインシリコスクリーニングを行い論文にまとめた(Mol. Inf., 2022; Sci Rep, 2022)．

さらに，生薬成分のターゲットタンパク質の推定と複合体構造モデル構築を目的として，COMPLIG を用いた生薬成分と PDB リガンドのマッチングを行い，1,629 個の分子を PDB リガンドに紐づけした．このアノーテーションにより，生薬成分-ターゲットタンパク質の複合体モデルのテンプレートを PDB 中で探索することが可能になったので，現在モデル構築を進めている．また，分子データが高度に蓄積されたことを受けて，機械学習による天然物･承認薬･ヒト代謝物･毒物分子の特徴抽出を行い予備的な結果を報告した(***蛋白質科学会*** 2021). また，AI 創薬の基盤として，深層学習の一つである，分子における原子の結合関係と原子の種類によるグラフ表現に基づいた分子グラフ・コンボリューション・ニューラルネットによる活性の有無によるインシリコスクリーニング法を確立した(***Mol Inf***, 2020; ***BMC Binformatics*** 2019).

COVID-19 パンデミックに対応するために，生理活性天然物･承認薬･ヒト代謝物･毒物分子構造データを利用することで SARS―CoV-2 タンパク質の構造モデリングを行い， main protease, S-glycoprotein, 2'-*O*-ribose methyltransferase と生理活性天然物･承認薬との複合体モデルを提唱した．この結果は， BINDS ホームページ http://harrier.nagahama-i-bio.ac.jp /dtx/SARS-CoV-2/, および BSM-Arc (https://bsma.pdbj.org/entry/ 15)から一般に公開するとともに論文報告した(***FEBS Lett***. 2020, ***生物物理*** 2021).

### 2）天然変性タンパク質 DB の拡充：代謝と転写の関連付け

細胞は外部からの刺激に応答し，転写制御を行うことでタンパク質の発現を調節している．シグナル伝達，転写制御には天然変性タンパク質の密接な関与が考えられるので， IDEAL にリン酸化を起点とした生命現象イベントを付加する方法を検討した．リン酸化から転写に至る過程の生命現象イベントのアノテーション結果をエクセルファイルとしてまとめ，説明文をつけてホームページより公開した．これで IDEAL エントリによるリン酸化ネットワークが表現されたので，代謝と転写の観点から KNApSAcK 生薬成分との関係性表現に取り組んだ．KNApSAcK における生薬情報を創薬に利用するために，KNApSAcK に登録されているリガンド（代謝）と IDEAL に登録されているタンパク質（転写）を KEGG の情報を介して連結させた．この結果を最適化ユニットで開発している DTX に提供し，KNApSAcK リガンドと IDEAL との接続を実行した．

### 3）KNApSAcK と IDEAL の登録，更新，公開

KNApSAcK DB においては，天然物化合物情報の充実を図る目的から，コロナ禍であってもデータを蓄積する手続きを確立し，データベースの天然物登録数についても 5,000 件増え，現在，56,829 件となっており，さまざまなインシリコスクリーニングに活用できる体制を構築した．現在、天然物化合物とそれらを生産する生物種の関係では 135,156 件となっており，生物種としての生薬の可能性を示すことができている。また，年間アクセス数は 100 万件を超えており世界の天然物・生薬研究の基盤としての役割を担っている．

IDEAL では，天然変性タンパク質のアノーテーション，および既存データのアップデートを予定通りに実施した．現在 995 エントリーについて 1 万を超える天然変性領域情報が提供されている．ライフサイエンス統合データベースセンターと共同で IDEAL の全データを RDF 化して公開し，SPARQL の EndPoint を構築した．データ共有が可能となったため UniProt とも相互リンクされるようになった．欧州でのデータ共有活動，ELIXIR に参画し，DisProt などと天然変性タンパク質のデータ統合についての連携を開始した．UniProt のアノテーションに基づき，液滴と関連する天然変性タンパク質に“LLPS”のアイコンと液滴の名称をつけて表示するよう，データを改訂，公開した．

「リン酸化情報のフォーマットと DB 化：リン酸化により相互作用や機能が制御される事例について， 情報の記述をフォーマット化し，公開する」については，リン酸化に伴うイベントの概念設計を行い， IDEAL から公開したので達成された．

「天然物化合物 DB，天然変性タンパク質 DB の拡充：科学文献の網羅的調査により生薬に含まれる天然物化合物と活性の情報の収集し KNApSAcK に格納する．IDEAL と IDEAL データベースサーバを更新しデータの安定的なサービス体制を整える」については、報告した通り，KNApSAcK，IDEAL ともにデータ拡張を実施した．目標以上に達成されている．

全体で 20 件の支援を実施し，これまでに 18 件を終了した．現在，進行中の研究は，0830：FANCM による RNaseH1 の DNA 架橋損傷修復複合体へのリクルート機構と生理的意義 0831：不溶性タンパク質と GroE リガンドの特徴抽出と識別法の開発であり，共に，依頼者が論文を作成するところまで進捗している．

