Machine Learning for Metabolite Identification with Mass Spectrometry Data(Abstract_要旨)

NGUYEN, DAI HAI

京都大学	博　士（　薬科学　）	氏 名	NGUYEN DAI HAI
論文題目	Machine Learning for Metabolite Identification with Mass Spectrometry Data （質量分析データによる代謝産物識別のための機械学習手法構築）		

Metabolites are small molecules which are used in, or created by, chemical reactions occurring in living organism. They play important functions such as energy transport, signaling, building block of cells and inhibition/catalysis. Understanding biochemical characteristics (or identification) of metabolites is an essential part of metabolomics to enlarge the knowledge of biological systems. It is also key to the development of many applications and areas such as biotechnology, biomedicine or pharmaceutical sciences. However, this still remains a challenging task with a huge number of potentially interesting but unknown metabolites. Mass Spectrometry is a common analytical technique that measures the mass-to-charge ratio of ions converted from a portion of a chemical sample. The results are typically presented as a mass spectrum, a plot of intensity as a function of mass-to-charge ratio. Another way to represent a mass spectrum is as a list of peaks, each is defined by its mass-to-charge ratio and intensity value.

Identification of metabolites based on mass spectra can be regarded as a retrieval task: given a query spectrum of an unknown molecule, we aim to find molecules which have similar spectra from a reference database. A traditional approach is to compare the query against reference spectra in the database. The candidate molecules from the reference database are ranked based on the similarity between their reference spectra and the query, and the best matched candidates are returned. However, the reference databases often contain spectra of a small fraction of molecules in reality, leading to unreliable matching results if the molecule of query spectrum is not in the reference database. Consequently, to mitigate the insufficiency of such databases, alternative approaches for the task are devised. In this thesis, we explore computational methods for metabolite identification from spectra data with a focus on machine learning (ML), which has two stages: (i) mapping a spectrum to an intermediate representation (usually a molecular fingerprint, which is a binary vector to encode the presence of predetermined substructures or chemical properties in a molecule) and (ii) retrieving candidate molecules from the reference database. The contributions of this thesis include: 1) we present a comprehensive survey on recent advances and prospects of computational methods for metabolite identification from mass spectra with an emphasis on ML approach; 2) we present SIMPLE, a method for predicting molecular fingerprints from spectra with ability to explicitly incorporate peak interactions and has interpretation, which are not addressed by the current cutting-edge methods for fingerprint prediction (stage (i)); 3) we present ADAPTIVE, a method for predicting chemical structures from spectra through learnable intermediate representations to overcome the drawbacks of molecular fingerprints: being very large to cover all possible substructures and redundant. We summarize each topic below in more detail.

In Chapter 1, we thoroughly survey computational methods for metabolite identification from mass spectra. The primary purpose of this survey is not only to summarize the proposed techniques in literature, but also to systematically organize them into groups according to their methodology and approaches. It would be beneficial for researchers to comprehend the key differences between techniques as well as rationale behind their groupings. We grouped computational techniques for the task into the following main categories: 1) spectra library; 2) *in silico* fragmentation and 3) ML. Given a query spectrum, spectra library is to compare it against a database of reference spectra of known molecules and rank the candidates based on their similarity to the query. In contrast, *in silico* fragmentation attempts to generate simulated spectra from the chemical structures in a compound database and then compare them with the query spectrum. ML is to predict intermediate representations between spectra and chemical structures of compounds and then use such representations for matching or retrieval. Our research focuses on developing ML models for predicting the intermediate representations with high accuracy and interpretation.

In Chapter 2, we present SIMPLE, a sparse learning based tool for fingerprint prediction. It takes a query spectrum of an unknown molecule as an input and predicts binary fingerprints as output, indicating which substructures or chemical properties are present in the molecule corresponding to the query spectrum. We then can use these predicted fingerprints to query candidate molecules with most similar fingerprints in the reference database. SIMPLE achieved around accuracy of 78.86%, which was comparable to the top-performance kernel based methods, which achieved around 76-80%, obtained by 10-fold cross validation on the MassBank dataset with 402 spectra. On the other hand, these kernel based methods needed around 1500 milliseconds, which is more than 300 times slower than that of SIMPLE, which required less than 5 milliseconds on the same dataset. This is a sizable difference when we process a huge amount of spectra produced by the current high-throughput mass spectrometry. One advantage of sparse learning models over kernel based methods is interpretation. SIMPLE clearly revealed individual peaks and peak interactions that contribute to enhancing the performance of predicting a particular fingerprint, shown by some case studies. In more technical detail, we formulate a sparse interaction model for spectra data. The model encourages sparsity over peaks and low-rankness over peak interactions while minimizing the classification errors for predicting the presence of fingerprints. The formulation of model is convex and guarantees global optimization, for which we develop an alternating direction method of multipliers algorithm.

