Knockoff-based target-decoy approach to compute similarities of real metabolomic spectra and decoys using Spec2Vec embeddings

Malgorzata Kurkiewicz (2145411k)

December 16, 2020

1 Status report

1.1 Proposal

1.1.1 Motivation

The motivation behind this project is to improve metabolite annotation. The techniques used in this work involve a knockoff-based target-decoy approach that uses spectral embeddings.

Background

Metabolomics is the study of metabolites – metabolites are small molecular products of cellular regulatory pathways. They are much smaller than proteins and lipids, however the analytical techniques used in proteomics and lipidomics still overlap in metabolomics. In the last few years, there have been a few advances in accurate metabolite identification from untargeted liquid chromatography tandem mass spectrometry (LC–MS/MS). The problem of identification for metabolites has been addressed by a few studies, all of which adapt their approaches from other fields such as proteomics and statistics [1]. Currently, the most common method of annotation for small molecules like metabolites is matching fragment spectra to already existing reference library spectra. Contemporary methods are heuristic-based and compatible with the commonly used similarity scores.

Recently, there has been an advancement in definition of spectral similarity [2] which introduced an unsupervised machine learning technique that has proven more successful in identification of metabolites. Cosine-based scores are the most popular measures of spectral similarity. They assume that molecules with high structural similarity result in spectra with a high spectral cosine similarity score. These scores are not ideal since there is a substantial overlap between the scores for correct and incorrect metabolite identifications (true and false positive matches). They function well with spectra that do not differ much but they do not handle molecules with multiple local chemical modifications. This is because these scores rely on a computationally intensive procedure focusing on aligning the fragment peaks from the two spectra. Noise filtering the data significantly improves the matching process. Spec2Vec is based on a fundamentally different concept since it finds relationships between similar spectra by creating embedded vectors that point in a similar direction. This allows for a better representation of a relationship between spectral and structural similarity.

Currently, there is no target-decoy method for metabolomic spectra compatible with Spec2Vec embeddings. The continuous vectors obtained from Spec2Vec could potentially be modelled as a Gaussian mixture which would provide a basis for a knockoff approach. This could help control False Discovery Rate (FDR) – the percentage of accepted spectral matches that are incorrect – which in turn could improve metabolite identification in the future metabolomic studies.

1.1.2 Aims

The aim of this work is to implement and evaluate a new approach for creating decoy databases which will allow for controlling FDR given metabolomics data. The new technique will use a knockoff target-decoy approach [3] and Spec2Vec embeddings for LC/MS and possibly gas chromatography mass spectrometry (GC/MS) metabolomic data to compute the similarities between real and decoy spectra. We will examine whether the statistical properties kept by creating knockoffs adhere to the methods already proposed in metabolomics. The deliverables of this project would include a comparison of the contemporary target-decoy cosine similarity-based techniques with Spec2Vec and knockoff-based approach for decoy creation.

The success of the project will be measured with popular statistical techniques that assess the quality of FDR estimations. An accurate evaluation can only be carried out when the true identity of all query compounds is known. First, we will calculate the estimated and true q-values (q-values are described as the adjusted p-values found using an optimised FDR approach) for the target-decoy approach for noise filtered and unfiltered data (LC/MS, GC/MS). We will then plot the estimated q-values (y-axis) against the true q-values (x-axis). We will be looking for a close resemblance or at least a significant improvement (considering the same dataset) from the study done by Scheubert et al. where the q-values were visibly underestimated. We would construct a q-value calculation for both filtered and unfiltered data since the current results do not differ significantly when using Spec2Vec similarity measure on the data provided by the Boecker laboratory [4]. If, upon further experimentation, it is proven that we do not need the noise filtering step, we would have a significant achievement to share in the field of spectral matching in metabolomics as the noise filtering of data is one of the main bottlenecks in identifying spectral similarities.

An additional verification measure of the final result is the distribution of p-values of false hits which, by definition, should be uniform.

To summarize, by the end of this project we should be able to understand if there is a more efficient and better way of creating decoy databases in different spaces that do not need to have a limited small size as in the other metabolomic methods. We should also be able to obtain further information regarding filtering of data and whether it is necessary when using Spec2Vec embeddings.

1.2 Progress

This progress bullet list is a condensed version of the timelog and plan available under:
`https://github.com/nitrozyna/FDR-Metabolomics/blob/main/timelog.md`
`https://github.com/nitrozyna/FDR-Metabolomics/blob/main/plan.md`

So far, I have:

- Created a literature review as a result of researching the topic.
- Created Level 2 and Level 3 summaries for seven most important papers to increase understanding on the topic.
- Familiarised myself with tools involved with the project (Spec2Vec, MatchMS).
- Created a general pipeline for data processing, cosine calculation, q-value calculation and FDR estimation.
- Recreated the technique provided by Scheubert et al.[4] for FDR estimation in Metabolomics.
- Changed the cosine similarity calculation for Spec2Vec similarity scoring in the current pipeline by training a Spec2Vec model and embedding the real and decoy spectra accordingly.

