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　　元 年　　６ 月 １４ 日現在

機関番号：８２６２６
研究種目：研究活動スタート支援
研究期間：2017～2018
課題番号：１７Ｈ０７３９２

研究課題名（和文）深層学習を用いた化合物とタンパク質の表現学習と創薬への応用

研究課題名（英文）Deep representation learning for drugs and proteins with neural networks

研究代表者
椿　真史（Tsubaki, Masashi）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：８０８０３８７４

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：機械学習分野における最大の国際会議である、Advances in Neural Information Processing Systems (NIPS 2017) のワークショップでベストペーパー賞を受賞した。また、バイオインフォマティクス分野の国際ジャーナルであるBioinformaticsに論文が採択された。同時に、開発したソフトウエアを一般に公開した。さらにこの成果に基づき、製薬会社と共同研究を行うことになった。基礎研究の部分である手法の考案から、論文採択とソフトウエアの公開、そしてそこから企業との共同研究という、産業応用までの一連の流れを作ることができた。

研究成果の学術的意義や社会的意義
この研究成果の学術的な意義としては、まず、グラフ構造のような離散データについても、深層学習の有効性を検証できたという点である。特に、これまで特徴量や記述子を使って、データを一旦変換した上で、つまり情報を人間の観点から削減した上で機械学習手法を適用していたものが、データのより原始的な情報を入力として扱えるようになった。また、社会的な意義としては、これまで新薬の開発が難しかった病気などに対して、コンピュータのアプローチから迫ることができる点である。特に、機械学習手法はシュミレーションなどの異なり、予測の精度が非常に速いことが大きな利点である。

研究成果の概要（英文）：Received the Best Paper Award at the workshop of Advances in Neural Information Processing Systems (NIPS 2017), the largest international conference in machine learning. It was also adopted by Bioinformatics, an international journal in the field of bioinformatics. At the same time, the developed software was released to the public. In fact, it was decided to conduct joint research with a pharmaceutical company. I was able to create a series of flows from industrial design to the application of the method which is a part of the foundation, to the dissertation and the release of software, and from there to joint research with companies, to industrial applications.

研究分野：機械学習

キーワード：深層学習　創薬

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

機械学習・人工知能（AI）技術による創薬研究の一つとして、深層学習を用いた創薬研究に取り組んだ。創薬の分野においては、実際これまで、様々な機械学習手法が使われており一定の成果も収めている。例えば、既存法では薬剤候補となる化合物の特徴量・記述子を人手で考えて設計し、SVM などの機械学習手法を適用する手法が代表的である。このような手法は一般に浸透している一方で、多くの解決すべき問題も生じた。その問題の中でも重要なのが、薬剤データをどのようにコンピュータ上で表現するかということである。最も基本的なものは、上に述べた創薬の専門家が考えて設計した薬剤の特徴量・記述子である。しかしこのような情報は、どのような薬を開発するかによって、また専門家の知識によって変わるため様々な問題に応じて変える必要があり、確実と言えるものがあるかもわからないことも多い。似たような問題は、画像処理や言語処理でもあったが、近年の深層学習では、特徴量自体をデータから自動的に学習できることが示されている。そこで、深層学習によって創薬研究における特徴量設計の問題も、解決できるのではないかと考えた。

２．研究の目的

特徴量や記述子を人手で設計することなく、end-to-end で化合物の薬剤活性を学習・予測できることを目的とした。図１にその全体の流れを示す。このような手法を開発する理由として、化合物の構造からその薬剤活性を予測するというのは、非常に複雑な関係（関数）で記述される現象であるため、人手で入力から関数まですべてを設計することは非常に困難であり、またできたとしても創薬に関して詳しい研究者がいないとできない。このような問題を、大規模な薬剤データを使って、そこから特徴量・記述子を自動的に獲得することを目標とし、より汎用的に使えるソフトウエアとして提供できるようにすることを目的とする。そのためにはまず、既存の創薬研究者の考えた特徴量を用いた機械学習手法と、自分の開発した手法との精度や速度などを比較して、その有効性を検証する。

