様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ９ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１２０６３

研究課題名（和文）苦味受容体におけるAI・シミュレーション・進化解析の融合解析フレームワークの構築

研究課題名（英文）Constructing a Fusion Framework of AI, Simulation, and Evolutionary Analysis in Bitter Taste Receptor

研究代表者
岩田　浩明（Iwata, Hiroaki）

京都大学・医学研究科・特定准教授

研究者番号：４０６１３３２８

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：データに乏しい状況において、学習データを生成しながらAIモデルを構築する取り組みが広く行われている。我々は、自己訓練に戻づく半教師付き学習フレームワークを提案した。まず、既知のデータを学習に用いて、AIモデルを構築する。次に、相互作用情報が未知の化合物-タンパク質ペアに疑似ラベルを生成し、学習データを増やし、モデルパラメータを改良していく。その結果、学習データのポジネガの不均衡が徐々に緩和され、さらに最終的に構築されたモデルは、既知の学習データセットのみを用いて構築された初期モデルを凌駕することが示された。

研究成果の学術的意義や社会的意義
様々な分野にAI技術が適用されてきており、成果が上がってきている。一方で、データが整備されていないことも多く、少量な学習データでよいモデルを作ることは学術的にも社会的にも求められている。本研究では、学習データを生成することで予測精度を高めることができた。今回は、化合物-タンパク質相互作用解析で手法の有用性を示したが、様々な分野で適用が可能である。学術的、社会的意義のある結果が得られた。

研究成果の概要（英文）：In situations where data is scarce, there have been widespread efforts to construct AI models while generating training data. We propose a semi-supervised learning framework based on self-training. First, known data is used for training to construct an AI model. Next, pseudo-labels are predicted for unknown compound-protein interaction information, thereby increasing the training data and improving the model parameters. As a result, the imbalance between positive and negative samples in the training data gradually diminishes, and furthermore, the final constructed model has been shown to surpass the original model constructed solely using the known training dataset.

研究分野：ケモインフォマティクス

キーワード：人工知能　シミュレーション　進化　苦味　ディープラーニング　分子動力学シミュレーション

１．研究開始当初の背景

苦味は、5 種類の味覚のうちの 1 つであり、口腔の味細胞にある苦味受容体によって感じ、毒物を苦味として受容し、身体を毒から守っている。苦味受容体は GPCR の一種であり、TAS2R 遺伝子によってコードされている。ヒトでは 25 種類の苦味受容体が存在しており、チンパンジーでは 28 種存在する。マウスは、35 種存在しており、種間のレパートリー数の差異が高く、また種内でも多様性を持っている。ファミリーを形成している苦味受容体であるがヒトのみならず、他生物種においても立体構造は 1 つも解明されていない。そのため、一般的な立体構造に基づいた解析が行えない。また、1 つの分子を複数の受容体が認識することもわかっており、同様に、1 つの受容体が複数の分子を認識することもわかっている。苦味分子と受容体は多対多の関係にあることも解析を困難にしている。

なぜ苦味受容体がファミリーを形成することになったか、苦味受容体の進化過程のなぞ、苦味受容体が苦味を認識するメカニズムの解明、など多くの謎が依然とある。リガンドがわかっていないオーファン受容体が多い。AI とシミュレーションを組み合わせた次世代のコンピュータ解析のフレームワークでこれらの学術的問いを解明されると期待されている。

２．研究の目的

苦味受容体に代表されるデータに乏しい状況において、AI とシミュレーションを組み合わせることでデータを生成しながら AI モデルを構築する取り組みが広く行われている。苦味受容体でも、オーファン受容体が多いことからも既知リガンド情報が少ない。本研究では、シミュレーションデータ等を加えて学習データを拡張し、精度のよい予測 AI モデルの構築手法を開発した。さらに、予測 AI モデルの精度を上げるため、効率的に特定のターゲットに関連する化合物を提案する手法も開発した。

３．研究の方法

本研究では、データに乏しい状況におけるデータ拡張手法として、シミュレーションを実施することで学習データ量を増やす自己学習によるデータ拡張手法と、ターゲットに活性がある化合物の化学構造を提案する新規化合物提案手法を開発した。

