平成 29 年度　利用報告書

課題名（タイトル）：

創薬プロセス効率化を目指した機械学習のための学習データの作成

利用者氏名：

〇千葉 峻太朗[1]、大田 雅照[2]、藤原 崇幸[3]、宮口 郁子[3]、中田 一人[3]、鹿島 亜季子[3]、小久保 裕功[3]、佐藤 美和[3]、古川 祐貴[4]、金重 光典[4]、長代 新治[4]、馬 彪[5]、井阪 悠太[5]

理研での所属研究室名：

[1]医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ 創薬バイオメディカルインテリジェンスユニット

[2]医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ

[3]医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ 分子設計インテリジェンスユニット

[4]医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ メディシナルケミストリーインテリジェンスユニット

[5]健康生き活き羅針盤リサーチコンプレックス推進プログラム 健康予測チーム

1. 本課題の研究の背景、目的、関係するプロジェクトとの関係

医薬品開発のためには多数のプロセスが存在し、承認薬はそのすべてを通過して初めて誕生する。承認までの 10～20 億 US＄とも見積もられているコストと 10 年以上という開発期間を削減するための手段として、各プロセスにおいて機械学習手法およびシミュレーションを利用することは検討の価値がある。本課題では、創薬開発のプロセスでの機械学習モデルの応用の可能性を調べるため、学習に必要となるデータの作成を実施している。

特に、本課題では「抗原・抗体の親和性の予測」、「分子動力学シミュレーションから生成されるトラジェクトリの解析」、「タンパク質構造予測」、「医薬品候補化合物の特性予測」に注目し、量子化学計算または分子動力学計算などによってデータを作成している

本課題のメンバーは、ライフサイエンス分野で機械学習モデルとビッグデータ利用を推進するコンソーシアム（ライフ インテリジェンス コンソーシアム（LINC）、代表：奥野恭史、事務局：京都大学大学院医学研究科人間健康科学系専攻ビッグデータ医科学分野、理化学研究所健康生き活き羅針盤リサーチコンプレックス推進プログラム、医薬基盤･健康･栄養研究所、公益財団法人 都市活力研究所）に所属し、各機関と連携の上で課題を実施している。

2. 具体的な利用内容、計算方法、結果

2-1-A. 抗原・抗体の親和性の予測に向けたデータ作成

2000 年以降、抗体医薬品が台頭をはじめ、今日では世界の医薬品の売り上げ Top10 のうち 5 個は抗体医薬品である。今後も、新しい抗体医薬品の開発に期待が寄せられている。抗体医薬品の開発プロセスでは、抗体分子の可溶性・安定性・抗原性などを調整するために、アミノ酸配列に変異を導入する。このときに、抗体の抗原に対する親和性を損なわない（もしくは向上させる）必要があるため、これを予測できると便利である。変異導入に伴う抗原・抗体の親和性変化（$\Delta\Delta G$）を予測するための、厳密な方法として自由エネルギー摂動法（FEP）の適用が考えられる。しかしながら、抗原・抗体複合体の$\Delta\Delta G$を 1 kcal/mol 程度の正確度で予測することは現時点では困難である。そこで、本課題ではまずFEP計算を$\Delta\Delta G$計算に適用するためのプロトコルを開発した。このプロトコルでは、抗原・抗体複合体の溶液中での構造アンサンブルを得ることができる。機械学習モデルへのインプットとして、この構造アンサンブルを利用することは検討の価値がある。

2-1-B. 変異導入に伴う抗原･抗体親和性変化の FEP 計算プロトコル

本課題では MPC-1（抗原）と 11KB (抗原)を抗原・抗体複合体のモデル系として採用し（複合体構造：PDBID: 2BDN）、重鎖 31 番のアスパラギンがグルタ

