IPRStats: visualization and analysis of InterProScan Results

David E. Vincent[1] and Iddo Friedberg*[1,2]
1. Department of Computer Science and Software Engineering
2. Department of Microbiology
Miami University, Oxford OH, USA
*Corresponding author: i.friedberg@muohio.edu

URL: http://github.com/idoerg/IPRStats

Introduction: InterProScan is a popular tool used in the functional analysis of protein sequences; it is a powerful tool for identifying protein families, predicting protein function, as well as other features included in InterPro member databases. While the information generated by InterProScan is extremely useful it can be difficult to manage and interpret because of the vast amount of data it produces when analyzing genomic and metagenomic data.

We present IPRStats, a web server and standalone program that accepts the output of InterProScan and provides chart and tabular summaries of the data. These can be downloaded for further use and data analysis pipelining. The user uploads the InterProScan XML output to the server. This output gets incorporated into a MySQL database which is then queried by a series of scripts producing the output. Additional features planned are the correlation of sequence signatures with the abiotic conditions of the habitats from which they were produced. Graphs are produced using Google graphs.

Conclusions: we provide a useful web tool which will allows users to upload data generated by InterProScan and receive a graphical display summarizing the results. A page of information and charts is generated for each database which was included in the InterProScan search. These pages allow the users to see features such as the most common protein families and sequence signatures found in PFAM, TIGRFAM, PANTHER and common structures reported by Gene3D. Each page also provides links to more information about the data from sites such as GO and EBI.