**@title yeastBetaSheetCalculator**

**#' @description**

#' NetSurfP-1.0 is a prediction tool for secondary structures using neural network.

#' NetSurfP-2.0 is an extension of NetSurfP-1.0 which utilized deep neural network to predict

#' secondary structures with the accuracy of 85%. In addition to accuracy, this tool presents

#' reduced computational time compared to other methods.

#' NetSurfP-2.0 is designed to be user friendly and efficient in calculation time of large number of

#' sequences. In addition to that the output of the calculation is available in many formats that would

#' make further data analysis even easier. This tool is available as a web-sever

#' ([http://www.cbs.dtu.dk/services/NetSurfP-2.0/)](http://www.cbs.dtu.dk/services/NetSurfP-2.0/%20)) which can accept up to 4000 sequences

#' at a time. NetSurfP-2.0 is an amazing tool for the prediction of secondary structures from

#' the primary structure. However, this tool like other prediction tool come with its own drawbacks.

#' In bottom-up proteomics, the proteins are enzymatically digested to peptides; sometimes

#' this process produces hundreds of thousands of peptides. First, NetSurfP-2.0 is not designed to

#' accept as many peptides at once, therefore the process of uploading the sequences and

#' waiting for the calculations to be complete is extremely time consuming. Second, even if all

#' sequences uploaded successfully and the results are back, it would be almost impossible to

#' combine the results that have been produced for each individual peptide

#' (hundreds of thousands of spread sheets) to get a coherent picture of the secondary

#' structure of the proteins.

#' In this document, an extension for NetSurfP-2.0 is presented which is specifically designed to

#' analyze the results of bottom-up proteomics that has primarily analyzed with MaxQuant.

#' We call this tool Yeast Proteome Secondary Structure Calculator (YPSSC).

#' YPSSC, on one hand benefits forms the accuracy of NetSurfP-2.0 to calculate secondary

#' structure and on the other hand addresses the issue of analyzing so many peptides with

#' NetSurfP-2.0 by eliminating the need for direct analysis of the peptides from bottom-up

#' proteomics.

#' Instead of direct analysis of peptides by NetSurfP-2.0 which raises the problem of combining

#' the results of peptides to proteins, the whole yeast proteome has been analyzed once by

#' NetSurfP-2.0 and kept as Secondary Structure Database for Yeast Proteome (SSDYP). Then

#' the peptides form the experiment are matched and compared to this database to extract

#' secondary structure of the peptides.

#' The SSDYP contains structural information for all amino acids of whole yeast proteome

#' (Over 3000,000 amino acids) which contains over 6700 proteins.

#' For a hypothetical protein, the SSDYP contains the ID of the protein, amino acids with numbers

#' and structural information for each amino acid.

#' Focusing on the hypothetical protein, in the real sample, there are many peptides identified from

#' the hypothetical protein. YPSSC first finds all the peptides that belongs to the hypothetical protein

#' and arrange them based on the numbers of the amino acids; then it removes the parts of the protein

#' that have been identified more than once in multiple peptides and collapses the population of

#' identified peptides in the sample into one sequence that represents the coverage of the

#' hypothetical protein. The result would show that which part of the protein is identified, and which

#' part is missing.

#' Then, YPSSC matches the the sequence that identified in the sample with SSDYP to find the

#' structural information about amino acids. In this example we are looking at beta-sheet structures.

#' The ration between number of beta-sheet

#' amino acids in the sample to SSDYP, determines the coverage of beta-sheet in the sample.

**#' @param pathFileInput**

#' MaxQuant is a quantitative proteomics software designed to analyze large mass-spectrometric

#' data. The input of MaxQuant is a raw file (.raw) from high-resolution mass spectrometers.

#' After analysis of the raw file in MaxQuant, the program generates a folder named “combined”.

#' In this folder there is another folder named “txt” which contains many files with text format (.txt).

#' One of the files called “peptides” which is the input of the YPSSC to calculate secondary structures.

#' YPSSC has been designed such a way that can analyzed and extract information regarding

#' the sample regardless of the name that user chosen for the sample.

**#' @return**

#' the output of the program is a csv file (.csv) that contains 5 columns, and the number of rows

#' depends on the number of proteins in the sample. First column contains the ID of the identified

#' beta-sheet proteins in the sample, second column contains the number of identified amino acids

#' from the corresponding protein, third column contains number of identified amino acids with

#' beta-sheet, fourth column contains the number of amino acids that the protein

#' originally has in the SSDYP, and fifth column contains the number of amino acids

#' with beta-sheet that the protein originally has in the SSDYP. These columns should

#' provide all information that the user needs to know about the protein and its beta-sheet

#' information as well as beta-sheet information about the parts of the protein that has been

#' identified in the sample.