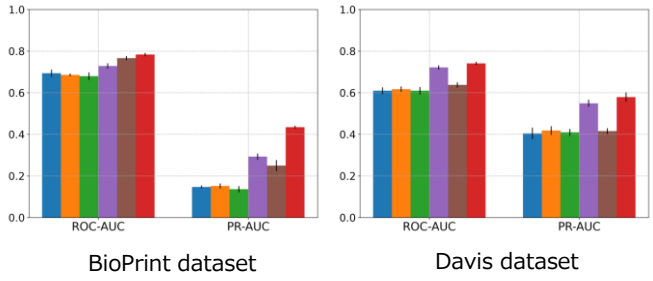
自己学習によるデータ拡張手法では、既知の学習データセットで構築した AI モデルを用いて、未知の化合物-タンパク質のペアを予測し、それを学習データに加えて AI モデルを構築する。さらにそのモデルを使って未知データを予測して、さらに学習データに加える。これを繰り返し行い、学習データを増やして AI モデルの精度を高めていく手法である。

新規化合物提案手法では、深層学習モデルである Variational Graph Auto-Encoder（VGAE）と強化学習モデルである Monte Carlo Tree Search（MCTS）を組み合わせた新たな分子生成モデル（VGAE-MCTS と呼ぶ）を開発した。

４．研究成果

①自己学習によるデータ拡張手法

構築した AI モデルの一般性を評価するため、外部データセットである GPCR ファミリーの BioPrint とキナーゼファミリーの Davis で AI モデルの性能を比較した。ここで、主要な指標としては PR-AUC スコアを採用した。これは、ネガティブサンプルが多い不均衡なデータセットでのモデル性能を評価するのに適している。BioPrint および Davis データセットでは、PR-AUC スコアがそれぞれ 0.4344 および 0.5792 となり、ベースラインモードよりも 28.7%および 17.5%向上し、モデルの性能が改善された。自己学習によるデータ拡張方法は、データの不均衡に対処するために、他の方法よりも優れた性能を発揮した。2 番目に優れた Random Negative モデルの PR-AUC スコアは、0.2927 と 0.5491 であった。これらの結果は、我々の手法が外部データセットにおけるモデルの頑健性を大幅に改善できることを示した。また、他のモデルは片方に大きく偏っていたため、Precision と Recall の調和平均である F1 スコアでも我々の方法は他のモデルを上回り、優位性を示せた。



②新規化合物提案手法

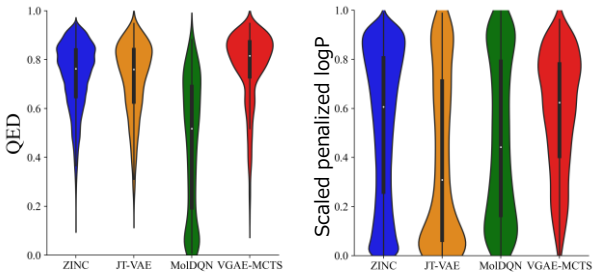
化合物提案生成モデルの基本性能を評価する指標である validity、uniqueness、novelty を含む GuacaMol フレームワークの中の Distribution-Learning Benchmarks を用いて我々の新規化合物提案手法（VGAE-MCTS）の評価を行った。VGAE-MCTS で生成した分子は validity、uniqueness、novelty のスコアは 1.000 を示した。すなわち、生成された分子はすべて価電子数

が守られて理論的にあり得る分子、生成された分子に重複がなく、学習データセットの中には存在しない新規の分子であった。これら 3 つの指標については先行研究のモデルと同等以上の結果となった。VGAE-MCTS の KL divergence のスコアは 0.659 となった。先行研究のモデルである Graph MCTS や VGAE の中で最も高い結果となった。VGAE-MCTS の FCD のスコアは 0.009 となった。FCD は、薬様の特性を持つ生理活性分子が登録されている ChEMBL 中の分子と生成した分子との生理活性予測値の分布の類似度を比較している。VGAE-MCTS の FCD のスコアは他のモデルと同様にスコアが低かった。

	GraphMCTS	VGAE	VGAE-MCTS
validity	1.000	0.830	1.000
uniqueness	1.000	0.944	1.000
novelty	0.994	1.000	1.000
KL divergence	0.522	0.554	0.659
FCD	0.015	0.016	0.009

