【報告様式Ａ】

作成／更新日：令和４年５月30日
課題管理番号：21am0101112j0005

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）関嶋政和
（英　語）Masakazu Sekijima

補助事業担当者　所属機関・部署・役職：
（日本語）国立大学法人東京工業大学・情報理工学院・准教授
（英　語）Tokyo Institute of Technology, School of Computing, Associate Professor

II 補助事業の概要

本研究における研究開発体制（代表者・分担者）及び、それぞれの担当する支援と高度化の分担は次の通りである。我々の支援の中心である共用ファシリティとしてのTSUBAME利用支援を関嶋が、インシリコスクリーニングに関わる支援・高度化を関嶋、タンパク質天然変性領域予測に関わる支援・高度化を石田、結合タンパク質探索・モデリングに関わる研究を秋山・大上が担当してきた。

TSUBAME利活用支援は、2017年度より順調に行われており、2019度の「京」の停止後も、またコロナ禍において期待されたCOVID-19治療薬探索研究に対してインシリコユニットの研究者の計算機リソースが不足しないよう、PSPOと連携しつつ支援を実施してきた。

また、高度化においては、PSPOより「機械学習による創薬開発手法の開発に期待する」というコメントを頂いていたが、相互作用エネルギーを学習するスクリーニング手法SIEVE-Scoreはインシリコ創薬におけるトップジャーナルに掲載されたほか、深層学習を用いたvisual inspectionを代替する手法であるVisINet（VISual Inspection NETwork）については特許出願を行い、現在はCOVID-19の創薬標的であるMain Protease(Mpro)の阻害化合物の獲得、さらにはMproについての創薬標的データベースを構築への応用を行っており、こちらもマイルストーンに対して、十分な成果を上げていると考えている。天然変性領域予測については、支援と高度化研究のいずれも計画通りマイルストーンが達成されている。タンパク質間相互作用予測については、高度化研究が計画以上に進展したため、高度化の内容について予定より早く支援の提供を開始した。東工大TSUBAMEスパコンの創薬支援分野における利活用支援、他方は独自ツール群による計算支援を行い、加えて新型コロナウイルス治療薬探索を実施した。TSUBAME 3.0は、14コアCPUを2基と最新Pascal-GPU(P100)を4基搭載した

計算ノードが計 540 台ノード接続された計算機である。インシリコユニットの研究グループに対し、上述計算機資源と分子シミュレーションや機械学習に必要な計算機資源、および各種ソフトウェア（量子化学計算プログラム：Gaussian, GAMESS, 分子動力学計算プログラム：Amber, NAMD, GROMACS の提供を行うほか、プログラムの更新支援やタンパク質立体構造予測プログラム AlphaFold の導入支援など、各グループの支援研究への貢献を行っている。もう一方は、TSUBAME を用いることで可能になる大規模分子シミュレーションを中心とした構造インフォマティクス技術によって、インシリコスクリーニング支援、具体的には、我々がこれまでに開発した高精度化合物スクリーニング法 SIEVE-Score や VisINet、世界最高水準のタンパク質天然変性領域予測システム PrDOS-CNF、大規模並列計算によるタンパク質間相互作用予測 MEGADOCK を最大限に活かした支援を行っている。

支援の例として「リガンド-タンパク質相互作用の動的解析とリガンド構造展開方法」について例を挙げる。脂肪酸は G 蛋白質共役型受容体(GPCR)やイオンチャネルを始めとした広範な膜タンパク質と相互作用することから、エネルギー代謝・免疫・中枢領域の創薬シーズとして大きなポテンシャルを持つ。本課題では、脂肪酸および脂肪酸の一部構造を閉環することによって固定（ノンリピッド化）した分子について、取り得る立体配座空間を分子動力学シミュレーションにより網羅的に探索し、特徴的な構造を抽出する解析を進めている。レプリカ交換法を用いて各分子の立体配座空間をサンプリングし、同じ部位に相当する二面角を比較したところ、脂肪酸 A) では多くの配座が出現しているが、構造を固定した分子 B) では一部の配座のみが出現していることがわかった。本研究により脂質リガンドのリゾホスファチジルセリン受容体への結合機構の解明を行い（Sayama et al., Biochemistry, 2020)、現在は得られた網羅的構造サンプリングデータと合成実験のデータを元とした論文出版を進めている(Jung et al.,ChemRxiv,2022)。