ミンに変異する場合のΔΔG計算をテーマとした。

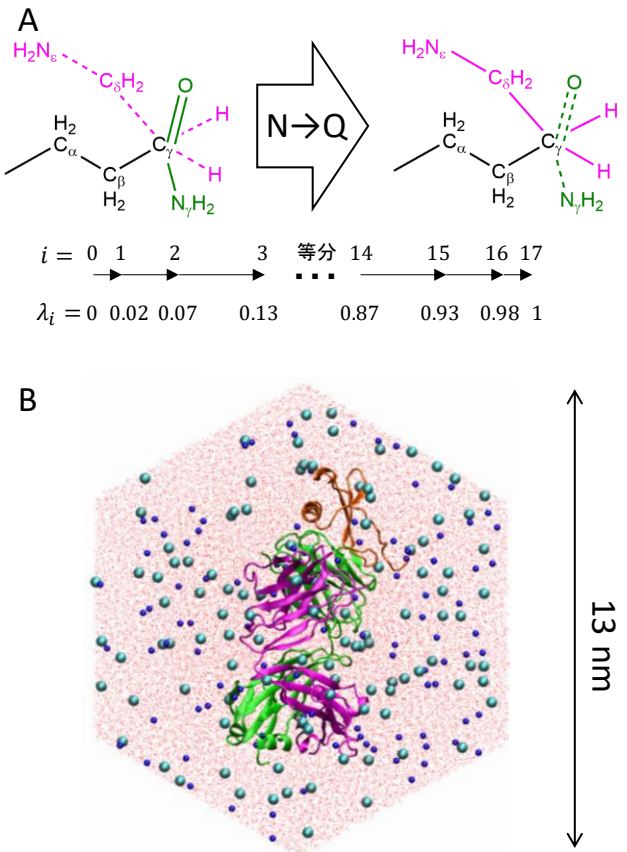

図 1. パラメータと初期構造

(A) Asn を Gln に変異させる場合を例として、pmx によって自動生成されたパラメータを模式的に示している。変異前（$\lambda_0 = 0$）は Asn のみ、変異後（$\lambda_{17} = 1$）は Gln のみのパラメータを採用し、中間状態（λ_1 –λ_{16}）では両者が混在するパラメータとなる。(B) 抗原抗体の複合体の初期構造。全 125571 原子、水 : 39257、Na^+: 106、Cl^-: 112 (150-mM NaCl 溶液）。抗原：オレンジ、重鎖：マジェンタ、軽鎖：緑、Na^+：青、 Cl^-：シアン、水分子：赤。

分子動力学計算（MD）エンジンとして GROMACS v.2016.3（Pronk et al., Bioinformatics 2013, 29, 845）を採用した。比較的新しい GROMACS は HOKUSAI の BigWaterfall に搭載されている Intel® Xeon® Gold 6148 CPU で計算を実行すると GPU を用いなくとも速い計算が可能である。変異導入に伴う FEP 計算に用いるパラメータファイルを作成するためには、pmx（Gapsys et al., J Comput Chem 2016, 36, 348）を使用した（変異前と変異後が混在した dual topology とよばれるパラメータ：図 1A）。PDB から取得した構造を dodecahedron 形のボックスに配置し周囲を 150 mM-NaCl 水溶液で満たすことで初期構造を作成した（図 1B）。作成したボックスに対して構造最適化、NVT、NPT アンサンブルによる平衡化計算を実施したのちに、100 ns の MD シミュレーションを 2 回実施し、構造サンプリングに利用した。

表 1. 結合ギブズエネルギー変化（ΔΔG）の計算結果と再現性の検証（kJ/mol）

	トラジェクトリ	Run1	Run2	Run1&2
ΔΔG	Forward@50 ns	41.3	42.0	41.6
	Reverse@50ns	43.0	43.2	42.9
	差	1.7	1.2	**1.3**

一回目と二回目のシミュレーション（Run1、Run2）の前半 50 ns（Forward@50ns）と後半 50 ns（Reverse@50ns）の構造アンサンブルを用いて FEP 計算を実施した値と Run1 と Run2 の構造アンサンブルを足し合わせた構造アンサンブルを用いて FEP 計算を実施した値を示している。Run1 と Run2 を足し合わせたアンサンブルでは前半 50 ns と後半 50 ns のΔΔGの差は 1.3 kJ/mol であった。

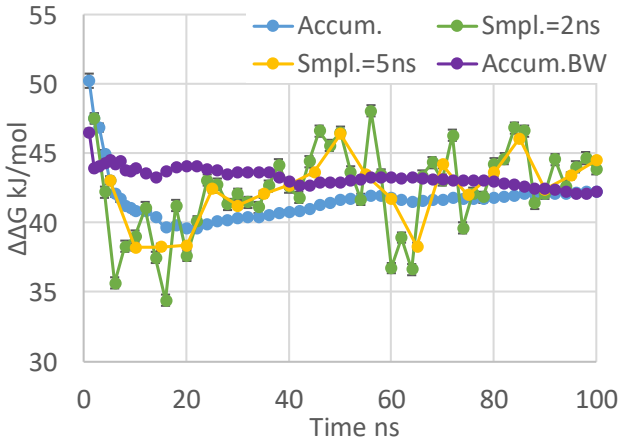

図 2. サンプリング領域とΔΔG の関係

サンプリング領域は以下の通りである。Accum. : [0, t], Smpl.=2 ns : [t–2, t], Smpl.=5 ns : [t–5, t], Accum.BW : [100–t, 100]。ここでは Run1 と Run2 の構造アンサンブルを合わせて FEP 計算を実施した。エラーバーは 68%信頼区間。

2 回の試行の結果として得られたΔΔG（表 1）によると、ΔΔGの再現性（精度）は 2 kJ/mol 以下であることが分かった。さらに、計算に使用するサンプリング領域を時間に対して順方向に増やしていった場合と、逆方向に増やしていった場合を比較することで、どの領域をサンプリングすれば、平衡状態の構造を取得でき