名古屋大学グループでは IDEAL をアップデートし，公開した．登録されているタンパク質は 1110 本となった．リン酸化などを受けて天然変性タンパク質の相互作用や複合体形成が遷移する様子をデータ化するための専用エディターの開発を行い，プロトタイプを作成した．データは SBGN と XML で出力され，前者をグラフツールに入力すると遷移が視覚化される．

長浜バイオ大では，潜水動物祖先型ミオグロビンの再現(iScience, 2021, 日本経済新聞 2021/9/19)，糖脂質型酵素 MPIase の構造モデリング(ACS Chem Biol, 2022)などの支援を実施した．また高度化課題として天然物・医薬品・代謝物の立体構造データベースを構築し，機械学習(GBDT)により薬効に結びつく分子パラメータを探索し，疏水性表面積・回転楕円体サイズなどいくつかの立体構造パラメータが有意に薬効識別に役立つことを示した．またこのデータベースを利用し，COVID-19 に対する天然物医薬品の探索を行い，感染研との共同研究により植物アルカロイドであるセファランチン・テトランドリンが有意な SARS-CoV-2 増殖抑制活性を持つことを示した(iScience 2021, FEBS OpenBio 2021)．

英文

In the task, we will upgrade KNApSAcK database consisting of species-metabolite relationships and the world's only crude drug database developed by the applicant toward drug discovery applications, especially for structure-based drug discovery (SBDD).

Specifically, with a focus on the human metabolism and gene control map, we performed (1) Expansion of natural compound data, proposal of new drug targeted compounds, prediction of ligand binding, (2) Expansion of intrinsically disordered protein DB: association between metabolism and transcription and (3) Registration, update and publication of KNApSAcK and IDEAL

**(1) Expansion of natural compound data, proposal of new drug targeted compounds, prediction of ligand binding**

For the purpose of proposing new drug skeletons and analyzing common drug designs by comparing human metabolites with natural products and approved drugs, we performed structural classification by three-dimensional structure clustering of

compounds by graph matching using COMPLIG using compounds including in KNApSAcK DB (physiologically active natural substance molecule, 3,055 molecule), approved drug (DrugBank, 1,985 molecule) for the purpose of proposing a new drug skeleton and analyzing a common drug design by comparing human metabolites with natural products and approved drugs, human metabolites (KEGG, 1,396 molecules), and toxic substances (517 molecules). In addition, in silico screening based on these natural product information was performed and summarized in papers (Mol. Inf., 2022; Sci Rep, 2022). Furthermore, for the purpose of estimating the target protein of the crude drugs and constructing a complex structure model, the crude drug substances and the PDB ligand were matched using COMPLIG, and 1,629 molecules were linked to the PDB ligand. We analyzed the characteristics of natural products, approved drugs, human metabolites, and toxic molecules by machine learning, and reported preliminary results (*蛋白質科学会* 2021). On the basis of AI drug discovery, we have established an in-silico screening method for predicting activity using molecular graph convolution neural network, which is one of deep learning. Here, molecules are represented by chemical bonds as bonds and atoms as vertices (Mol Inf, 2020; BMC Binformatics 2019).

In the COVID-19 pandemic, structural modeling of SARS-CoV-2 protein was performed by using physiologically active natural products, approved drugs, human metabolites and toxic molecule structural data, then, we proposed complex model for interaction between proteins such as main protease, S-glycoprotein, and 2'-O-ribose methyltransferase and ligands for physiologically active natural products and so on. This results are available from BINDSweb sites (http://harrier.nagahama-i-bio.ac.jp /dtx/SARS-CoV-2/), and BSM-Arc (https://bsma.pdbj.org/entry/ 15) (***FEBS Lett***. 2020, *生物物理* 2021).

**(2) Expansion of intrinsically disordered protein DB: association between metabolism and transcription**

Intrinsically disordered proteins provide a role for signal transduction, and transcriptional regulation. IDEAL database focused on molecular cell event starting from phosphorylation and expanded to metabolite-related transcriptional regulation based on phosphorylation in proteins. Those data provide DTX database which make it possible to examine between proteins and ligands and protein phosphorylation.

**(3) Registration, update and publication of KNApSAcK and IDEAL**

In KNApSAcK DB, for the purpose of enriching information on natural products, we have established a procedure for accumulating data even if COVID-19 pandemic, and the number of registered natural products in the database has increased as 56,829 substances. We also have built a system that can be used for various in silico screening.

Currently, there are 135,156 relationships between natural products and species that produce them, indicating the possibility of crude drugs as biological species. In addition, the annual number of accesses exceeds 1 million, and it plays a role as the basis of research on natural products and crude drugs in the world.

IDEAL carried out notations of intrinsically disordered proteins and updates of existing data as scheduled. Currently, more than 10,000 intrinsically disordered region information is provided for 995 entries.

In addition to (1)-(3), we supported a total of 20 researches in five years project.