分担者の秋山および大上らが、独自のタンパク質間相互作用（Protein-Protein Interaction, PPI）予測システムである MEGADOCK と、本 BINDS 高度化成果である MEGADOCK-Web データベースを用いた PPI の結合相手となるタンパク質の探索・モデリング支援および支援のため開発整備を実施した。H30 年度までの目標であった TSUBAME 3.0 の GPU をフルに活用した MEGADOCK 実行環境の整備を完了しており、また当初 R1～2 年度で予定していたウェブ版の PPI 予測結果検索データベース MEGADOCK-Web を、高度化研究の進捗が当初予定より大幅に進んだため、支援に関しても H30 年度に提供することができた。R2~3 年度で新たにフォーカストデータベースである MEGADOCK-Web-Mito を開発・公開するなど、これまでの計画を当初予定以上に達成した。コンテナ型仮想化環境を含めたパッケージとしての提供や、クラウド計算環境でのサービスプラットフォームの開発、GUI クライアントの開発を残期間で取り組むとともに、TSUBAME 3.0 およびウェブシステム双方で複数の PPI 予測支援を引き続き実施する。R3 年 9 月に実施予定の AMED-BINDS 講習会「創薬のためのタンパク質構造インフォマティクス 2」にて、大上による MEGADOCK 利用講習も実施する。

高度化においては、機械学習を用いてドッキング計算の結果から高精度に化合物の活性を予測する手法である SIEVE-Score の開発を行った。概要を右図に示す。まず標的タンパク質に対して活性が既知の化合物をドッキングし、その結果からタンパク質とリガンド間の相互作用エネルギーを抽出したものである相互作用エネルギーベクトルを得る。次に Random forest によりこれらの相互作用を学習し、活性の有無を予測するモデルを作成し、スクリーニングを行う。立体構造の特徴を 0 か 1 かではなくエネルギーという連続な形で評価できる点が新規な点である。比較実験では、ドッキングに Glide SP mode, データセットに DUD-E を用い、5-fold クロスバリデーションでの ROC 曲線下の面積である AUC の平均値を用いてリランキング前の Glide SP モードとの比較を行った。この結果、DUD-E の全 102 タンパク質において、SIEVE-Score, Glide SP mode に対して 97 標的で AUC が向上する結果となった。いずれもランキング上位の予測精度が特に改善しており、上位を正しく順位付けすることが求められるバーチャルスクリーニングにおいて有用であることがわかった。

本成果を報告した論文は、インシリコ創薬におけるトップジャーナル Journal of Chemical Information and Modeling 誌の Machine Learning Special Issue において Supplementary Cover Art（表紙）にも選ばれた（N.

Yasuo and M. Sekijima, JCIM, 2019)。本成果はgithubを通じて、プログラムの公開を行い、広く研究者が利用可能となっている（https://github.com/sekijima-lab/SIEVE-Score)。

また、ドッキングを用いたスクリーニングにおいては、早い段階から人間がドッキング結果の構造（ドッキングポーズ）を目視することで良い構造を選別する visual inspection が広く行われてきた。しかし、visual inspection は属人性が高く人によって評価の基準が異なる点、また数百万にも上る化合物ライブラリに対して全てのドッキング構造を目視で判別することは困難である点が問題であった。本高度化研究では、ドッキング結果の三次元構造を画像化し、画像認識分野に多く用いられている深層学習手法である convolutional neural networkを用いてドッキング結果をリランキングすることにより、visual inspectionを代替する機械学習手法であるVisINet（VISual Inspection NETwork)の開発を行った。

VisINetでは、タンパク質とドッキングされた化合物を全方位から81枚の画像にし、深層学習のモデルであるResNetを用いて学習を行う。これまでに行ったDUD-Eデータセットのうちの8タンパク質を対象として行った評価実験では、1標的タンパク質あたり158,760から2,045,574枚の画像をスーパーコンピュータTSUBAMEを用いて学習することで、Glide SP modeによるドッキングおよび上述したSIEVE-Scoreを上回る精度の活性予測を実現した。また、画像認識で用いられる解釈手法GradCAMを用いることにより、画像上のどの部位を認識して活性予測を行なっているか特定することも可能である。