そうか、ある程度調べることができる。図 2 によると、最初の 10 ns 程度は$\Delta\Delta G$（Forward）が急激に変化するためまだ平衡化途中であると予想される。そのため、この領域は計算から除くことが望ましいことが示唆された。また、短いシミュレーション（2〜5 ns）のサンプリングでは、どのような領域を構造アンサンブルとして採用しても収束する値を得られないことがわかった。

2-2-A. 長時間 MD トラジェクトリに対する AI を利用した解析ツールの開発

タンパク質に対する長時間 MD の解析において、現状では研究者の主観に頼った解析となり、重要な情報を見落としている可能性は排除できない。しかしながら、客観的な解析をしようにも、MD トラジェクトリデータ自体が巨大であることから、網羅的な解析が困難である。また、長時間 MD とはいうものの、現状の計算資源ではナノ〜マイクロ秒スケール程度の挙動を追うのが限界であるため、ナノ秒スケール程度のトラジェクトリ情報から、長周期のダイナミクス予測のニーズもある。本課題においては、既存の AI パッケージを調査しながら現状で得られている MD トラジェクトリに対して本質的な情報を抽出できるような解析ツールを整備していく。

2-2-B. MD トラジェクトリデータをインプットとした解析ツールのプロトタイプ作成

MD トラジェクトリデータをインプットとして、Distance Matrix や二面角、RMSD などの記述子を算出できるスクリプトを開発した。このスクリプトをベースにした MD トラジェクトリ解析ツールのプロトタイプ構築において、Gini Importance の手法がトラジェクトリの解析において研究者の主観を排除する一つの手段として利用できることを見出し、これを利用した解析用スクリプトを作成した。

本課題では、MSM Builder に含まれる解析スクリプト群の中から、独立成分分析（tICA）や隠れマルコフモデル（HMM）などを用いて低次元に射影したダイナミクスの情報を教師データとして試験解析を実行するためのスクリプトの構築を行い、任意の MD トラジェクトリファイル（現段階では xtc 形式のみに対応）に対して、Gini Importance 解析を可能にした。

2-3. タンパク質立体構造予測

X 線結晶構造解析はタンパク質の詳細な構造を知る有力な手段であるが、低分解能のデータしか得られない場合、得られる構造の確度は低下する。本研究では低分解能データから確度の高い構造を機械学習により選び出すことを目的とする。

学習データを作成するため、Protein Data Bank に登録されている高分解能タンパク質の構造を正解とするような学習セットを作成するために、類縁タンパク質からのホモロジーモデリング技術を用いてタンパク質構造を多数発生した。今後は、X 線結晶構造解析用ソフトウェアなどのツールを用いて、これらのモデル構造の全体構造、部分構造を順次評価するとともに、この方法をより多くの構造に対して適用し学習データを蓄積する。得られた各種パラメータは記述子として構造予測モデルに学習させる予定である。

2-4. 医薬品候補化合物の特性予測のための深層学習モデル

医薬品開発プロセスにおいて、新しく候補となる化合物を検討する場合は薬理活性、物性、および毒性などの化合物特性を実験的に取得する必要がある。これらの特性を計算機上で予測できると便利である。そこで、本課題では化学構造式畳み込み手法を用いた深層学習プログラムを作成している。深層学習モデルを使用することによって、ネットワーク内部に構造式および原子・結合の特徴量から予測対象に適した記述子が学習の過程で生成され、従来法よりも予測性および解釈性が向上すると期待される。

まず、深層学習プログラム DeepChem を用いて、深層学習モデルを溶解度パラメータ予測に適用した。学習データおよびテストデータのためには Tox21 および PubChem AID1996 のデータ（合計 18013 種）を用いた。学習の結果、DeepChem のオリジナルの手法と同等以上の正確度での予測に成功している（ROC-AUC を用いて比較）。より正確度を向上させるために、原子特徴量、分子特徴量などのハイパーパラメータを変更できるようにプログラムを改変し、さらにハイパーパラメータ探索プログラムを作成した。今後はこれを用いて探索を実施する予定である。

3.　まとめ

本課題では、創薬プロセスの効率向上を目指して、いくつかのプロセスでの機械学習モデル作成およびデータ生成を実施している。今年度は、特に分子動力学シミュレーションを用いた抗原・抗体のアミノ酸変異導入に伴う結合ギブズエネルギー変化の計算において、再現性のある計算結果を得ることができる構造アンサンブルの作成方法を開発した。さらに、このようなシミュレーションから得られる大量のトラジェクトリを解析するための機械学習モデルの準備においても、解析上重要となる手法を見出した。また、分子動力学シミュレーションで必要となるタンパク質立体構造予測にも取組んでいる。

また、医薬品候補化合物の特性予測（溶解度パラメータ予測）に取組み、従来法と同等以上の予測正確度を得ることに成功した。

4.　今後の計画・展望

機械学習モデルのためのデータ作成とともに機械学習モデル開発の実施を進める。また、今回対象としなかったプロセスに関しても、機械学習モデルによる最適化の適用を検討する。