

Abstract

Technology has been the driving force behind the design and development of many analytical methods such as NMR, Chromatography and Mass Spectrometry leading to major advancements in proteomics and other life science disciplines. The increasing importance of high-throughput techniques has resulted in the generation of enormous amounts of data that pose new challenges in their analysis and interpretation. Therefore, the development of sophisticated algorithms and computational tools is of utmost importance to effectively manage, analyse and interpret the experimental results.

In this thesis, two database-centered systems MMsINC and MASPECTRAS 2.0 have been implemented: the chemogenomics system MMsINC predicts the pharmacological profile of newly isolated or synthesized compounds, based on the similarity to active compounds in the MMsINC chemical space. The MAss SPECTRometry Analysis System 2.0 (MASPECTRAS 2.0) organises, clusters and visualises mass spectrometry based protein identifications for sophisticated biological interpretations. The features include a command line interface which enables integration into automatic mass spectrometry production pipeline, an annotation fetching mechanism, a versatile user interface, and logical comparison of samples. Additional clustering algorithms enable grouping of proteins based on mass spectrometry evidence grouping and gene ontology.

In conclusion, the developed tools are a significant contribution to the scientific community, which allow researchers to perform analysis on large-scale proteomics experiments and evaluation of the activity of new compounds.

Keywords: proteomics, sample organisation, protein clustering, external databases, gene ontology, mass spectrometry, cheminformatics, chemogenomics, virtual screening, lead discovery