３．研究の方法

情報科学としては、化合物データはグラフ構造で表現できる。このことから、離散構造データに対する機械学習・深層学習手法が使えることがわかる。特に近年は、離散構造データの中でもグラフ構造に対する深層学習手法の研究・開発が進んでおり、それを上記の end-to-end に拡張することを手法のメインとした。化合物グラフの構造を適切に組み込んだアルゴリズムを考えた。特に、創薬分野で広く用いられている、化合物の部分構造をモデルの中に組み込んで、それを学習するような最適化アルゴリズムにした（図２）。また、創薬研究では、化合物だけではなくタンパク質の配列や立体構造も入力とする、つまり２つの異なる構造を持つデータを同時に扱う必要があり、そのような手法を開発した（図３）。そしてその手法を用いて、大規模なデータセットを使って、実装・実験を行った。

４．研究成果

まずは、化合物のみを入力とした薬剤活性予測手法については、機械学習分野における最大の国際会議である、Advances in Neural Information Processing Systems (NIPS 2017)のワークショップである、Machine Learning for Molecules and Materials で発表し、そこでベストペーパー賞を受賞した。また引き続き、グラフ構造と配列構造の 2 つの深層学習手法を組み合わせて、高精度かつ高速な薬剤スクリーニング手法を開発した。この成果は、バイオインフォマティクス分野の国際ジャーナルである Bioinformatics に採択された。同時に、開発したソフトウエアを一般に公開した。さらに、この成果を基に、製薬会社などの企業、大学、研究機関の集まる創薬インフォマティクス研究会において、講演を行った。この講演と、開発・公開したソフトウエアをきっかけに、実際に製薬会社と共同研究を行うことになった。論文執筆時は、創薬のパブリックなベンチマークデータセットを用いて手法の評価を行うだけに留まっていたが、製薬会社では実データを用いてより実践的に手法の評価、さらにはそのアップデートが可能となると考えている。このように研究実績としては、基礎の部分である手法の考案から行い、論文採択とソフトウエアの公開、そしてそこから企業との共同研究という、産業応用までの一連の流れを作ることができた。

５．主な発表論文等

〔雑誌論文〕（計　4　件）

1, Masashi Tsubaki, Kentaro Tomii, and Jun Sese, Compound-protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences, Bioinformatics.

2, Tatsuro Kawamoto, Masashi Tsubaki, and Tomoyuki Obuchi, Mean-field theory of Graph Neural Networks in Graph Partitioning, Advances in Neural Information Processing Systems.

3, Masashi Tsubaki and Teruyasu Mizoguchi, Fast and Accurate Molecular Property Prediction: Learning Atomic Interactions and Potentials with Neural Networks, The Journal of Physical Chemistry Letters.

4, Shin Kiyohara, Masashi Tsubaki, Kunyen Liao, and Teruyasu Mizoguchi, Quantitative estimation of properties from core-loss spectrum via neural network, Journal of Physics: Materials.

〔学会発表〕（計　2　件）

1, Masashi Tsubaki, Masashi Shimbo, Atsunori Kanemura, and Hideki Asoh, End-to-end Learning of Graph Neural Networks for Latent Molecular Representations, Advances in Neural Information Processing Systems (NIPS 2017) Workshop, Machine Learning for Molecules and Materials, Best paper award.

2, 椿真史,
深層学習を用いた化合物とタンパク質の相互作用予測,
創薬インフォマティクス研究会（招待講演）

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

グラフ構造データに対するニューラルネットワークの実装
https://github.com/masashitsubaki/GNN_molecules

化合物タンパク質相互作用予測ソフトウエア
https://github.com/masashitsubaki/CPI_prediction

化合物物性値（薬剤活性やエネルギー等）予測ソフトウエア
https://github.com/masashitsubaki/QuantumGNN_molecules

６．研究組織

(1)研究分担者

研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者

研究協力者氏名：麻生英樹
ローマ字氏名： Hideki Asho

研究協力者氏名：兼村厚範
ローマ字氏名： Atsunori Kanemura

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。