Design and implementation of analysis pipeline for single cell type proteomics data

By

Lukas Gamp

in partial fulfilment of the requirement
for the degree of MSc
in Bioinformatics

mm yy

Abstract

(the spacing is set to 1.5)

no more than 250 words for the abstract

- a description of the research question/knowledge gap – what we know and what we don't know
- how your research has attempted to fill this gap
- a brief description of the methods
- brief results
- key conclusions that put the research into a larger context

Contents

Acknowledgements

Thank you for following this tutorial!

I hope you'll find it useful to write a very professional dissertation.

1   Introduction

1.1   Proteomics

The proteome is referred to the sum of all proteins of a given sample at a given time. In the past several quantitative and qualitative assays were used to enlighten the protein composition of a sample.

An early approach of qualitative analysis of the cellular proteome involved labeling with fluorescent antibodies and imaging. The major disadvantage of this technique was the limitation to only stain a few proteins per cell. For quantification procedures such as single-cell Western blots, immunoassays or CyTOF have been used. Other disadvantages are the ability to permeate cells, accessibility and binding of the epitope and the creation of specific antibodies for a given protein (Budnik et al. 2018).

One of those techniques involved RNA-sequencing. Since RNA involves also non-coding RNA, the amount of RNA is often not proportional to the content of proteins in a cell. So the proteinaceous content of a cell was only predicted and quantitative analysis was not possible.

1.2   Mass Spectrometry

Mass spectrometry enables qualitative and quantitative analysis of the entire repertoire of a biological sample. The availability of gene sequences in databases and the ability to match proteins against those sequences with computational methods makes it possible to identify alterations of a sample on a protein level. These alterations can rely on the sequence level or could be to post-translational modifications (PTMs) such as phosphorylation, methylation or else (Aebersold & Mann 2003).

Mass to charge ratio (m/z)

1.2.1   Instrumentation

1.2.1.1   Cell Isolation

1.2.1.2   Lysis

1.2.1.3   Digestion

1.2.1.4   Liquid chromatography   In order to separate proteins according to their chemical properties, size or species a liquid chromatography (LC) is recommended before ionization.

1.2.1.5   Ionization   In order to analyze a biological sample consisting of proteins in solution the liquid needs to be vaporized into gas phase. Two techniques are capable of this procedure. Electrospray ionization (ESI) pushes the analyte through a capillary and applies an electric current to the liquid, vaporizing the sample to a charged aerosol. Biomolecules are fragmented according to their chemical properties and can be further handled in the mass spectrometer. The fragmented biomolecules are now in charged droplets separated by their charge on the surface, splitting further into smaller droplets until they become a gas phase ion. Two physical models describe the process from gas phase to ion called "The ion evaporation model" (IEM) and "The charge residue model" (CRM). In the ion evaporation model (by Iribarne and Thomoson) the droplets shrink by evaporation until ions are expelled. The model had its limitation by explaining same evaporation rate constant among ions with different chemical properties. In the charge residue model the assumption of one molecule per droplet leads to an ionization rate constant, which is independent of the ion itself and relies solely on the generation of the droplet and the efficiency of the solvent (Wilm 2011).

Matrix-assisted laser desorption/ionization (MALDI)

1.2.1.6   Coupled mass-spectrometry (MS/MS)   In order to enhance sequence identification, two MS devices are built in series. In the first run (MS1) the m/z is determined and the molecules are passed to the next device. Upon passing the molecules are fragmented into smaller ions and analyzed by the second MS. The fragmentation highly depends on the chemical bonds found in the molecule. The majority of these breaks occur on the peptide bond of the protein, although this is not guaranteed for all bonds and so it can happen that certain peptide ions have a low abundance (Budnik et al. 2018). These low abundant peptides will not be detected, hence the problem needs to be faced with another approach . A solution for this problem is molecular barcoding with labeling mentioned in the chapter labeling.

1.2.1.7   Labeling techniques   For differential analysis proteins need to be labeled to compare mass to charge intensities. Since mass spectrometry is not a quantitative technique by itself, the peak height or area does not reflect the abundance of a peptide. Physicochemical properties of the proteins can change the ionization efficiency and detectability of the target. However, when comparing the same analyte between multiple runs of labeled proteins, differences in the mass spectrum reflect the

abundance of those. Labels should be chosen to change solely the mass of the sample and to not affect folding or other inherent properties of the protein.

1.2.1.8   Metabolic labeling   Feeding cells with aminoacids containing heavy isotopes, is the method of choice in order to label peptides at the earliest possible level. This atoms can be heavy nitrogen in aminoacids or salts in fertilizer for plants. Mass shifts are proportional to the isotopes incorporated during biomass production and are visible after proteolytic cleavage. Stable isotope labeling in cell culture (SILAC) was presented in the early 2000s. This method used heavy aminoacid enriched media to feed cells, in order to quantitatively analyze expression profiles.

1.2.1.8.1   Isobaric labeling   Tandem mass tag (TMT)

Tandem mass tag (TMT) reagents enable to differentiate multiple samples analyzing in one MS run. The samples are labeled individually and pooled afterwards, this procedure is called multiplexing. All TMTs have the same mass and can be found at the same peak in all samples, these peaks are called reporter ions (RI). The identification of these RI leads to an enrichment and identification of low abundance peptide ions which is common especially in single-cell techniques. With this technique it is possible to quantify proteins and differ low abundant proteins from background noise. The disadvantage of isobaric labeling is, that the co-fragmentation signals can be observed in the spectrogram and the data needs to be normalized in order to remove unwanted contribution (Marx 2019, Budnik et al. 2018).

