The IHP-PING Package Manual

Gaston K. Mazandu [1,2,3*] *et al.*

Type Package
Title: An integrated human protein-protein interaction network generator
Version 2.4.1

Contributors: Christopher Hooper[1], Kenneth Opap[2], Funmilayo L. Makinda[2,3], Victoria Nembaware[1], Nicholas E. Thomford[3,4], Emile R. Chimusa[1], Ambroise Wonkam[1] and Nicola J. Mulder[2], Gaston K. Mazandu[1,2,3*]

[1] *Division of Computational Biology, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Medical School, Anzio Road, Observatory 7925, Cape Town, South Africa.*

[2] *African Institute for Mathematical Sciences (AIMS), Melrose Road, Muzenberg 7945, Cape Town, South Africa.*

[3] *Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town (UCT), Medical School, Anzio Road, Observatory 7925, Cape Town, South Africa.*

[4] *School of Medical Sciences, University of Cape Coast, PMB, Cape Coast, Ghana.*

Maintainer Mazandu GK <gaston.mazandu@uct.ac.za, gmazandu@gmail.com, kuzamunu@aims.ac.za>

General description

In the current 'big data' driven 'post-genomic' era, in which all available sources of information can be explored at the systems level, protein-protein interactions (PPIs) are contributing to elucidation of the complex genetic architecture of organisms. These PPIs are experimentally or computationally predicted, stored in different online resources and updated regularly. As with many biological datasets, this continuously renders older protein-protein interaction (PPI) datasets potentially outdated. The integrated human protein-protein interaction network generator (IHP-PING) tool is a flexible python package, which generates a human PPI network from freely available online resources. This tool extracts and integrates heterogeneous PPI datasets to generate a unified PPI network, which is stored locally for further user applications.

Depends	Python ($\geq$ 2.7)
requires	[local ncbi-blast, python-selenium, chromium-browser and chromium-chromedriver]
License	GLP (`https://www.gnu.org/licenses/gpl-3.0.en.html`)
URL	`http://web.cbio.uct.ac.za/ITGOM/post-analysis-tools/ihp-ping-dev/` and `https://github.com/gkm-software-dev/post-analysis-tools`

Release date Thursday 21st May, 2020–05:51

*To whom correspondence should be addressed. Tel: +27 21 650 3463; Email: gaston.mazandu@uct.ac.za

Contents

IHP-PING Package	An integrated human protein-protein interaction (PPI) network generator tool, a user-friendly and accessible tool, easing integration of PPI datasets from multiple sources into a unified PPI network on-the-fly, which is stored locally for further user applications.

Description

A repository of python modules for easing integration of existing human PPI datasets from multiple sources into a unified PPI network on-the-fly (in real time), which is stored locally for further user applications. This provides a platform that enables the retrieval of human PPI datasets stored in different online resources and produced a real time up-to-date integrated human PPI network with increased accuracy, confidence and coverage.

Details

In light of the current post-genomic era, much work is being done to explore the human interactome for a systems level based analysis using protein-protein interaction networks. Though these PPI datasets are continuously curated and stored in several online freely accessible resources [1], integrating these datasets to produce a unified PPI network is still challenging. In addition, these databases are regularly updated, resulting in constantly shifting source data for these network analyses. platforms [2, 3]. These web platforms offer some advantages, e.g., ease of access and being more user friendly than terminal application interfaces. However, depending on server hosts, a web service may become unavailable at any time. Furthermore, users are required to trust the developer update and curation processes.

IHP-PING has been designed, a software for downloading and integrating currently available human PPIs, alongside using protein sequence information to predict further interactions. This software is implemented in Python modules, which can be invoked from the terminal interface on any computer or any operating system running Python, using a single command-line (refer to **Appendix 1**). It has been shown to produce an human PPI network possessing the properties of biological networks from the integration of these multiple online protein interaction datasets. The use of this software allows users to generate a human PPI network based on the most current protein interaction datasets available for humans, which likely increases the confidence or reliability of downstream network analysis. IHP-PING is a flexible tool for the generation of integrated protein-protein interaction networks on the fly.

