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　４ 年　 ５ 月 ２４ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１８４８

研究課題名（和文）統計的グラフ解析による分子構造の解析と応用

研究課題名（英文）Molecular Structure Analysis using Statistic Graph Model and Its Applications

研究代表者
宮崎　智（Miyazaki, Tomo）

東北大学・工学研究科・助教

研究者番号：１０７５５１０１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：情報通信の発達により、インターネット上に様々な大規模分子科学データが構築されている。これら大規模な分子データから有用な知識（例えば薬理活性を決定する分子構造）を発見することが期待されている。しかし、その発見は一般的に専門家に頼っており、大規模データからの重要構造の自動抽出は挑戦的な課題である。本研究は、大規模な分子データを基に、分子の化学的特性を決める重要な構造を統計的に自動解析する手法を開発した。主な成果として、深層学習による分子の化学的特性を高精度に推定する手法を開発し、物理化学などの分子の解析に有効であることを示した。さらに、分子解析法を画像処理に活用して、魚の脂肪率推定を行った。

研究成果の学術的意義や社会的意義
本研究の成果は、分子の化学的特性を特徴づける構造を統計的に自動抽出するアルゴリズムを確立した点である。これまで専門家が人手で行っていた分子構造の発見を機械学習により自動化することで、客観的かつ高効率な発見が可能となった。これにより、客観的でより信頼性が高く、多様な分子構造を極めて効率よく抽出することが期待できる。大規模な分子データベースが利用可能となっている現在、本研究成果である統計的な解析手法により、今まで人間が想像もしなかったような新たな知見（構造）を発掘できる可能性がある。よって本研究成果は、インターネット上に構築された大規模分子データの統計解析の実用化に大きく貢献するものである。

研究成果の概要（英文）：With the development of information and communications, various large-scale molecular science data have been constructed on the Internet. It is expected that useful knowledge (e.g., molecular structures that determine pharmacological activity) can be discovered from these large-scale molecular data. However, such discovery generally relies on experts, and the automatic extraction of essential structures from large-scale data is challenging. In this study, we developed a method to statistically and automatically analyze important structures that determine the chemical properties of molecules based on large-scale molecular data. As a major achievement, we developed a method to estimate the chemical properties of molecules with high accuracy by deep learning and showed that the method is effective for analyzing molecules in physical chemistry and other fields. Furthermore, the molecular analysis method was applied to image processing to estimate the fat content of fish.

研究分野：パターン認識

キーワード：化学特性推定　グラフ解析　分子特性認識　グラフマッチング　統計的グラフモデル　構造データ解析　分子解析　部分構造

１．研究開始当初の背景

分子の化学的特性を決定づける部分構造は薬理活性の予測などに用いられる重要な特徴であるが[①]、その発見は一般的には専門家に頼っており、統計的手法による分子の構造解析や重要な部分構造の自動抽出は未だ挑戦的な課題である。これは、分子が元素とその結合で表されるシンボリックなデータ構造であるため、主にベクトル形式のデータを扱う統計手法とは相性が悪かったことに起因している。したがって、これまで統計学的手法による真の意味での客観的な分子構造の解析は実現できていなかった。

データ科学を活用した分子解析の分野では様々な研究が行われているが、分子を数値ベクトルに変換して扱うものが多く、変換の過程で構造情報が失われる場合が多い。数えあげにより構造を抽出する手法[②]もあるが、部分構造の組み合わせは膨大であるため、人間の知識に基づいた選別処理が必須である。分子の化学的特性はどのような部分構造によって特徴付けられるのであろうか。大規模な分子データベースが利用可能となっている現在、統計的な解析手法により、今まで人間が想像もしなかったような新たな知見（構造）を発掘できる可能性がある。

２．研究の目的

機械学習により分子の化学的特性を特徴づける本質的な部分構造を抽出する手法を開発することを本研究の目的とする。分子の化学的特性を特徴づける部分構造は現在人手により発見されているが、人間の主観が入る上、高度な専門知識を必要とし、膨大な手間と時間がかかり、調査対象となる分子データ数にも限界があるといった問題がある。これを機械学習により自動化することで、膨大な分子データから予測や解析に適した構造を抽出することが可能になり、分子特性の解析に大いに貢献できる基盤技術となることが期待できる。さらに、分子解析手法を画像処理の諸問題に応用することで、機械学習によるグラフパターン解析技術の深化を目指す。

