論 文 の 内 容 の 要 旨

論文題目　　化学ー生物情報を協奏的に用いる創薬システムの開発

氏　　名　　海東　和麻

1. 序論

超高齢社会の到来や、感染症による公衆衛生の危機などの諸課題に対応するために、新薬の創出が強く望まれている。一方、創薬の難易度は非常に高く、効率的な創薬を実現するために、初期段階における緻密な医薬品分子設計が求められる。医薬品分子の設計は、標的タンパク質に結合する出発点となるヒット化合物と、構造展開により生物活性と薬物動態の双方が改善されたリード化合物の設計が焦点となる。ヒット化合物導出のために大規模スクリーニングが行われているが、その有用性は化合物ライブラリの質に大きく依存する。標的タンパク質の高次情報があることが望ましいが、結合部位を含めたその情報を実験的に取得することは困難である。

近年の有機合成化学の発展に伴い、新奇化学構造を導入した化合物の SAR 取得が盛んになっている。例として、bicyclo[1.1.1]pentane（BCP）は benzene の bioisostere とされており、化合物中への BCP 導入により、代謝安定性などの物性を改善することができ、リード化合物の導出とその最適化への寄与が期待されている[1]。

深層学習を中心とする機械学習技術の発達に伴い、創薬効率化のために人工知能（AI）を導入する試みがなされている。分子設計に用いられる AI として、構造生成器が挙げられる。構造生成器は、初期条件に従い、分子の構造式を自動的に出力するアルゴリズムである。Variational autoencoder（VAE）をはじめとする深層学習を用いた構造生成器は多数考案されている[2]。一方、生成した化学構造が包括的な生物応答を考慮しているとは言い難い。薬物代謝反応における代謝部位（site of metabolism: SoM）を予測する機械学習ベースの AI も考案されている[3]。SoM あるいは代謝物の構造を予測するには、限られたデータから、化学的な根拠を含めた予測が必要となる。

次世代シーケンサーやマイクロアレイなどの技術の発達と、データ集積環境の整備により、大規模なトランスクリプトームデータを解析することが可能となりつつある。遺伝子が転写されて生じる mRNA の発現量データである遺伝子発現プロファイルを解析することで、創薬標的の予測や、ドラッグリポジショニング予測などが行える[4]。このことは、遺伝子発現プロファイルは生物応答情報を包括的に有していることを示している。

本研究の目的は、化学ー生物情報を協奏的に用いる創薬システムの開発である。この創薬システムは、以下の 3 つのサブシステムから構成される。

・遺伝子発現プロファイルを入力してヒット化合物の構造を生成するシステム
・特異な化学構造を重点的に生成するシステム
・化学的解釈が可能な代謝物予測システム

化学情報と生物情報を効果的利用により、生体応答を考慮した医薬品分子設計が実現する。

2. オミクスデータを用いた構造生成器の開発

遺伝子発現プロファイルを入力することで、標的タンパク質に対するヒット化合物の構造式を出力する構造生成器、**TR**anscriptome-based **I**nference and generation **O**f **M**olecules with desired **PHE**notypes by machine learning（TRIOMPHE）[5]の開発を目的とした。TRIOMPHEは、発現類似化合物の選択と、VAE による構造展開の 2 つのステップから構成される。最初のステップでは、分子生物学的手法により標的タンパク質の遺伝子をノックダウンさせることで得られた標的摂動プロファイルと、化合物添加により遺伝子摂動を与えることにより取得した化合物応答プロファイルとの相関係数を算出する。高い相関係数を示した化合物応答プロファイルに対応する化合物の構造式を、発現類似化合物として選択した。次のステップでは、発現類似化合物を VAE に入力し、潜在空間座標をサンプリングし、decoder で復元することで新規化合物を生成した。

新規化合物生成に有利な VAE を構築するために、化合物表記法として SMILES と SELFIES[6] のどちらが有用か検証した。SELFIES-based VAE は SMILES-based VAE と比較して、生成した文字列が化学構造として成立する割合である validity と重複を除いた割合である uniqueness について高い値を示した。

