

# PROSPERITI

## Protein Network Significance Permutation Testing – Graphical README

The python script associated with this project provides a worked example for the textbook:

Proteome Bioinformatics, Methods in Molecular Biology.  
Edited by Shivakumar Keerthikumar and Suresh Mathivanan.  
© Springer Science+Business Media LLC 2017.  
ISBN: 978-1-4939-6738-4  
DOI: 10.1007/978-1-4939-6740-7\_15

### **Chapter 15: Permutation testing to examine the significance of network features in protein-protein interaction networks.**

Written by Joe Cursons & Melissa Davis.

In particular, this script should regenerate Figure 3 presented within the aforementioned chapter.

For further details, please refer to the GitHub project page:

<http://github.com/DavisLaboratory/PROSPERITI>

Or contact the authors:

#### **Dr. Melissa J. Davis**

Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Australia  
davis.m (at) wehi.edu.au

#### **Dr. Joe Cursons**

Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Australia  
cursons.j (at) wehi.edu.au

## Overview

Section	Page
Project outline and contact details	1
Programming dependencies	2
Data dependencies	3
Script/code structure	4
Extended instructions	5

***Please note that this graphical README is still in a draft format. If you have any queries which are not addressed, please contact Joe Cursons (cursons.j (at) wehi.edu.au).***

## Programming dependencies

This script has been written using python;

For Unix users who do not have a current installation of python, it can be easily installed using an appropriate package manager such as apt-get.

For Windows users who do not have a current installation of python, you may find it easier to install a set of pre-compiled packages such as those provided by:

- WinPython       ::       <http://winpython.github.io/>

This script uses a number of python packages, including:

- pandas           ::       <http://pandas.pydata.org/>
- numpy            ::       <http://www.numpy.org/>
- networkx         ::       <http://networkx.github.io/>
- matplotlib       ::       <http://matplotlib.org/>

For experienced python users, these packages can be installed using pip. For example, from the command line, execute:

```
python -m pip install pandas
```

Alternatively pre-compiled binaries can be downloaded from Christoph Gohlke's page:

<http://www.lfd.uci.edu/~gohlke/pythonlibs/>

**NB:** some python packages require the installation of Visual C++ (available with Visual Studio 2016), and some dependencies have issues with 64-bit Windows installations.

Once installed, the python script provided with this project can be executed from the command line:

```
python.exe <scriptPath>\prosperiti.py
```

where <scriptPath> is the absolute file path (operating system dependent) for the python script – see the Graphical Description on page 5.

**NB:** to execute python from the command line in Windows systems the appropriate directory must be listed within the PATH environment variable. If multiple python installations are present, it may be easiest to execute this from within the folder (where python.exe is located) of the appropriate installation.

Alternatively, readers who wish to modify the code and control its execution (i.e. debug it) are encouraged to install an appropriate IDE, such as:

- JetBrains PyCharm       ::       <http://www.jetbrains.com/pycharm/>
- PyDev/Eclipse           ::       <http://www.pydev.org/>
- ::       <http://eclipse.org/downloads/>

## Data Dependencies

Execution of this script is dependent upon several publically available data sources, included with scientific publications/reports. The location of these data files must be specified within the data script (as the parameters strDataPath and strPINA2Path). The default values for these (i.e. data paths which will not require editing of the python script) are:

```
strDataPath = C:\doc\methods_in_proteomics
```

```
strPINA2Path = C:\db\pina2
```

### Protein-protein interaction data

The protein-protein interaction data used to generate network structures was released as part of the Protein Interaction Network Analysis (PINA) platform. In particular, we use the PINA v2.0 MITAB .tsv (tab-separated value text file) file which contains an edge list. Readers are encouraged to read the corresponding manuscript:

Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J. **PINA v2.0: mining interactome modules**. *Nucleic acids research*. 2012;40 (Database issue):D862-5.

<http://dx.doi.org/10.1093/nar/gkr967>

Data are directly available from:

<http://cbg.garvan.unsw.edu.au/pina/download/Homo%20sapiens-20140521.tsv>

**NB:** As noted above, Windows users are encouraged to create a folder at C:\db\pina2 for storing this file, to avoid the need for editing the corresponding python script.

### Phospho-protein abundance data

The phospho-protein abundance data used here come from a published phospho-tyrosine enriched quantitative MS/MS data set. Readers are encouraged to read the manuscript:

Hochgrafe F, Zhang L, O'Toole SA, Browne BC, Pinese M, Porta Cubas A, Lehrbach GM, Croucher DR, Rickwood D, Boulghourjian A, Shearer R, Nair R, Swarbrick A, Faratian D, Mullen P, Harrison DJ, Biankin AV, Sutherland RL, Raftery MJ, Daly RJ. **Tyrosine phosphorylation profiling reveals the signaling network characteristics of Basal breast cancer cells**. *Cancer Research*. 2010 Nov 15; 70(22): 9391-401.

<http://dx.doi.org/10.1158/0008-5472.CAN-10-0911>

Data (Table S3) are directly available from:

<http://cancerres.aacrjournals.org/content/70/22/9391/suppl/DC1>

**NB:** As noted above, Windows users are encouraged to create a folder at C:\doc\methods\_in\_proteomics for storing this file, to avoid the need for editing the corresponding python script.