３．研究の方法

(1)　深層学習を用いた分子の化学的特性の推定手法の開発[③]を行なった。近年、深層学習は分子解析にも活用されている。しかし、既存手法は分子のある特定の特徴のみを用いるという問題があった。そこで本研究では、「複数モデルによる分子解析」を深層学習に組み入れた。これにより、分子から幅広い特徴抽出が可能となった。本手法は、複数特徴を同時に考慮できる点が特徴的である。具体的には図１に示すように、三つの処理手順に分岐した後に統合することで、複数の特徴を同時に考慮しながら、化学的特性を推定できる。Node Path は分子の原子の特徴に着目し、Edge Path は結合種類、3D Path は三次元空間における分子の座標をそれぞれ集中的に考慮して特徴を抽出した。最終的に、全結合層により特性値を推定した。

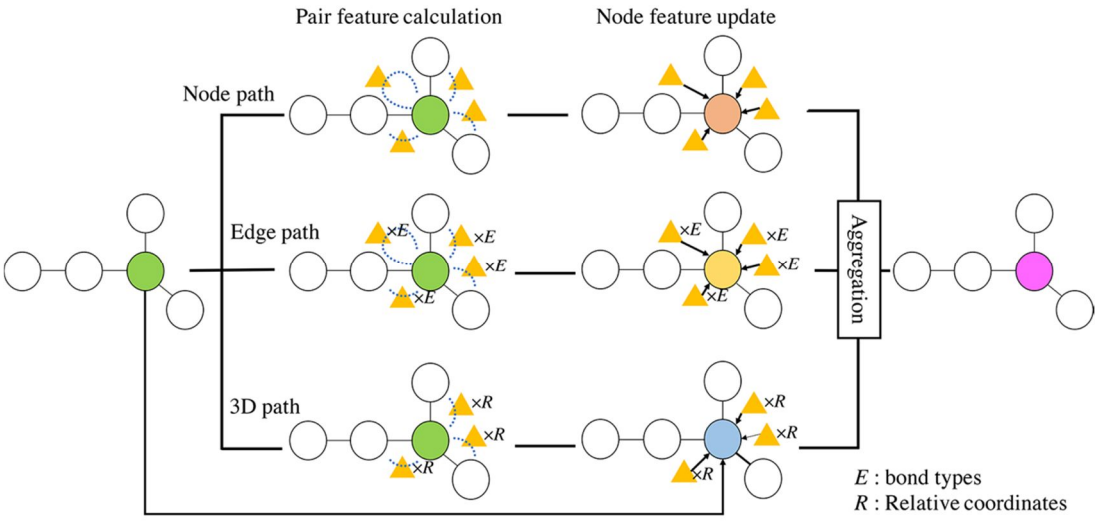

図１．深層学習を用いた分子の化学的特性の首位定手法

(2) 開発した分子解析法の独創点である「さまざまな特徴を同時に考慮する」というアイデアを画像処理のタスクに応用する研究を行った。具体的には、テキストによる動画検索[④]と魚画像の脂肪率推定[⑤]を行った。動画検索では、動画像とテキストを複数の埋め込み空間（複数の特徴を表す空間）にマッピングすることで、複数の特徴から動画とテキストの関連づける手法を開発した。魚画像の脂肪率推定では、魚の全体と局部（頭部、腹部、ヒレ部）から特徴を抽出した。さらに、RGB 画像に加えて深度画像を用いることで、魚の三次元的な形状も考慮して推定した。

４．研究成果

(1)　分子の化学的特性の推定では主に Freesolv と ESOL の二つの分子データセットを用いて、提案手法を評価した。Freesolv と ESOL ではそれぞれ水分子の自由エネルギーと溶解度を推定した[③]。表１の評価結果から、それぞれの特徴を単独に用いた場合と比較して、すべての特徴を統合する提案手法が最も良い評価結果を得た。他にも、様々な分子（量子力学から物理化学、生物物理学､生理学)を用いて評価した結果､提案手法が最も高精度に推定できることを確認した。この結果は様々な構造特徴を同時に考慮することの有効性を示しており、幅広い特徴を基に分子の統計解析ができることを明らかにした点が特徴的である。

表１．化学的特性の推定結果（数値は平均絶対誤差で、小さいほど良い。）

	原子特徴に着目	結合特徴に着目	三次元特徴に着目	全特徴を統合（提案手法）
Freesolv	0.764	0.817	0.743	0.717
ESOL	0.503	0.665	0.531	0.498