IHP-PING retrieves PPI datasets from eight different online resources shown in Fig. 1 with complete descriptions in Table 1, including protein sequence information, which consists of protein sequences and InterPro domains [4] retrieved from the UniProt database [5] and used to predict further interactions with scores computed using an information theory-based scheme described in [6]. Each PPI is integrated with its score from its source or estimated for sources with no PPI scores, depending on the source (refer to **Appendix 2**). These interaction scores provide an indication about the confidence of predicted interactions. This is important due to relatively high noise related to high-throughput data or experiments from which interactions are inferred. So, the protein-protein interaction network produced may contain incorrectly classified interactions, i.e., fails to detect interactions (false negatives) or wrongly identifies some other interactions (false positives), which is technology-dependent. The likelihood of incorrectly classified interaction may be minimized computationally by: (1) using a data integration model, combining information from multiple interacting data sources into one unified network, and (2) applying a strict interaction reliability or confidence score cutoff. These techniques are expected

to significantly reduce the false negative and positive rate of the network produced, leading to a PPI network of high confidence interactions with an increased coverage [7].

Table 1: The resources used by IHP-PING to retrieve PPI datasets for building the an integrated human PPI network. In column 2, Arg stands for argument for each tool as used in IHP-PING package when running the tool (refer to **Appendix 1**).

Scheme	Arg	Resource	Description	URL	Ref.
STRING	`stringdb`	Search Tool for Retrieval of Interacting Genes/Proteins	A database of known and predicted protein interactions, with interactions extracted from literature, large-scale experiments, other databases, genomic context, and co-expression.	`https://string-db.org`	[8]
BioGRID	`biogrid`	Biological General Respository for Interaction Datasets	A source of literature curated human protein interactions, with high throughput and literature curated protein interactions for other species such as S. cerevisiae.	`https://thebiogrid.org`	[9]
DIP	`dip`	Database of Interacting Proteins	A collection of protein interactions obtained from literature curation of experimental data.	`https://dip.mbi.ucla.edu/dip`	[10]
HPRD	`hprd`	Human Protein Reference Database	A resource specifically focused on manual literature curation of interactions within the human proteome.	`http://www.hprd.org`	[11]
IntAct	`intact`	Open source molecular interaction database	A database of protein interactions extracted largely from large-scale experiments, with some interactions obtained through literature curation in collaboration with Swiss-Prot.	`http://www.ebi.ac.uk/intact`	[12]
MINT	`mint`	Molecular INTeraction Database	A database of protein interactions manually curated from literature, with the majority of interactions extracted from large-scale experiments.	`https://mint.bio.uniroma2.it/`	[13]
MPPI-MIPS	`mips`	The Munich Information Centre for Protein Sequences Mammalian Protein-Protein Interaction Database	A database of protein interactions manually curated from literature, with a specific focus on interactions within mammals.	`http://mips.helmholtz-muenchen.de/proj/ppi`	[14]
UniProt	`sequence`	Universal Protein knowledgebase	Centralized resource for protein sequences and functional information: A collection of manually and automatically curated protein sequences and annotations, which includes a list of human proteins reviewed by Swiss-Prot and mapping information to convert protein IDs in other databases to the IDs used by UniProt.	`http://www.uniprot.org`	[5]

It is worth noting that different sources provided here are freely accessible and, generally, different datasets are free to use for non-profit organizations. However, depending on the use of a unified PPI network being generated from these resources, the user need checks specific conditions required for each datasets. Also the user should bear in mind that there is a need of a credential identifier (ID) for validation and ID protection (VIP) services provision from the DIP Team for free access and use for non-profit organization.

IHP-PING Environment Management

The IHP-PING system is composed of one main high level folder: PyPING and one main python module, `ihppinbuilder.py`, which serves as an interface processing the human PPI interaction network and storing this network locally for further user applications. The PyPING folder includes two Python modules: (1) `networkgenerator.py`, which downloads PPI datasets requested by a user and builds the integrated PPI network and (2) `sequenceprocessing.py`, processing and scoring sequence dataset, which includes protein sequences and InterPro domains.

