IPRStats User Documentation

Introduction

IPRStats is a statistical tool that facilitates the analysis of InterProScan results by generating graphs and tables with links to additional information.

InterProScan is a popular tool used for 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 metagenomic data. This is because when analyzing metagenomic data, the analysis is that of a population of genes and their functions, and the results should be displayed as a statistical analysis

We present IPRStats, an application 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.

Installation

In order to make installation easier, we have developed several binary installers for IPRStats. By going to http://github.com/devrkel/IPRStats and clicking on “Downloads”, you can find the most current installers for your system. We currently have Windows, Mac, and Ubuntu installers (tested on Vista, Snow Leopard, and Lucid Lynx respectively).

Windows

Before installing IPRStats, you must install several dependencies. Installation of these dependencies is very straightforward:

- Python 2.6 available at http://www.python.org/download/releases/2.6.5/.
- wxPython 2.8 runtime (unicode) available at http://www.wxpython.org/download.php

Once you have these two dependencies installed, you can download and install IPRStats. Navigate to http://github.com/devrkel/IPRStats and click the “Downloads” link. Choose the IPRStats-0.4.win32 installer and run it. This will install IPRStats and add a shortcut to your Windows start menu.

Mac OS X

The Mac OS X application can be downloaded from http://github.com/devrkel/IPRStats by clicking the “Downloads” link and downloading the Mac application. IPRStats comes packaged in a .dmg disk image. You must open the image and drag the application to your Applications folder (or anywhere on your computer) before double clicking it to open the application.

Ubuntu/Debian

A debian installer is provided at http://github.com/devrkel/IPRStats in the download section. Download the installer and double click it. This will open the installer which will download and install any missing dependencies you may have. The installer will add a menu item under the Science section in Gnome.

Other

It should be possible to run IPRStats on any *nix, Solaris, or BSD system on which you can install its dependencies. These dependencies are:

- Python 2.6
- wxPython 2.8

You must also have an Internet connection if you wish to download charts. After the dependencies are installed, download the source code from http://github.com/devrkel/IPRStats as either a zip or tarfile by clicking “Download Source.” Extract the contents to a location on your computer and run “iprstats.py” inside the “iprstats” folder.

Optional

You can optionally install the following Python modules for additional chart types:

- NumPy and Matplotlib available at http://sourceforge.net/projects/numpy/files/ and http://sourceforge.net/projects/matplotlib/files/matplotlib/matplotlib-1.0/ respectively.

Using MySQL

It is highly recommended that you use MySQL for large XML files. By default IPRStats uses SQLite as the database to store and query data extracted from a given InterProScan XML file. IPRStats gives you MySQL connection settings in its properties dialog, however, you must first install the Python module MySQLdb.

On Windows, you must compile MySQLdb from source, which requires several development libraries. System engineer Yun Fu has written a very good article on installing MySQLdb on Windows which can be found at http://www.fuyun.org/2009/12/install-mysql-for-python-on-windows/

On Mac, you must also compile MySQLdb from source. It may be necessary to install Xcode from your OSX installation disk to have Apple's gcc installed. Mangoorange has written very clear instructions for doing so at http://www.mangoorange.com/2008/08/01/installing-python-mysqldb-122-on-mac-os-x/

Many Linux distributions have MySQLdb available through their package management systems. Using aptitude, you can install it via the command prompt by typing:

```
sudo apt-get install python-mysqldb
```

Note that if you want to install a local MySQL database, you will need to install the MySQL server. The MySQL Community server can be downloaded at http://www.mysql.com/downloads/mysql/. IPRStats has been tested with version 5.1 but should work with other versions.

InterProScan

In order to run IPRStats, you must supply the XML output file from an InterProScan run.

Installing InterProScan

Because IPRStats is intended for populations of genomic data and the shared InterProScan server (http://www.ebi.ac.uk/Tools/InterProScan/) is so limited, it is important to have access to your own installation of InterProScan, preferably running on a cluster. If you do not have access to an InterProScan installation, the software can be downloaded from ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/index.html. EMBL-EBI offers very good documentation on installing and running InterProScan.

Running InterProScan

Note, it is very important that certain parameters are set when running InterProScan in order for IPRStats to correctly parse the InterProScan output XML. You must enable goterms and iprlookup. Running from the command line, a run would look something like this:

```
     iprscan -cli -goterms -iprlookup -format xml -i <input fasta
file> -o <ouput xml file>
```

where `<input fasta file>` is the fasta file containing the metagenomic data and `<output xml file>` is the output file that you would open in IPRStats.

Running IPRStats

IPRStats has a number of functions and options that help you analyze InterProScan XML results. This section should help familiarize you with these functions.

Opening a file

To examine the XML result of an InterProScan run, choose File → Open from the menu bar and select the XML file. IPRStats will immediately begin parsing the file and display a progress bar. Note that when the progress bar says “Retrieving results,” the application may look frozen but is actually still running. It is recommended that you install and configure MySQL if you are opening XML files larger than 100Mb (see Installation for more details).

To open a previously saved IPRStats file (.ips), chose File → Open, change the extension filter to “IPRStats file (*.ips)” and select the file. This option will not show you a progress bar and should finish within a few minutes at most. This option should be considerably faster than opening a XML file and is useful for sharing or storing IPRStats data.

Saving a file

IPRStats gives you the option to save the data that you are viewing. This is convenient for sending the data to another person or viewing the data again at a later date. It is much faster to open an IPRStats file than to reopen an XML file.

To save a file, chose File → Save from the program's menu bar. Select the folder where you would like to save the file and give it a name. You can open this file later by choosing File → Open from the menu bar.

Properties

IPRStats has a number of different options that you can change to affect both how the data shows up in the application and what it looks like when it's exported. To change these settings, choose File → Properties from the application's menu bar.

General Settings

Chart type lets you choose whether you want to view a pie chart or a bar graph.
Chart generator specifies the style of chart.
'google' requires an Internet connection
'pylab' requires the Python modules NumPy and Matplotlib to be installed.
Max chart results specifies the top *n* number of results to be displayed in the chart.
Max table results says how many rows should be shown at a time; -1 means all rows
(note that this also limits the number of rows exported to another format)

Database Settings

Use SQLite tells whether to use the built in database (checked) or MySQL (unchecked)
Use GO lookup specifies whether to retrieve names from a Gene Ontology MySQL database

Export as HTML

Static HTML output works well to illustrate IPRStats output to people who do not have IPRStats installed, providing the chart and tabular data shown in the application. It also contains an HTML menu for navigating between different InterProScan member database pages.

From the application menu, choose File → Export as HTML... and choose the directory you wish to export to. Notice that exporting as HTML creates many files in the directory that you choose, so you may wish to create a separate folder for the output. There is no index file, but you can access all the exported data by opening any of the generated HTML files in a web browser. The easiest way to change the look of the HTML files is to edit “style.css.”

Export as XLS

IPRStats can export the tabular data from the application as a spreadsheet viewable in Excel, OpenOffice and similar programs. Choose File → Export as XLS from the application menu and type a name for the spreadsheet. The resulting spreadsheet will have a separate tab for each of the InterProScan member database.