VGAE-MCTS の物性値を最適化した際の分子生成能について評価を行った。最適化する物性値は、Quantitative Estimate of Drug-likeness（QED）と Penalized logP である。QED は薬らしさを定量的に測る指標であり、Penalized logP は脂溶性、合成のしやすさ、大きい環に対するペナルティの 3 つの物性値を組み合わせた指標である。

まず、QED 最適化においては、ZINC から抽出した学習データセットの分子（平均：0.732、中央値：0.762）と比較したところ、VGAE-MCTS で生成した分子（平均値：0.772、中央値：0.815）は明らかに高スコアの分子が生成できていた（Mann-Whitney U test: P=2.52×10-7）。つまり、VGAE-MCTS が MCTS による探索において創薬物性値 QED が良くなる方へ分子を展開できていることが示唆される。また、先行手法である JT-VAE（平均：0.720、中央値：0.760）や MolDQN（平均：0.455、中央値：0.518）と比較しても、高スコアの分子を生成できていた（Mann-Whitney U test: $P=4.77\times10^{-11}$, <0.01）。次に、Penalized logP の最適化においては、学習データセットの分子（平均：0.536、中央値：0.606）と比較したところ、VGAE-MCTS で生成した分子（平均値：0.572、中央値：0.610）は高スコアの分子が生成できていた（Mann-Whitney U test: P=0.6077）。加えて、VGAE-MCTS は、ZINC データセットの分子や先行手法のモデルで生成した分子よりも Penalized logP の値が小さい分子を生成している割合が少ないことが分かった。これは、VGAE-MCTS が MCTS による探索において Penalized logP が良くない方へ分子を展開することを避けていることが示唆される。また、JT-VAE（平均：0.392、中央値：0.309）や MolDQN（平均：0.472、中央値：0.442）と比較しても、高スコアの分子を生成できた（Mann-Whitney U test: $P=4.26\times10^{-14}$, 1.9×10^{-3}）。



＜今後の展望＞

様々な分野に AI 技術が適用されてきており、成果が上がってきている。一方で、データが整備されていないことも多く、少量な学習データでよいモデルを作ることは学術的にも社会的にも求められている。本研究で開発した自己学習によるデータ拡張法や新規分子提案手法は、学習データを生成することで予測精度を高めることがでる。今後は、様々な分野にこれらの手法を適用していきたい。

＜引用文献＞

① Takuto Koyama, Shigeyuki Matsumoto*, Hiroaki Iwata, Ryosuke Kojima, Yasushi Okuno*, “Improving Compound-Protein Interaction Prediction by Self-Training with Augmenting Negative Samples,” *ChemRxiv*.

② Hiroaki Iwata*†, Taichi Nakai†, Takuto Koyama, Shigeyuki Matsumoto, Ryosuke Kojima, Yasushi Okuno*, “VGAE-MCTS: a New Molecular Generative Model combining Variational Graph Auto-Encoder and Monte Carlo Tree Search,” *ChemRxiv.*

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Iwata Hiroaki	４．巻 71
２．論文標題 Application of <i>in Silico</i> Technologies for Drug Target Discovery and Pharmacokinetic Analysis	５．発行年 2023年
３．雑誌名 Chemical and Pharmaceutical Bulletin	６．最初と最後の頁 398～405
掲載論文のDOI（デジタルオブジェクト識別子） 10.1248/cpb.c22-00638	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Iwata Hiroaki、Matsuo Tatsuru、Mamada Hideaki、Motomura Takahisa、Matsushita Mayumi、Fujiwara Takeshi、Kazuya Maeda、Handa Koichi	４．巻 110
２．論文標題 Prediction of Total Drug Clearance in Humans Using Animal Data: Proposal of a Multimodal Learning Method Based on Deep Learning	５．発行年 2021年
３．雑誌名 Journal of Pharmaceutical Sciences	６．最初と最後の頁 1834～1841
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.xphs.2021.01.020	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 Hiroaki Iwata and Yasushi Okuno
２．発表標題 Machine learning-based method for integrating phenotypic and target-based approaches in drug discovery
３．学会等名 2021 International Chemical Congress of Pacific Basin Societies（国際学会）
４．発表年 2021年

１．発表者名 岩田浩明
２．発表標題 創薬開発プロセスへのIT技術適用
３．学会等名 一般社団法人製剤機械技術学会，第30回講演会（招待講演）
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関