## Script/Code Structure

The script associated with this graphical README (prosperiti.py) can be broadly separated into two sections. The python script is heavily annotated, and users are encouraged to read through the embedded comments.

The first half of the script (approx. lines 100 → 630) contains functions which have been written to process the input data files (**Extract**), create network structures (**Build**), and then calculate quantitative metrics from those networks (**Test**), together with corresponding null distributions from randomly permuted network structures.

### class Extract:

- hochgrafe\_supp\_table\_3(): extracts Supplementary Table 3 from Hochgrafe et al (2010)
- hochgrafe\_lists(): extracts specific protein lists (UniProt ID) from processed Table S3
- pina2\_mitab(): extracts the PINA v2.0 MITAB file from Cowley et al (2012)

### class Build:

- ppi\_graph(): create a networkx graph using UniProt transcript lists

### class Test:

- network\_features(): calculate quantitative metrics on the network and perform permutation testing to estimate corresponding null distributions
- edge\_correlation(): calculate the Pearson's correlation for phospho-protein abundance between edges in the PPI, and use permutation testing to estimate a corresponding null distribution

The second half of the script (approx. lines 630 → 1160) controls the execution of the functions and processing of the output to produce Figure 3 from the corresponding textbook chapter.

Lines	Contains
0630 - 0690	specify parameters (file paths, plotting parameters) for execution
0690 - 0730	execute the functions specified below
0730 - 0820	process the data
0820 - 1165	create and output the figure with the script results

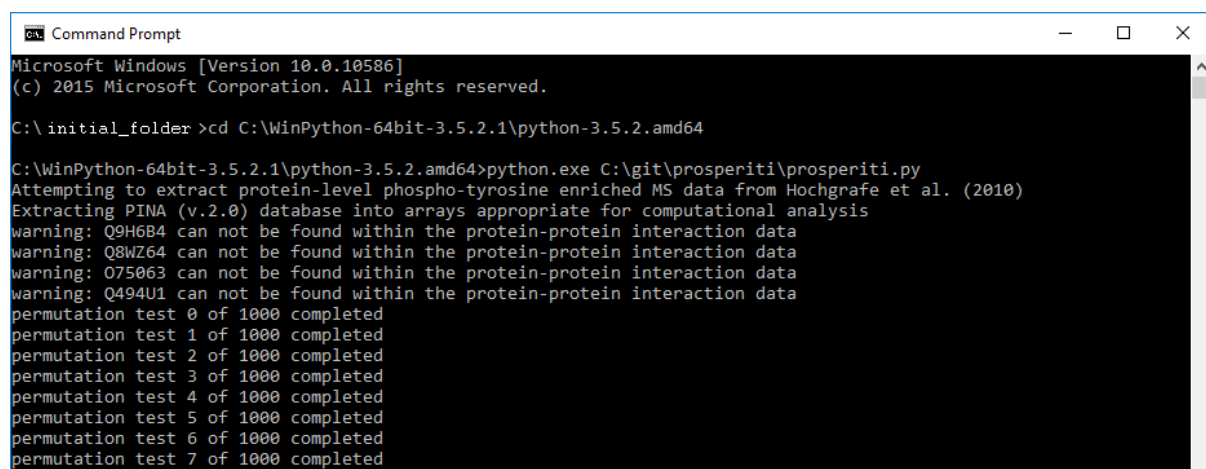
## Extended Instructions

### Command line execution

For users who have configured their system as described above (i.e. dependencies are installed correctly, data files are located in the appropriate location of the python script has been modified accordingly), this script can be executed from the command line.

In this instance, prosperiti.py can be found at “C:\git\prosperiti”.

Note that this machine has several python installations, thus this call is executed from the WinPython folder containing the python.exe file “C:\WinPython-64bit-3.5.2.1\python-3.5.2.amd64”



```
Microsoft Windows [Version 10.0.10586]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\initial_folder>cd C:\WinPython-64bit-3.5.2.1\python-3.5.2.amd64

C:\WinPython-64bit-3.5.2.1\python-3.5.2.amd64>python.exe C:\git\prosperiti\prosperiti.py
Attempting to extract protein-level phospho-tyrosine enriched MS data from Hochgrafe et al. (2010)
Extracting PINA (v.2.0) database into arrays appropriate for computational analysis
warning: Q9H6B4 can not be found within the protein-protein interaction data
warning: Q8WZ64 can not be found within the protein-protein interaction data
warning: O75063 can not be found within the protein-protein interaction data
warning: Q494U1 can not be found within the protein-protein interaction data
permutation test 0 of 1000 completed
permutation test 1 of 1000 completed
permutation test 2 of 1000 completed
permutation test 3 of 1000 completed
permutation test 4 of 1000 completed
permutation test 5 of 1000 completed
permutation test 6 of 1000 completed
permutation test 7 of 1000 completed
```

### Script output

This script produces a figure containing an array of histograms, which is comparable to Figure 3 from the textbook chapter mentioned above. Please note that the output figure was generated using nPerm=10000; however the default value within the script is nPerm=1000. This may result in minor differences for the empirical p-values calculated by the script.

By default the output figure will be saved to the same location as strDataFolder, producing a file called “CombinedHists.png”.