In Chapter 3, we present ADAPTIVE, a tool for metabolite identification with learnable intermediate representations from given pairs of spectra and corresponding chemical structures of known molecules. It takes a spectrum of an unknown molecule as input and outputs a list of candidate compounds from the reference database. Instead of using fingerprints as in existing methods, ADAPTIVE could learn intermediate representations (called molecular vectors) between spectra and chemical structures of compounds. The benefits of learning molecular vectors are: 1) specific to both given data and task of metabolite identification and 2) more compact than molecular fingerprints, leading to a significant improvement in terms of both predictive performance and computational efficiency. ADAPTIVE with the molecular vector size of 300 achieved top-10 and -20 accuracies of 71.1% and 78.52%, which are 4% and 5% higher than those of the current best method, input output kernel regression (IOKR), respectively, obtained by 10-fold cross validation on a benchmark dataset with 4138 spectra. Furthermore, ADAPTIVE took 1000 milliseconds for retrieving one spectrum, while IOKR needed more than 3000 milliseconds on the same dataset, meaning that ADAPTIVE was three times faster than IOKR. Technically, ADAPTIVE has two parts for learning two mappings: (i) from chemical structures to molecular vectors; (ii) from spectra to molecular vectors. The first part learns molecular vectors for molecular structures by maximizing the correlation between given spectra and molecular structures. The second part uses input output kernel regression, the current cutting-edge method for mapping spectra to molecular vectors obtained by the first part.

（論文審査の結果の要旨）

近年、社会の様々な領域において、計算機による自動化がさらに進んでいる。例えば、将棋や碁等の伝統的ボードゲームでは人工知能によるソフトウェアがトップレヴェルのプロ棋士でさえ圧倒的に凌駕し、自動車等の運転でも人工知能による自動操縦技術の実現が間近と言われている。このような最近の人工知能技術の革新を先導しているのが「機械学習」と呼ばれる技術である。機械学習は、所与のデータから内在する仮説・規則・パタンを抽出し、それらを使って将来を予測する技術である。特に、昨今のビッグデータ時代を迎え、データの豊富さが計算機の高速化と相まって、様々な分野に適用されている。本研究は、質量分析における「代謝物同定（Metabolite Identification)」に着目し、従来技術より「高精度な」、あるいは「高速な」、また、より「解釈可能な」同定を行うために、相応しい機械学習技術の構築を行う試みである。

代謝物同定は、化合物の混合に対して質量分析を行ったスペクトラムを入力し、内在する化合物を出力する問題であり、本研究では一貫してこの問題を扱っている。この問題に対して、構築した手法は目的が異なる以下の２つである。

１，
より「解釈可能な」モデルの構築を目的とし、スペクトラム内のピークの相互作用項を加えた線形モデルを構築し、学習・予測を行った。相互作用項は、化合物構造式における部分構造の含有関係を考慮している。その結果、構築モデルは、カーネル学習に基づく既存手法に較べ、予測精度は、やや劣るものの、はるかに高速で、かつ、ピークの解釈が可能な手法であることを大規模実験により示した。

２，
より「高精度かつ高速な」モデルの構築を目的とし、既存手法の代謝物同定では、スペクトラムに関わらず内在化合物のすべてを常に考慮していることに着目した。すなわち、スペクトラムに含まれている化合物数はそれほど多くない（例えば5から10）にも関わらず、代謝物同定ではすべての化合物（例えば1,000種類）を考慮するが、これは非常に無駄となる。従って、一つ一つの化合物がスペクトラムに含まれているかを考察すると同時に、それらを考慮対象とするかどうかを考察することが可能なモデルを構築した。このモデルには、最新の機械学習技術、例えば、情報量基準であるHSIC (Hilbert-Schmidt Independence Criterion)やグラフニューラルネットワークを利用している。大規模実験の結果から、構築モデルは既存手法を上回る精度とまたはるかに短時間での予測を可能とすることが示された。

前述のように、この２つのモデルは目的が違うため、実際の代謝物同定応用では、上記２つのモデルを目的に応じて使い分けることが可能である。

上記、２つのモデルは各々論文にまとめられ、それぞれ2018及び2019年に、バイオインフォマティクス（生命情報科学）のトップ国際会議であるISMB (Intelligent Systems for Molecular Biology) に採択され、同国際会議の予稿集である、Bioinformatics誌の特別号に掲載された。実際、上記２つのモデルともに、Computational metabolomicsにおいて、機械学習研究に従事している研究者では、現在よく熟知されており、モデルの完成度と有効性が高く評価されている。さらに、この２つのモデルのみならず、既存の機械学習による代謝物同定手法を調査し、それらの性能比較等をまとめた論文がBriefings in Bioinformatics誌に掲載された。

これら計3報の内容は、生命情報科学及び機械学習の研究成果として、一定の基準を十分に満たしているとみなすことができる。

よって、本論文は博士（薬科学）の学位論文として価値あるものと認める。また、令和２年８月２０日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。

要旨公表可能日：　　　　　年　　　月　　　日以降