A family of open-source tools for deep data mining in large-scale mass spectrometry-based proteome analyses

Lev I. Levitsky, Mark V. Ivanov, Irina A. Tarasova, Anna A. Lobas, Marina L. Pridatchenko, Julia A. Bubis, Elizaveta M. Solovyeva, Mikhail V. Gorshkov

Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

Talrose Institute for Energy Problems of Chemical Physics, RAS, Moscow, Russia

Proteome characterization reveals important biological information which is not provided by DNA or RNA sequencing. It provides crucial insights into protein expression in cells, posttranslational modifications, as well as protein-protein interactions. This characterization is performed by mass spectrometry (MS) which provides sequence specific information about proteins and their proteolytic peptides and thus allows unambiguous identification and quantitation of proteins in a sample. Recent advances in MS instrumentation have multiplied the amounts of data that need to be processed in a single proteomics experiment, forcing the data processing algorithms to adapt in terms of sensitivity/specificity and computational performance. The typical data processing workflow for protein identification and/or quantification in MS-based shotgun proteomics consists of a number of non-trivial procedures, including denoising and deisotoping of MS data, peptide identification using a protein database search, post-search processing and validation, protein inference and quantification. Each of these stages may have a critical impact on the overall performance of the data processing and the validity of biological discoveries made. Although the proteomics field has benefited greatly from the emergence of new highly performant mass analyzers, such as those based on the Orbitrap™ technology, these changes may have, to some extent, aggravated the key problems of proteomics data processing software development as an ecosystem: excessive diversity of data representation formats, lack of flexibility and universal applicability of data processing algorithms, etc. Some of the commonly used tools also have strict licensing limitations, rendering the employed algorithms opaque and the results hard to validate.

In this report, we present a family of data processing tools developed in Python programming language and aiming to address all above-mentioned challenges of deep proteome

analysis. Being an open-source high-level dynamically typed interpreted language, Python perfectly fits the current trend toward flexibility, simplicity, openness, and cross-platform compatibility of proteomic data processing software.

Almost any piece of proteomics software relies on a certain set of primitives representing the key objects (peptides, proteins, mass spectra) and their basic properties, as well as common operations on those. We have put together a collection of such primitives in our Pyteomics library. We find that the library greatly facilitates software development for MS-based proteomics, regardless of its specific application. Pyteomics is the core of every software application we develop, and can be recommended in everyday use for rapid script-based data analysis by proteomic laboratory bioinformaticians.

Pyteomics includes parsers for common data formats used in MS-based proteomics, which can help bridging a gap between different formats. This use of Pyteomics will be exemplified in this report by the pepxmltk package, which can be used for conversion of X!Tandem search engine output to the more common pepXML format, enriching it with quantitative information.

Since Pyteomics allows calculation of peptide properties, it is tempting to utilize them further for post-search validation. In this report we demonstrate this idea in our MPscore software, which validates MS/MS-based peptide identifications by considering complementary experimental data, such as retention times, cleavage specificity, precursor ion mass error, etc. This allows recovering a significant number of peptide identifications of modest quality, increasing the overall sensitivity of identification. MPscore also implements a number of label-free quantitation methods for accurate quantitative analysis.

Finally, we make full use of Pyteomics MS data parsers, format conversion tools, and the post-search validation software MPscore in IdentiPy, an experimental proteomic data search engine we are building. We are making it as flexible and universal as possible by allowing to tweak the most defining part of any search engine: the scoring function. Since the scoring algorithm is essentially what makes search engines different, and applicable to a limited range of mass analysers, we can unify workflows for different data sources, leaving all differences to the corresponding configuration files. IndetiPy is also used as a research platform, allowing us to gain insights into the mechanisms in which different search parameters affect the apparent results

- something too troublesome for a “black box” tool that most search engines effectively are. Complementary peptide properties and statistical analysis of their distributions also come extremely useful in a search engine, since it is only there that one can evaluate the whole search space and affect the result of peptide matching, rather than deal with its outcome afterwards. The same statistical techniques can be also applied for self-optimization of search parameters, improving search specificity and making the software much friendlier to non-expert users.

Protein identification and quantitation is a crucial step in proteomics-based systems biology. The above-mentioned data analysis procedures affect the conclusions that can be drawn from biological studies. The importance of appropriate MS data analysis is however often underestimated by biologists. In this report we will emphasize the need to handle the proteomics data carefully and consciously in order to get the best of it, and present a set of flexible tools based on the open-source framework.

1. A.A.Goloborodko, L.I.Levitsky, M.V.Ivanov, M.V.Gorshkov (2013) Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics, *Journal of The American Society for Mass Spectrometry*, **24(2)**:301-304.

2. M.V.Ivanov et al. (2014) Empirical Multidimensional Space for Scoring Peptide Spectrum Matches in Shotgun Proteomics. *Journal of Proteome Research*, **13**:1911–1920.