Overview

Purpose: Extraction of biologically meaningful information from proteomics datasets
Method: Development of a novel bioinformatics tool
Results: Implementation of KEGG pathway repository into existing bioinformatics tool

Introduction

The mass spectrometry-based approach to proteomics studies has significantly enhanced our understanding of complex biological processes. However, a number of challenges remain, one of which is how to quickly extract biologically relevant information from large datasets. Despite the wide variety available, most bioinformatics tools have been designed to deal with genomic data and do not take into consideration the specific characteristics associated with proteomics data.

Here we present a new development in the Thermo Scientific ProteinCenter bioinformatics tool: Implementation of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway repository. The KEGG pathway repository has been implemented into the ProteinCenter software specifically designed for the interpretation of proteomics data, supporting direct import into ProteinCenter.

Methods

ProteinCenter™ software is specifically designed for the interpretation of proteomics data, supporting direct import into multiple database search engines and protein repositories. Recently, a pathway analysis tool based on the KEGG pathways repository has been implemented. Proteomics datasets were downloaded from literature sources, in particular, data from Olsen, JV., et al., Cell. 2008. 137(1):635-48.

Data loaded into ProteinCenter software included protein accession codes, peptide sequences and quantitative information. Protein lists were clustered according to an indistinguishable proteins group algorithm. IP and Swiss-Prot databases were used as reference databases. Statistical analysis was performed based on Benjamin-Hochberg correction of p-values to determine which proteins are over-represented.

Visualization of Genes Involved in the ErbB Pathway

Figure 4 shows a heat map of genes in the ErbB pathway colored according to the median of their respective phosphopeptide ratios. The pathway maps show 0, 5, 10, 20 minutes after EGF treatment of HeLa cells.

Conclusion

Modern proteomics studies generate an enormous amount of peptide and protein identifications. One of the remaining challenges in proteomics data analysis, especially when trying to interpret the characteristics of protein datasets based on the involvement of identified proteins in different pathways.

Here we presented a software tool that enables characterization of protein datasets based on the involvement of identified proteins in different pathways. Using the study of global site-specific phosphorylation dynamics in EGFR signaling network (Olsen et al., Cell 2008) as an example, we concluded that the dominant over-represented pathway was the ErbB signaling pathway (20% of pathway coverage).

As anticipated, after 10 min. of stimulation, top over-represented pathways included mTOR and MAPK. This is in agreement with the current model of EGFR signal transduction.

References