Shared Peptides For Accurate Protein Quantification and Predictive Biomarker Selection

Mateusz Staniak

Proteomics is a field of research concerned with a *large-scale characterization of the entire protein complement of a cell line, tissue, or organism.* The study of proteins provides knowledge that cannot be deduced from the study of genes, or transcripts, as it is the proteins, and not the genes that determine the phenotypes of cells due to processes such as post-translational modifications (PTMs). The core technology of proteomics capable of both identification and quantification of proteins in biological samples is mass spectrometry (MS). Quantification allows a differential analysis of protein expression, which is necessary for biomarkers discovery - finding proteins whose expressions are associated with a disease.

In the bottom-up approach to MS, peptides - smaller segments of proteins - enter the mass spectrometer and thus measurements are made on a peptide level. Some peptides may be assigned to several proteins. This problem cannot be solved by technological advancement, as it is caused by sequence homology, often related to protein families (groups of evolutionarily-related proteins) or protein variants (similar proteins originating from one gene or gene family). Such peptides are referred to as *shared* or *degenerate* peptides. Presence of these peptides, along with proteins identified only by a single peptides, complicates both finding a reliable list of proteins present in the sample (*protein inference*), and estimation of their abundance (*protein quantification*).

Currently, the most prevalent approach to handle shared peptides is to ignore them. This leads to several problems in protein quantification. Firstly, fewer proteins are identified, as proteins identified only by shared peptides are discarded from the analysis when shared peptides are removed. Secondly, protein abundance estimates are altered by shared peptides removal (or grouping of proteins whose abundance is estimated jointly). Lastly, the precision of protein abundance estimates is decreased due to the lower sample size.

The relationship between identified peptides and potential proteins can be represented with a bipartite graph. Thus, we will propose a novel statistical learning approach to the analysis of bipartite graphs based on regularization. Using quantitative profiles of peptides as input information in peptide nodes, we will jointly estimate protein abundance and select proteins that are actually present in the sample by removing edges between proteins and peptides to control the false discovery rate of protein identifications without sacrificing statistical power.