As a package, IHP-PING can be imported in another Python module, however, a user can directly produce a unified human PPI network in three logical steps: PPI Extraction, Mapping Process and Integration Process, via user interface and input processing using a simple single command-line terminal on any computer or any operating system running Python as described in Figure 1. The only required argument is the list of human PPI datasets to be incorporated into the unified PPI network, with output format parameters according to user preferences. Each user-requested database is downloaded from specific uniform resource locations (URLs). Eight different resources used are shown in Figure 1 and described in Table 1.

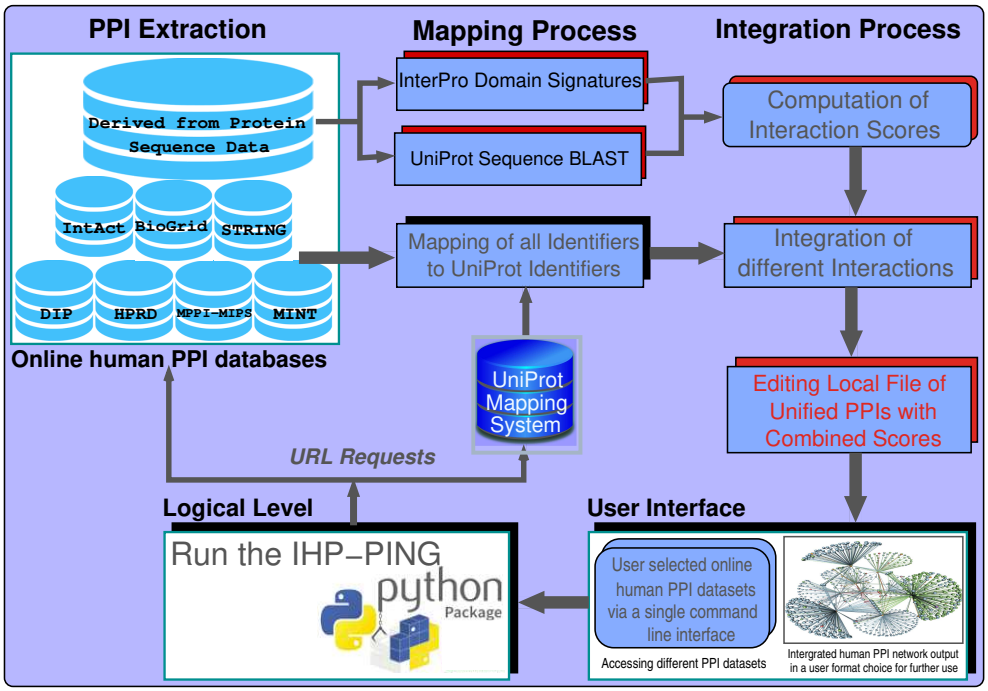


Figure 1: **Overall workflow of the IHP-PING tool**. The scheme goes through three main steps from user input to a generated human PPI network: Input is parsed via a simple single command-line terminal, then the selected human PPI datasets are retrieved and network generated in tsv, csv or csv2 format.

IHP-PING generates an output file in a tabular format, the number of columns depending on the PPI datasets retrieved with each row representing a unique PPI. The first two fields contained the IDs of the two proteins involved in the interaction, remaining columns showing the scores for the interaction from each source. The last value in the row contained the combined score of the interaction (see **Appendix 2**). Once IHPPIG has run successfully, the resultant PPI network can be used in any form of network analysis.

Appendix-1 IHP-PING Administration and Usage

2.1 IHP-PING administration

The main website for the IHP-PING package is `http://web.cbio.uct.ac.za/ITGOM/post-analysis-tools/ihp-ping-dev` where users can find essential information about obtaining IHP-PING. It is freely downloadable under GNU General Public License (GPL), pre-compiled for Linux version and protected by copyright laws, a free software and comes with ABSOLUTELY NO WARRANTY. Users are free to copy, modify, merge, publish, distribute and display information contained in the package, provided that it is done with appropriate citation of the library and by including the permission notice in all copies or substantial portions of the module contained in this package.