TRIOMPHE により生成した化合物の標的タンパク質に対するリガンドらしさを、既知リガンドの化学構造と比較することで検証した。TRIOMPHE により生成した化合物は、GAN を用いた既往手法[7]と比較して、リガンドとして重要な部分構造を有する化合物を生成することができた（Figure 1）。提案手法である TRIOMPHE は、遺伝子発現プロファイルを元に、標的タンパク質のリガンドとして機能する化学構造を有した化合物を生成することができた。

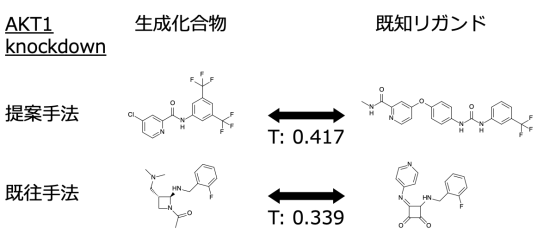

Figure 1 生成した化合物の例

3. 特異な化学構造を重点的に生成するハイブリッド型構造生成器の開発

化合物の中心骨格である scaffold を保持しつつ、特異な化学構造を有する高品質な virtual library（VL） を構築する構造生成器 Exhaustive Molecular library Production In a scaffold-REtained manner (EMPIRE) 開発を目的とした。 EMPIRE では、VAE を用いた深層学習型構造生成器と、予め指定した部分構造を付加するビルディングブロック型構造生成器のそれぞれに fragment を入力し、新規 fragment を生成する。この fragment を scaffold に組み合わせることで新規化学構造を生成した。特異構造をビルディングブロックリストに含めることで、新奇な部分構造を含む化合物を重点的に生成することができる。また、深層学習型構造生成器を併用することで、膨大なビルディングブロックリストを用意することなく、入力した fragment と類似した新規 fragment を生成し、多様な化合物を生成することができる。5 種類の scaffold それぞれを入力して化合物を生成した。EMPIRE では、scaffold の種類によらず、安定して化合物を生成することができた。また、提案手法では scaffold そのものと類似度の高い構造を多く生成することができた。極端に巨大な fragment を付加しないことから、scaffold に対して微小な変化を加えられた化合物が多く、virtual screening に有利な VL を構築できたと考えられる。BCP を重点的に生成することができるかを検証した。EMPIRE は、入力化合物の scaffold を保持しつつ、部分構造に BCP を含む化合物を網羅的に生成することができた（Figure 2）。

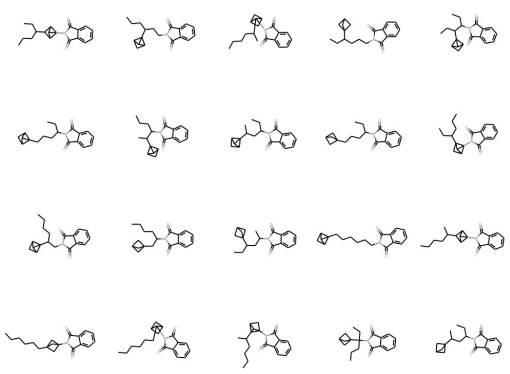

Figure 2 EMPIRE により生成した化合物の例

4. 新規記述子を用いた代謝反応予測モデルの構築

化学的に解釈可能な記述子を用いた SoM 予測モデルの構築及びその結果を用いて代謝物の構造式を予測するシステム **ME**tabolite **R**ule-based **C**onverter by **I**ndex（MERCI）の開発を行った[8]。新規記述子として、周辺構造とMulliken 電荷から算出される原子の電気的性質である Gasteiger Charge[9] (GC) と分子のトポロジカル的性質の両方を反映させた GC-path 記述子と GC-skeleton を考案した。GC-path 記述子は、化合物の構造式における原子間のトポロジカル距離と GC を元に算出した記述子である。GC-skeleton 記述子は、指定した部分構造内における GC の特徴を元に算出した記述子である。既往の記述子セットである ordinal set、提案した記述子である proposed set、既往記述子、提案記述子を組み合わせた combined set の三種類の記述子グループで CYP3A4 に対する SoM 判別予測モデルを構築した。構造変換システム MERCI は、SoM 予測モデルの結果と SoM 周辺の化学構造情報に従い、ルールベースで反応の種類を予測し、代謝物の構造式を自動的に出力する。

