

論文 / 著書情報
Article / Book Information

題目(和文)	機械学習とシミュレーションを組み合わせた薬物代謝酵素(CYP3A4)の結合様式予測法の研究
Title(English)	
著者(和文)	佐藤敦子
Author(English)	Atsuko Sato
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第11541号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:山村 雅幸,小野 功,青西 亨,瀧ノ上 正浩,関嶋 政和,小長谷 明彦, 本間 光貴
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第11541号, Conferred date:2020/3/26, Degree Type:Course doctor, Examiner:,,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

論文要旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位 (専攻分野)： Academic Degree Requested	博士 Doctor of	(学術)
学生氏名： Student's Name	佐藤 敦子		指導教員 (主)： Academic Supervisor(main)	山村雅幸 教授	
			指導教員 (副)： Academic Supervisor(sub)	本間光貴 特定准教授	

要旨 (和文 2000 字程度)

Thesis Summary (approx.2000 Japanese Characters)

本論文は、「機械学習とシミュレーションを組み合わせた薬物代謝酵素 (CYP3A4) の結合様式予測法の研究」と題し、6 章から構成される。

第 1 章「序論」では、研究の社会的、技術的な背景、研究の目的と全体像について述べている。本研究ではタンパク質の薬物結合ポケットの動きを考慮して薬物の結合様式を予測する方法の開発を目的とし、題材としてポケットが大きく、主鎖レベルで構造の揺らぐ CYP3A4 を用いて以下 2 つのアプローチを検討した。1 つ目のアプローチはタンパク質と化合物複合体の 3 次元構造を用いた機械学習による結合様式予測である。2 つ目のアプローチは複数の短時間 MD シミュレーションを並列させた結合様式予測である。また本研究の背景として、創薬研究における結合様式予測効率化の重要性と、予測の難しいターゲットタンパク質に対する技術的な課題について述べている。

第 2 章「関連研究」では、結合様式予測の難しいターゲットに対する技術的な課題を、「機械学習による結合様式予測」と「シミュレーションによる結合様式予測」に分け、関連研究を整理し、本研究の位置づけを示した。機械学習による結合様式予測では、3 次元構造情報を保持したタンパク質構造および化合物構造の入力手法と、訓練データとなる 3 次元構造情報が不十分な場合の学習方法ならびに高品質データセットの作成方法論の必要性について述べた。シミュレーションによる結合様式予測では、結合する薬物の形状に応じて結合ポケットが柔軟に変化する予測難度の高いターゲットタンパク質に対するアプローチを整理しつつ、十分な構造をサンプリングする方法論の必要性を示した。

第 3 章「機械学習による結合様式予測」では、3 次元構造情報を用いた機械学習スコアリングとして 3D-CNN を利用した CYP3A4 の結合様式予測に取り組んだ。従来研究では、ターゲットに依らないデータセットでの予測モデル構築が検討されたが、本研究では、訓練データが不十分である場合のデータセットとしてさまざまな特徴とデータサイズを有する 4 つの訓練セットを選択し、CNN モデルの事前学習とファインチューニングを検討した。その結果、従来の結合様式評価法 (エネルギーベースのドッキングポーズ評価、ターゲットに依らないデータセットでの予測モデルによる評価) と比較して大幅に予測精度が向上し、その要因の考察を通じてデータセット選択の方法論を構築した。

第 4 章「MD シミュレーションによるタンパク質の動きを考慮した結合様式予測」では複数の MD 初期構造を用いた短時間シミュレーションによる CYP3A4 の結合様式予測に取り組んだ。従来研究では、化合物の位置に着目して MD 初期構造が選択されたが、本研究ではより柔軟な構造を持ち、結合様式予測の難度が高いケースにおいて、タンパク質と化合物間の相互作用に着目し

て初期構造の選択を行った。その結果、従来法とは異なり、MD 初期構造の選択には化合物側よりも相互作用に着目した選択が有効であり、また、短時間の MD シミュレーションを複数並行して実施する手法が結合様式予測の難度が高い化合物においても有用であることを示した。

第 5 章「総合討論」では第 3 章、第 4 章で構築した手法の応用および機械学習で用いたデータセットの考察について述べた。まず、機械学習とシミュレーションを組み合わせることで結合様式予測を効率化させる方法の提案に向け、第 3 章で得られた機械学習モデルを MD 初期構造の選択に適用した結果を確かめた。その結果、従来法で選択した方法と比較して結合様式を正しく予測できることが示された。第 3 章の機械学習アプローチで作成したデータセットに関する考察では、オキシドリダクターゼの比率、ポケットサイズに注目した比較の他に、複数の観点でデータの分布を調査した。そのうちタンパク質側に注目した切り口では、タンパク質ファミリーセットが標的に依らないデータセットをよく補完していることを再確認し、一方で化合物側に注目した切り口では、データセット間の分布に偏りが無いことを示した。

第 6 章「結論」では、第 5 章までの主要な結果をまとめ、今後の展望や課題について述べた。本研究では、結合サイトが柔軟で大きな高難度ターゲットの結合様式予測を目的に、CYP3A4 を題材として機械学習による結合様式予測に向けたデータセット作成のための方法論と、複数初期構造を用いた MD シミュレーションによる結合様式予測法を示した。また、機械学習とシミュレーションそれぞれについて、今後の課題と展望を示した。将来、シミュレーションと機械学習を真に融合させた高難度ターゲットの結合様式予測として、1) シミュレーションの強みである実験では得られないような仮想大量データを生成、2) その大量データによる結合様式予測機械学習モデルの訓練、3) その結合様式予測機械学習モデルによる MD 初期構造の選択、というサイクルを繰り返すことにより、予測手法の高速化、高精度化が期待され、汎用的な手法として創薬研究に貢献すると考えている。

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note: Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1 copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

論文要旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース
学生氏名： Student's Name	佐藤 敦子	

申請学位 (専攻分野)： Academic Degree Requested	博士 Doctor of	(学術)
指導教員 (主)： Academic Supervisor(main)	山村雅幸 教授	
指導教員 (副)： Academic Supervisor(sub)	本間光貴 特定准教授	

要旨 (英文 300 語程度)

Thesis Summary (approx.300 English Words)

The subject of this study is to predict the binding mode of proteins and compounds. Traditionally, docking calculations have been used to predict protein-compound binding modes. However, the accuracy of the prediction is not enough for difficult targets that having large and flexible binding pockets such as drug-metabolizing enzymes (CYP3A4).

In this research, binding mode prediction that combines machine learning and molecular dynamics calculations for target proteins that are difficult to predict by conventional methods, such as CYP3A4 have been developed. In this study, four training sets with various features and data sizes were selected as datasets when training data was insufficient, and pre-training and fine-tuning of CNN models were examined. As a result, the prediction accuracy has been greatly improved compared to the conventional method of binding mode evaluation, and a data set selection methodology has been established through consideration of the factors.

In the binding mode prediction considering the flexibility of the protein by MD simulation, the binding mode prediction of CYP3A4 by short time simulation using multiple MD initial structures was examined. In this study, the MD initial structure was selected by the interaction between the protein and the compound when the compound structure was more flexible than previous studies and it was difficult to predict the binding mode. As a result, the initial MD structure selection was more effective when focused on the interaction than on the compound coordinate. In addition, it was shown that the technique of performing multiple short-time MD simulations in parallel was useful even for compounds that were difficult to predict the binding mode.

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).