The whole package itself is relatively small with a total of about 0.2MB, dynamically increasing when retrieving different PPI datasets from different resources. IHP-PING contains one main module and one main folder containing modules required for generating an integrated human PPI network and formatting results to be written into a file. It is currently maintained by one member of the core-development team, Gaston K. Mazandu <gmazandu@gmail.com, gaston.mazandu@uct.ac.za, kuzamunu@aims.ac.za>, who regularly updates the information available in this package and makes every effort to ensure the quality of this information.

2.2 IHP-PING usage

IHP-PING v2.4.1 requires Linux operating system and Python ($\geq$ 2.7), requiring the installation of the `NCBI BLAST` software locally when retrieving interactions predicted from sequence data. It also requires the selenium Python package, as well as chromedriver and chromium-browser, for retrieving the DIP dataset. These needs to be installed prior to the use of IHP-PING.

To use IHP-PING, the user needs to download the 'tar.gz' file and extract all files as follows:

```
tar xzf ihp-ping-tool.tar.gz
```

or alternatively, it can also be retrieved from the github public platform using git clone command line as follows.

```
git clone https://github.com/gkm-software-dev/post-analysis-tools.git
```

After downloading and/or uncompressing, move to the folder post-analysis-tools/ihp-ping-dev/, which should be set as a working directory where IHP-PING and related commands are executed using the following terminal command:

```
cd post-analysis-tools/ihp-ping-dev/
```

2.3 IHP-PING licence and version

As pointed out previously, the IHP-PING package is free to use under GNU General Public License. You are free to copy, distribute and display information contained herein, provided that it is done with appropriate citation of the library. Thus, by using the IHP-PING package, it is assumed that you have read and accepted the agreement provided and that you agreed to be bound to all terms and conditions of this agreement. Please, use the following command line to see the package licence:

```
python setup.py --licence
```

To check the current version of the IHP-PING interface, use the following terminal command:

```
python setup.py --version
```

2.4 Running IHP-PING

As pointed out previously, IHP-PING can be processed through one main python module, `ihppinbuilder.py`, which serves as an interface. Get help on how to run IHP-PING through this interface module using the following command:

```
python ihppinbuilder.py -h
```

The above command should produce the following output:

```
usage: ihppinbuilder.py [-h] [-r list [list ...]] [-o FILE] [-f str]
```

with different tags explained below:

-h, --help	show this help message and exit
-r list [list ...], --resources list [list ...]	Database to be integrated or considered (default: all)
-o FILE, --dir FILE	Folder which will contain the PPI produced (default: current working folder)
-i str, --identifiers str	Identifier outputs: uniprot or genename (default: uniprot)
-f str, --outformat str	Output format tsv, csv or csv2 (default: tsv)

As highlighted by the `help` option, IHP-PING is run using the following one line command:

```
python ihppinbuilder.py -r resources -o outputfolder -i outputProtID -f outputfileformat
```

1. `resources`: represent different online PPI resources to be included in the unified human PPI network to be generated. Different resource arguments are shown in Table 1 and by default all resources are included.
2. `outputfolder`: The path to the folder where the outputs should be written. If not provided, the current working directory is used.
3. `outputProtID`: The identifier (ID) system to be used in the output file. If not provided, the UniProt ID system is used. IHP-PING provides two ID systems: UniProt and genename.
4. `outputfileformat`: The format of the PPI network file produced and this depends on user preferences. Three possible formats are provided: `tsv` (tab separated values), `csv` (comma separated values) and `csv2` (semi-column separated values).

2.5 Illustrating IHP-PING usage

As pointed out previously, move to the ihp-ping-dev folder and run following commands for illustration. Please type these commands manually using the computer keyboard, do not use `copy` and `paste`:

```
python ihppinbuilder.py -r sequence mips
```

This produces a tsv format file of human PPI network derived from sequence data and MPPI-MIPS online database saved in the working directory, which is ihp-ping-dev folder.

