データ駆動・AI 駆動を中心としたデジタルトランスフォーメーションによる生命科学研究の革新　　2022 年度年次報告書

2021 年度採択研究代表者

岡田 眞里子

大阪大学 蛋白質研究所
教授

自然言語処理とシミュレーションによる細胞制御探索法の構築

主たる共同研究者：

下平 英寿（京都大学 情報学研究科 教授）
泰地 真弘人（理化学研究所 生命機能科学研究センター 副センター長・チームリーダー）

研究成果の概要

本研究では、計算科学主導の細胞制御の方法論の確立を目的として、(1)自然言語処理による遺伝子相互作用の抽出、(2)数式フリーの細胞数理モデリング手法の構築、(3)細胞制御化合物デザインの自動化といった一連の要素技術を開発する。(1)に関しては、PubTator と公共データベースを組み合わせて教師データを作成することにより、細胞内反応における遺伝子間の相互作用と制御様式の関係を、70-90%の高い成功率で得ることができた。また、一方で、大規模言語モデルのタスク表現を工夫することで、一本の原著論文から、Text2Model(論文発表済)フォーマットを吐き出し、BioMASS ソフトウェア(論文発表済)を用いたシグナル伝達系のシミュレーションの実行に成功した。さらに、KEGG データベースの遺伝子相互作用情報を Text2Model フォーマットに変換させることで、KEGG Pathway から半自動的にモデル構築とシミュレーションが可能(KEGG2Model)になった。このように、Text2Model を中間に置くことで、様々なデータソースから、数理モデルの構築とシミュレーションが実現できた。さらに、BioConceptVec を用いて、遺伝子の単語ベクトルと薬剤の単語ベクトルの足し算・引き算によって遺伝子を選んだ場合、かなり良い精度でその関連候補を絞れることが分かった。(2)に関しては、プロトコル論文を発表したほか、グラフ描画や KEGG2Model 機能を追加した BioMASS のアップデートを進めている(論文準備中)。(3)に関しては、機械学習による薬物生成モデルと分子シミュレーションを連携させ、細胞老化モデル(論文投稿中)により同定された老化関連代謝酵素について、阻害剤の設計と実験検証を進めている。また、肺がん細胞モデルにおいて重要と予測された EGFR の特定の変異による蛋白質相互作用への影響を予測するため、3−4種の変異型 EGFR に対して、MD シミュレーションを実行中である。研究はおおむね順調に進展している。

【代表的な原著論文情報】

1) Iida K, Kondo J, Wibisana JN, Inoue M, Okada M. ASURAT: functional annotation-driven unsupervised clustering of single-cell transcriptomes. ***Bioinformatics.*** 2022 Sep 15;38(18):4330-4336. doi: 10.1093/bioinformatics/btac541. PMID: 35924984; PMCID: PMC9477531.
2) Wibisana JN, Inaba T, Shinohara H, Yumoto N, Hayashi T, Umeda M, Ebisawa M, Nikaido I, Sako Y, Okada M. Enhanced transcriptional heterogeneity mediated by NF-κB super-enhancers. ***PLoS Genet.*** 2022 Jun 1;18(6):e1010235. doi:10.1371/journal.pgen.1010235. PMID: 35648786; PMCID: PMC9191726.
3) Imoto H, Yamashiro S, Murakami K, Okada M. Protocol for stratification of triple-negative breast cancer patients using in silico signaling dynamics. ***STAR Protoc***. 2022 Aug 11;3(3):101619. doi: 10.1016/j.xpro.2022.101619. PMID: 35990741; PMCID: PMC9389415.