- Critically analysed the differences between FDR estimation results obtained using Spec2Vec and results discussed in Scheubert et al. [4]
- Started working on a knockoff technique for generating decoys that is compatible with Spec2Vec embeddings.

1.3 Problems and risks

1.3.1 Problems

At the beginning of the project, I needed to understand concepts from other fields of science. As the project is heavily based on biology and chemistry knowledge, I had to spend a significant amount of time educating myself to be able to understand the meaning behind this work. The process is ongoing and has taken more time than I initially expected.

The recreation of methods from Scheubert et al. [4] was particularly challenging. In the process, I understood there are many factors that I have not considered before. Since the techniques described in that work are not very clear, I needed to refer to their codebase for answers. Even then, I could not obtain similar results. Although I was not able to recreate the results, we were able to identify the main differences and critique both approaches to make sure the basis for the future experiments is understandable and easy to recreate.

When working with metabolomic datasets, I realised that real data coming from a wet lab is very noisy. Working with such data needed some adjustments in the pipeline that I did not plan for. I have realised that results obtained from the contemporary spectral similarity methods heavily depend on the purity of the dataset which needs to be considered carefully.

1.3.2 Risks

Since I am at the stage of the project where I am introducing a novel technique, I expect to encounter a lot more obstacles. The pace of the progress in the project will be slower and require deep understanding of underlying concepts behind knockoff methods for decoy creation. To increase my understanding, I will need to construct experiments that work with small datasets and fewer dimensions. This will help me in creating a more abstract model with a higher complexity.

I will also need to make sure to allocate enough time for experiments related to the best choice of dimensions in the embedded space given spectra represented as vectors created by Spec2Vec. I will need to define appropriate measures of success (such as ratio of true positives and false positives) and present my reasoning in writing.

As this project will be further expanded to tackle GC/MS data input, I will need to make sure to understand how this input differs from the LC/MS datasets I have seen so far. I will choose appropriate design methods for easy switching between different sets of input in the FDR estimation pipeline.

1.4 Plan

The following plan includes a potential extension to the project which is the adjustment of the current pipeline to work with GC/MS datasets. As we cannot foresee all the problems that may appear in this project, it is possible several adjustments will have to be made to deliver a fully functioning product.

Week 13

- Experiment with Spec2Vec dimensionality to find the most optimal parameters.

Week 14

- Continue experimenting with the knockoff technique, progress on the abstract of the thesis.

Week 15

- Continue experimenting with the knockoff technique and evaluate the current results, progress on the background for the thesis.

By week 15 we should have substantial results for LC/MS datasets. If this is the case, we can move on to the optional extension of considering GC/MS datasets.

Week 16

- Adjust the pipeline to account for GC/MS data, progress on the introduction for the thesis.

Week 17

- Progress on including GC/MS data, progress on the experimental part of the thesis.

Week 18

- Progress on including GC/MS data, progress on the experimental part of the thesis.

Week 19

- Test the FDR estimation pipeline with different set of data inputs (LC/MS, GC/MS, libraries, denoised data sets etc.), progress on the experimental part of the thesis.

Week 20

- Continue testing the pipeline, refactor code and write documentation.

Week 21

- Evaluate all results and progress in writing evaluation for the thesis.

Week 22

- Write the conclusion for the thesis, discuss handover of the project and talk to interested parties that could benefit from knowing about these results.

Week 23

- Create a presentation on the project.

Week 24

- Dissertation submission deadline and presentations.

References

[1] Lukas Käll et al. "Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases". In: *Journal of Proteome Research* 7.2 (2008), pp. 29–34. DOI: 10.1021/pr700600n.

[2] Florian Huber et al. "Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships". In: *bioRxiv* (2020). DOI: 10.1101/2020.08.11.245928.

[3] Jaime Roquero Gimenez, Jamirata Ghorbani, and James Zou. "Knockoffs for the mass: new feature importance statistics with false discovery guarantees". In: *arXiv.org* 1 (2019). DOI: arXiv:1807.06214.

[4] Kirsten Scheubert et al. "Significance estimation for large scale metabolomics annotations by spectral matching". In: *Nature Communications* 8.1494 (2017). DOI: 10.1038/s41467-017-01318-5.