よって、本研究の成果は、分子の化学的特性を特徴づける構造を統計的に自動抽出するアルゴリズムを確立した点である。これまで専門家が人手で行っていた分子構造の発見を機械学習により自動化することで、客観的かつ高効率な発見が可能となった。これにより、客観的でより信頼性が高く、多様な分子構造を極めて効率よく抽出することが期待できる。大規模な分子データベースが利用可能となっている現在、本研究成果である統計的な解析手法により、今まで人間が想像もしなかったような新たな知見（構造）を発掘できる可能性がある。

(2)　テキストによる動画像検索の実験結果を表２に示す。埋め込み空間数に応じて検索結果も向上しており、異なる複数の特徴を用いることの有用性を示している。提案手法は埋め込み空間を容易に追加することができる。よって、テキストに加えて、音声や動画中の物体などの関係性を同時に考慮することで更なる性能向上が見込まれる。また、魚画像の脂肪率推定の結果を表３に示す。提案手法は全ての既存手法を上回る推定精度を達成しており、複数の特徴を組み合わせることで、魚画像から脂肪率を高精度に推定できることが明らかとなった。

表２．動画の検索結果（再現率は大きいほどよい。中央値は小さいほど良い）

埋め込み空間数	再現率（1位）	再現率(5位以内)	再現率(10位以内)	検索順位の中央値
1	5.6	18.4	28.3	41
2	6.8	20.2	30.6	30
3	6.7	21.2	32.4	29

表３．魚画像の脂肪率推定の結果（MAE、RMSE は小さいほど、その他は大きいほど良い。）

	MAE	RMSE	R2	Correlation
SVR	3.82	4.73	0.18	0.45
Random Forest	3.81	4.65	0.21	0.46
VGG16	2.54	3.38	0.58	0.77
VGG19	2.50	3.30	0.61	0.79
提案手法	2.25	2.91	0.69	0.83

＜引用文献＞

① Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, Aron Walsh, “Machine Learning for Molecular and Materials Science,” Nature, 559, pp. 547-555, 2018

② David Rogers, Mathew Hahn, “Extended-Connectivity Fingerprints,” Journal of Chemical Information and Modeling, 50, 5, pp.742-754, 2010

③ Sho Ishida, Tomo Miyazaki, Yoshihiro Sugaya, Shinichiro Omachi, “Graph Neural Networks with Multiple Feature Extraction Paths for Chemical Property Estimation,” Molecules, vol.26, no.11, 3125, 2021

④ Huy Manh Nguyen, Tomo Miyazaki, Yoshihiro Sugaya, Shinichiro Omachi,“Multiple Visual-Semantic Embedding for Video Retrieval from Query Sentence,” Applied Sciences, vol.11, no.7, 3214, 2021

⑤ Shuya Sano, Tomo Miyazaki, Yoshihiro Sugaya, Shinichiro Omachi,”Mackerel Fat Content Estimation using RGB and Depth Images,” IEEE Access, vol.9, pp.164060-164069, 2021

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Nguyen Huy Manh、Miyazaki Tomo、Sugaya Yoshihiro、Omachi Shinichiro	４．巻 11
２．論文標題 Multiple Visual-Semantic Embedding for Video Retrieval from Query Sentence	５．発行年 2021年
３．雑誌名 Applied Sciences	６．最初と最後の頁 3214～3214
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/app11073214	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 Ishida Sho、Miyazaki Tomo、Sugaya Yoshihiro、Omachi Shinichiro	４．巻 26
２．論文標題 Graph Neural Networks with Multiple Feature Extraction Paths for Chemical Property Estimation	５．発行年 2021年
３．雑誌名 Molecules	６．最初と最後の頁 3125～3125
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/molecules26113125	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 Sano Shuya、Miyazaki Tomo、Sugaya Yoshihiro、Sekiguchi Naohiro、Omachi Shinichiro	４．巻 9
２．論文標題 Mackerel Fat Content Estimation Using RGB and Depth Images	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 164060～164069
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3134260	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1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 石田 聖，宮崎 智，菅谷至寛，大町真一郎
２．発表標題 グラフの構造的特徴を考慮した分子特性の認識
３．学会等名 2019年度電気関係学会東北支部連合大会
４．発表年 2019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