1.2.2   MS-Spectrum

Each peptide is reflected by its' indivual fingerprint in the ms-spectrum. The fingerprint is based on the chemical properties and modifications. In order to identify proteins, fingerprints are matched against a sequence database (Cox & Mann 2008).

1.2.3   Interpretation of the data

Since ms data has a high resoultuion, algorithms are used to convert peaks to an interpretable form. These algorithms find local minima of the function to seperate peaks from each other. Peaks include all isotopes of the proteins `containing atoms. MaxQuant \citep{Cox2008} is one of the sofware packages to process the data and provides it for further analysis and statistical testing. To find the istopic distribution of a biomolecule, MaxQuant`

creates a vertex of every single peak and connects them with their possible isotopic counterparts by finding the proportion of mass of an average aminoacid to its respective isotope (averagine (Senko et al. 1995)). After this procedure a large amount of noise is reduced and a single peak reflects a small biomolecule. These biomolecules can now be searched in a database in forward and reverse direction. The peptide identification (P-) score indicates the fit of the data to the found sequence in the database according to the length of the peptide and is used to calculate the posterior error probability (FDR). The calculation of the false discovery rate is then calculated by taking FDR into contrast. After these calculation the peptide peaks are joined to a protein and can be quantified.

1.2.3.1 Principle component analysis (PCA)   Protein levels can be projected to their principle components (PC) and clustered to their specific cell type. So cell types are distinguished by their proteinaceous composition.

1.2.3.2 Testing for differential expression   According to the experiment and sample groups, individual statistical methods should be used.

follow sctranscr. analysis pipe

signatures of cell clustures

monocyte data acvailiabe

tmt label $->$ noise reduction

tools for clustering

plst, knm,

2   Methods

3   Materials and Methods

3.1   Materials

For analysis two types of cells were used. One type is the Jurkat-based cell line (J-lat) with integrated HIV.

The other type of cells are macrophages with a sample size of 72 cells. The analysis is done with two groups. A HIV negative (HIV-) control group and a HIV positive (HIV+) group.

3.2   Data

3.2.1   Acquisition

Analysis of the data was done with MaxQuant (Cox & Mann 2008)

4 Results

Some more guidlines from the School of Geosciences.

This section should summarise the findings of the research referring to all figures, tables and statistical results (some of which may be placed in appendices). - include the primary results, ordered logically - it is often useful to follow the same order as presented in the methods. - alternatively, you may find that ordering the results from the most important to the least important works better for your project. - data should only be presented in the main text once, either in tables or figures; if presented in figures, data can be tabulated in appendices and referred to at the appropriate point in the main text.

Often, it is recommended that you write the results section first, so that you can write the methods that are appropriate to describe the results presented. Then you can write the discussion next, then the introduction which includes the relevant literature for the scientific story that you are telling and finally the conclusions and abstract – this approach is called writing backwards.

5 Discussion

the purpose of the discussion is to summarise your major findings and place them in the context of the current state of knowledge in the literature. When you discuss your own work and that of others, back up your statements with evidence and citations. - The first part of the discussion should contain a summary of your major findings (usually $2-4$ points) and a brief summary of the implications of your findings. Ideally, it should make reference to whether you found support for your hypotheses or answered your questions that were placed at the end of the introduction. - The following paragraphs will then usually describe each of these findings in greater detail, making reference to previous studies. - Often the discussion will include one or a few paragraphs describing the limitations of your study and the potential for future research. - Subheadings within the discussion can be useful for orienting the reader to the major themes that are addressed.

6 Conclusion

The conclusion section should specify the key findings of your study, explain their wider significance in the context of the research field and explain how you have filled the knowledge gap that you have identified in the introduction. This is your chance to present to your reader the major take-home messages of your dissertation research. It should be similar in content to the last sentence of your summary abstract. It should not be a repetition of the first paragraph of the discussion. They can be distinguished in their connection to broader issues. The first paragraph of the discussion will tend to focus on the direct scientific implications of your work (i.e. basic science, fundamental knowledge) while the conclusion will tend to focus more on the implications of the results for society, conservation, etc.

7 Bibliography

Aebersold, R. & Mann, M. (2003), 'Mass spectrometry-based proteomics', *Nature* **422**, 198–207.

Budnik, B., Levy, E., Harmange, G. & Slavov, N. (2018), 'Scope-ms: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation', *Genome Biology* **19**, 161.

Cox, J. & Mann, M. (2008), 'Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nature Biotechnology* **26**, 1367–1372.

Marx, V. (2019), 'A dream of single-cell proteomics', *Nature Methods* **16**, 809–812.

Senko, M. W., Beu, S. C. & McLaffertycor, F. W. (1995), 'Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions', *Journal of the American Society for Mass Spectrometry* **6**, 229–233.

Wilm, M. (2011), 'Principles of electrospray ionization', *Molecular & Cellular Proteomics* **10**, M111.009407.

8 Appendix(ces)

8.1 Appendix A: additional tables

Insert content for additional tables here.

8.2   Appendix B: additional figures

Insert content for additional figures here.

8.3   Appendix C: code

Insert code (if any) used during your dissertation work here.