Karolinska Institutet

Department of Oncology-Pathology

Multivariate analysis of cancer proteomics data – towards a biological systems view and understanding

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Föreläsningssal Samuelsson, Tomtebodavägen 6

Torsdagen den 19:e september 2013, kl. 9.30

av

Lina Hultin Rosenberg

Huvudhandledare:
Docent Janne Lehtiö
Karolinska Institutet
Institutionen för Onkologi-Patologi
Cancer Proteomics Mass Spectrometry
SciLifeLab Stockholm

Bihandledare:
Dr. Jenny Forshed
Karolinska Institutet
Institutionen för Onkologi-Patologi
Cancer Proteomics Mass Spectrometry
SciLifeLab Stockholm

Fakultetsopponent:
Professor Lennart Martens
Ghent University
Faculty of Medicine and Health Sciences
Department of Biochemistry
Computational Omics and Systems Biology

Betygsnämnd:
Professor Johan Gottfries
Göteborgs Universitet
Institutionen för Kemi och Molekylärbiologi

Docent Per Andrén
Uppsala Universitet
Institutionen för Farmaceutisk Biovetenskap

Professor Erik Sonnhammer
Stockholms Universitet
Stockholm Bioinformatics Centre
SciLifeLab Stockholm

Stockholm 2013

ABSTRACT

Important aims of cancer proteomics include gaining better understanding of cancer biology and identifying cancer biomarkers. Mass spectrometry (MS) based shotgun proteomics allow for identification and quantification of thousands of proteins in complex human samples. However, proteomics discovery research in clinical material faces many challenges. The biological differences between groups are often expected to be rather small, at the same time the human proteome is highly complex and there is large biological variation between clinical samples. To be able to extract meaningful results from proteomics data derived from biological and clinical material, care has to be taken to all the critical steps in the data analysis workflow. First of all we need to have robust methods to extract good quality data. A proper statistical analysis is then of outmost importance, taking into account risks of over-fitting and false positives. In addition, we also need system based approaches to relate the data to clinical and biological questions.

The main goal of this thesis was to generate robust methods for selection of key proteins, networks and pathways relevant for answering biological and clinical questions. The work includes development and evaluation of workflows for quantitative analysis of proteomics data.

In paper I, a multivariate meta-analysis workflow was developed to link existing proteomics data from human colon and prostate tumours. The aim was to identify proteins distinguishing between normal and tumour samples independent of tissue origin, as well as to find unique markers. The bioinformatics workflow for meta-analysis developed in this study enabled the finding of a common protein profile for the two malign tumour types, which was not possible when analysing the data sets separately. The purpose of paper II was to generate a basis for the decision of what protein quantities are reliable and find a way for accurate and precise protein quantification. We developed a methodology for improved protein quantification in shotgun proteomics and introduced a way to assess quantification for proteins with few peptides. The experimental design and developed algorithms decreased the relative protein quantification error in the analysis of complex biological samples. In paper III, we presented SpliceVista, a tool for splice variant identification and visualization based on MS proteomics data. SpliceVista identifies splice variant specific peptides and provides the possibility to perform splice variant specific quantitative analysis. SpliceVista was applied in two experimental datasets to exemplify its capability of detecting differentially expressed splice variants at the protein level. The aim of paper IV was to develop a network based analysis workflow for proteomics data to identify protein subnetworks with different activity between groups of samples. The methodology, which is based on a multivariate model directed by the network, was applied to several of our clinical mass spectrometry datasets. The output from the subnetwork analysis was functional subunits of proteins, rather than a collection of sparse proteins, which were shown to more readily provide a model of the biological mechanisms studied, and thus aid in the biological interpretation.

ISBN 978-91-7549-243-8