本手法は既に特許を申請しており（H31年1月31日申請, 特願2019-015086)、新型コロナウイルス治療薬探索でもMain Proteaseの新規阻害化合物の獲得に貢献している。また、研究を担当していた大学院生の依田（当時M2、補助事業参加者）が2019年日本薬学会第139年会及び情報処理学会第18回全国大会で行った当該研究の発表が、学生優秀発表賞及び学生奨励賞を受賞した。

新型コロナウイルス（SARS-CoV-2）に関して、Main Proteaseと既知リガンドとの結合様式を分子動力学シミュレーションなどのシミュレーション手法等で調べ、COVID-19に対する治療薬候補化合物に求められる物理化学的な特徴であるファーマコフォアを明らかにした（Yoshino et al., Sci Rep, 2020)。本論文は、出版1年未満である令和3年5月5日時点で33回引用されているなど、国内外から多くの注目を集めている。二つの水素結合ドナーと二つの水素結合アクセプター、ペプチド様阻害剤の主鎖のカルボニルとアミンを認識、相互作用しているアミノ酸残基はSARS-CoV, SARS-CoV-2間で保存されているというこれらの特徴は、医薬品候補とされるα-ケトアミド阻害剤（Zhang et al., Science, 2020）に対する検証でも適合することが確認を行った。また、型コロナウイルス治療薬探索では、これまで探索されている SARS-CoV-2 のウイルス複製に必要な酵素の一つである Main Protease の阻害剤にペプチド様化合物が多いため、ペプチド様化合物を避けた化合物ライブラリを構築し、それに対してバーチャルスクリーニングを実施した後にアッセイ試験を行うことで、Main Protease を阻害するペプチド様ではない既知の阻害剤とは異なる空間に属し構造新規性が高い化合物を 6 個発見した(Yamamoto et al., JCIM,2022)。

英文：

The research and development structure of this research (principal investigators and collaborators) and their respective responsibilities for support and enhancement are as follows. Sekijima has been in charge of TSUBAME supercomputer utilization support, which is the core of our support; Prof. Sekijima has been in charge of support and advancement related to in silico screening; Prof. Ishida has been in charge of support and advancement related to protein natural denaturation region prediction; and Prof. Akiyama and Prof. Ohue have been in charge of research related to binding protein discovery and modeling.

We supported the utilization of the Tokyo Tech TSUBAME supercomputer in the field of drug discovery support, and on the other hand, we provided computational support using our own tools, and in addition,

we conducted a search for new coronavirus therapeutics.

TSUBAME 3.0 is a computer with two 14-core CPUs and four state-of-the-art Pascal-GPUs (P100), with a total of 540 nodes connected. We provide the above-mentioned computer resources, computer resources necessary for molecular simulations and machine learning, and various software (quantum chemistry calculation programs: Gaussian, GAMESS, molecular dynamics calculation programs: Amber, NAMD, GROMACS) to the research groups in the in silico unit. In addition, we contribute to the research supported by each group by assisting with program updates and the introduction of AlphaFold, a program for protein conformation prediction.

On the other hand, the structural informatics technology centered on large-scale molecular simulations made possible by TSUBAME is used to support in silico screening, specifically, to maximize the use of SIEVE-Score and VisINet, the high-precision compound screening methods we have developed, the PrDOS-CNF, the world's best system for predicting natural protein denaturation regions, and MEGADOCK, a large-scale parallel computation system for predicting protein-protein interactions. We also provided support by maximizing the use of PrDOS-CNF, the world's best prediction system for naturally denatured proteins, and MEGADOCK, which predicts protein-protein interactions using massively parallel computations.

In terms of advancement, PSPO commented that they were looking forward to the development of drug development methods using machine learning. SIEVE-Score, a screening method that learns interaction energies, was published in a top journal in in-silico drug discovery. We have also applied for a patent for VisINet (Visual Inspection NETwork), an alternative to visual inspection using deep learning, and are currently working on acquiring inhibitors of Main Protease (Mpro), the drug target of COVID-19, as well as on the development of a drug target database for Mpro. We are now working on the acquisition of inhibitors of Main Protease (Mpro), a drug target of COVID-19, and the application to the construction of a drug target database for Mpro. We believe that we have made satisfactory progress toward our milestones.

For the prediction of naturally degenerate regions, milestones for both support and advanced research were achieved as planned. As for protein-protein interaction prediction, the advanced research made more progress than planned, so we started providing support for the content of the advanced research earlier than planned.