Towards off-the-grid demultiplexing of LC-MS/MS data

Gowtham Sankarananth Seenivasaharagavan*[1,2], Thomas Burger[3], and Antoine Chatalic[1]

[1]GIPSA Pôle Géométrie, Apprentissage, Information et Algorithmes – Grenoble Images Parole Signal Automatique – France

[2]CEA Grenoble – INSERM U1038, Université Grenoble Alpes – France

[3]CEA Grenoble (BIG, Biologie à grande Echelle, EDyP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France

Abstract

The proteome, which is the set of all proteins expressed by a tissue, is a wellspring of biochemical information. This repository is predominantly tapped by the biotechnology termed LC-MS/MS (liquid-chromatography coupled to tandem mass-spectrometry). Here, the proteins are digested into simpler biomolecules called peptides which are then sorted by liquid chromatography (LC) and, finally, encoded as mass spectrograms. However, interpretable yet accurate inference of proteomes from LC-MS/MS data remains an open challenge as each mass-spectrogram does not encode a single peptide, but rather a mixture of several peptides. Consequently, every spectrum must be computationally unmixed (aka demultiplexed) to recover the identity of the constituent peptides. In practice, to leverage the continuity of spectrograms along the sorting by LC, demultiplexing parses the entire collection of spectra produced along the LC's elution time (which can be visualized as an image if they are stacked together).

Existing approaches to infer the proteome are either accurate or interpretable – not both. An example of the former is DIA-NN, where a deep neural network bestows accuracy at the cost of interpretability. In contrast, algorithms like non-negative matrix factorization (NMF) are interpretable as they employ linear unmixing models, but, by the same token, are less accurate. While the need for accuracy is self-evident, interpretability is equally crucial as it determines the utility of the demultiplexed output in downstream biology related tasks (e.g., protein roll-up, differential analysis, etc.).

To this end, we seek to improve the accuracy of NMF by encoding the relevant physics of LC-MS/MS. NMF models the LC-MS/MS data matrix as a sum of non-negative rank-one factors corresponding to individual peptides. Concretely, each factor is the outer-product of the peptides' mass spectrum and elution profile. Since these factors are solutions to block-convex optimization problems, they can be efficiently recovered by block coordinate descent (aka alternating minimization). Even so, NMF's rank-one assumption is at odds with the MS precision being finite (MS signal can be pictured as a collection of Dirac that jitters along time). Consequently, NMFs' estimates of peptide mass spectra can be forced to a poorer resolution than the measurements themselves.

*Speaker

Specifically, we account for the unmodelled effect of mass precision by augmenting the variational formulation of NMF with additional decision variables that model the support of each peptides' mass spectrogram as a finite collection of points. By ensuring the number of support points are below the maximum predicted by theory, we hope to improve the resolution of the estimated spectra.

However, the resulting optimization problem is no longer block-convex.

So, we leverage the block successive upper bound minimization framework to develop an approximate alternating minimization algorithm that is still guaranteed to converge to stationary points.

Future work aims at improving the physical relevance of these stationary points with appropriate regularizers that promote sparsity of the peptide spectra weights (to account for varying peptide lengths) and smoothness of elution profiles. Subsequently, we will validate the resulting unmixing algorithm on synthetic and actual LC-MS/MS data.

Keywords: Proteomics, liquid chromatography mass spectroscopy, non negative matrix factorization, alternating minimization, block convexity