The second and the third commands building some human PPI networks are given below:

```
python ihppinbuilder.py -f csv
python ihppinbuilder.py -r stringdb biogrid dip -i genename -f csv2
```

The second command will generate a unified human PPI network derived from all sources under consideration currently, and save under the working directory (default) in csv format. For the third command, only STRING, BioGrid and DIP databases are used and the network is saved as a csv2 (semi-column separated value) file with the gene name ID system.

2.6 Running IHP-PING as a Python Package

As any python library or package, IHP-PING can be imported and used in another Python models. For accessing and learning about different classes of the two main modules under PyPING, networkgenerator providing functions for downloding and integrating the human PPI network and sequenceprocessing for processing sequence datasets (protein sequences and InterPro datasets). Please access the python interpreter or the command shell for interactive computing (IPython) and run following commands:

```
>>> from PyPING import *
>>> help(networkgenerator)
>>> help(sequenceprocessing)
```

2.7 Important notes

- To efficiently use the IHP-PING library and to maximally benefit from its use, make sure that you have carefully read this PDF package documentation file, which is provided in the library.
- In some cases, you may need or be required to provide the folder in which interaction file should be. Please make sure that the full path to the folder target is provided.
- Make use of the full screen mode when displaying results on it for a nice and more adapted visualization.

2.8 Contributors

Christopher Hooper, Kenneth Opap, Funmilayo L. Makinda, Victoria Nembaware, Nicholas E. Thomford, Emile R. Chimusa, Ambroise Wonkam and Nicola J. Mulder, Gaston K. Mazandu

Maintainer: Mazandu GK <gmazandu@gmail.com, gaston.mazandu@uct.ac.za, kuzamunu@aims.ac.za>

2.9 Main references

1. Mazandu GK, Chimusa ER, Rutherford K, Zekeng EG, Gebremariam ZZ, Onifade MY, Mulder NJ. Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. Brief Bioinform 2018, 19(6), 1141–1152.

2. Mazandu GK, Mulder NJ. Using the underlying biological organization of the Mycobacterium tuberculosis functional network for protein function prediction. Infection, Genetics and Evolution 2012, 12(5), 922–932

3. Mazandu GK, Mulder NJ. Generation and Analysis of Large-Scale Data-Driven Mycobacterium tuberculosis Functional Networks for Drug Target Identification. Advances in Bioinformatics 2011, 2011(Article ID 801478), 14 pages. Doi:10.1155/2011/801478.

4. Mazandu GK and Mulder NJ. Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One 2011, 6(4), e18607.

2.10 Questions, Comments and Report Bugs

The IHP-PING team is striving to aggregate knowledge about the human protein-protein interaction online datasets on-the-fly in realistic timeframes, providing a solution to producing a real time or up-to-date unified human PPI network. However, IHP-PING does not guarantee the quality or accuracy of different result outputs. Thus, if it happens that you find errors, please contact the primary source of data set used for more information. If you feel that the errors may be due to some systematic error in the PySML library, please contact the library maintainer at <gmazandu@gmail.com, gaston.mazandu@uct.ac.za, kuzamunu@aims.ac.za>.

2.11 IHP-PING copyright and license

The IHP-PING library is free to use under GNU General Public License (GLP: `https://www.gnu.org/licenses/gpl-3.0.en.html`). You are free to copy, distribute and display information contained herein, provided that it is done with appropriate citation of the tool.

2.12 Citing IHP-PING.

The manuscript is being prepared for publication, before its publication you can cite the preliminary report:

"Hooper C, Opap K, Makinda FL, Nembaware V, Thomford NE, Chimusa ER, Wonkam A, Mulder NJ, Mazandu GK. IHP-PING–A flexible tool for generating integrated human protein-protein interaction networks on-the-fly". Technical report 2020, H3ABioNet-AIMS node and SADaCC, AIMS & UCT, South Africa. `http://web.cbio.uct.ac.za/ITGOM/post-analysis-tools/ihp-ping-dev/`.