各記述子グループについて 10 個の SoM 判別予測モデルを構築し、それらのテストデータに対する予測性能を Matthews correlation coefficient（MCC）により評価した。RF、GBDT いずれの手法を用いた場合でも、提案した記述子を用いた場合に高い MCC を示した。特に、184 種類の記述子からなる combined set を元に構築した SoM 予測モデルのうち、最も高い MCC は 0.618 となった。

MCC 0.618 を示した SoM 予測モデルの寄与率を元に、SoM 予測に有用な記述子を評価した。寄与率の大きい上位 10 種類の記述子の内、4 種類が提案した GC-path 記述子であり、SoM 予測における提案記述子の有用性が示された。

CYP3A4 に対する基質 75 分子のうち、51 分子について、実験的に同定されている代謝物の構造を取得できた。さらに、24 分子については、全ての代謝物の構造を過不足なく得ることができた。生成代謝物の種類も、脱アルキル化やヘテロ原子酸化をはじめ、エポキシ化など様々な反応に由来する構造を取得できた（Figure 3）。

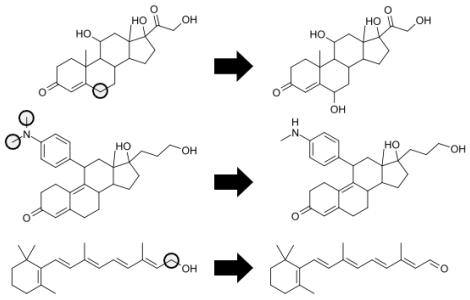

Figure 3 MERCI により得られた代謝産物の例

5. まとめと今後の展望

本研究では、新薬開発の初期に相当する分子設計を効率化するための創薬システムを構築した。このシステムにより、創薬標的タンパク質に関する遺伝子発現プロファイルから、ヒット化合物の創出と新奇構造を有する構造展開が可能となる。標的タンパク質の高次構造情報を用いることなく、分子生物応答を考慮することで、所望の生物活性を有するヒット化合物を創出できる。このヒット化合物を、新奇化学構造を含む構造展開を行うことで、生物活性や薬物動態が改善された、リード化合物創出に寄与する。生成した新規化合物を、SoM 予測を介して代謝産物の構造式を取得することで、代謝毒性を回避した分子設計の指針を取得することができる。構築したシステムは、創薬の効率化や化学物質設計への展開により、基礎化学・薬学の進展と社会への安全・安心創出に貢献することが期待される。

6. 参考文献

[1] Stepan, A. F. *et. al.*, *J. Med. Chem.* **2012**, *55*, 3414-3424.

[2] Gómez-Bombarelli, R. *et. al.*, *ACS Cent. Sci.* **2018**, *4*, 268-276.

[3] Hasegawa, K. *et. al.*, *Mol. Inform.* **2010**, *29*, 243-249.

[4] Iwata, M. *et. al.*, *J. Med. Chem.* **2018**, *61*, 9583-9595.

[5] Kaitoh, K. and Yamanishi, Y. *J. Chem. Inf. Model.* **2021**, *61*, 4313-4320.

[6] Krenn, M. *et. al.*, *Math. Learn.: Sci. Technol.* **2020**, *1*, 045024.

[7] Méndez-Lucio, O. *et. al.*, *Nat. Commun.* **2020**, *11*, 10.

[8] Kaitoh, K. *et. al.*, *Mol. Inform.* **2019**, 201900010.

[9] Gasteiger, J. and Marsili, M. *Tetrahedron* **1980**, *36*, 3219-3228.