2.13 Other information about IHP-PING development

Please refer to `http://web.cbio.uct.ac.za/ITGOM/post-analysis-tools/ihp-ping-dev/PKG-INFO` (See some other details about the IHP-PING development)

or

go to the local ihp-ping-dev folder and type the following command line for the short description:

```
python setup.py --description
```

or alternatively,

```
python setup.py --long-description
```

for the long description.

Appendix-2 Human PPI network in the post-genomic context

3.1 The post genomic context

This modern period of science has been described as the post-genomic era [15], where all available sources of information are explored at the biological systems level [16, 17]. This is in stark contrast to the older genocentric view [15], which sought to describe simple links connecting individual genes to particular phenotypes [18]. Considering then the many genes which may contribute to disease, it is understandable that modern attempts to associate phenotypes with genetic variations do so on a genomic scale, using interactome, especially in the case of complex diseases, such as cancer [16]. Initially, the discovery of disease-associated genomic variations was achieved through the use of linkage analysis and genetic association tests based on low numbers of genetic markers [19]. Later, the number of genetic polymorphisms that could be used as genetic markers in such tests exploded with, for example, DNA microarray chips enabling the identification of thousands of novel genetic variants in the form of single-nucleotide polymorphisms for any particular species [20].

3.2 Protein-Protein Interaction network generation and analysis

The complete set of physical protein-protein interactions (PPIs) within a cell is defined as the interactome [21, 22, 23]. Currently, one approach to exploring the interactome is through the generation and analysis of PPI networks [24, 25, 26, 27]. These networks display proteins and the interactions between protein pairs in the form of a mathematical graph, which consists of edges and nodes [7, 17]. Interactions between protein pairs can be predicted experimentally by high-throughput yeast two-hybrid screens and/or mass spectrometry [28]. Alternatively, interactions can be inferred from literature [29, 30], or predicted from sequence data [6]. These PPI datasets are continuously curated and stored in several online resources [1], including STRING [8], IntAct [12], MINT [13], BioGrid [9], DIP [10], HPRD [11], and MPPI-MIPS [14], and updated frequently. Thus, new versions are often released, rendering older networks potentially outdated. Thus, the generation of an aggregate network in real-time would aid researchers in producing research outputs that are continually based on current information.

The above sources generally list pairs of proteins, usually accompanied by references to the literature that documents these interactions with a confidence score for that interaction. There are some differences between the data stored within these resources, but they each contain the interaction data required to perform analyses of human PPI networks. Current research and analysis involving these networks produce results which are potentially based on reactions that are likely outdated as they may not include novel identified interactions in the results released or published. Researchers are unaware of any software which generates complete human PPI networks on demand, while possessing a tool to generate an aggregate network in real-time would allow researchers to more readily apply known protein-protein interaction data in the generation and testing of hypotheses. Specifically, PPIs can be applied in, for example, protein function prediction [22], candidate gene or target prediction [31] and prioritization, post genome-wide association analyses [32], prediction of disease phenotype trends, and identification of disease related genetic patterns or properties [17].

3.3 Integrating PPI Datasets and Different Platforms

The use of any network is likely to contain incorrectly predicted PPIs [7] due to relatively high noise related to high-throughput data, source of these PPIs. PPI network coverage and likelihood of false negative interactions may be optimized by combining data sources, thus some have attempted to

generate new databases which integrate PPI information from some or all of the above sources. Such attempts include the Protein Interaction Network Analysis (PINA) platform [33], the Integrated Interactome System (IIS) [34], the Unified Human Interactome database (UniHI) [35], Protein InteraCtion KnowLedgebase (PICKLE 2.0) [2] and Molecular Interaction Search Tool (MIST) [3].

The common thread linking most of the above approaches to integrated PPI network generation and analysis is firstly, the access via a web interface and secondly, the existence of prebuilt databases which are reportedly updated regularly. Though these approaches have some advantages, they also present some limitations to the users of these services. Web platforms are considered to be easily accessible and more user-friendly when compared to terminal interfaces or application software that requires an initial installation. Web platforms allows for users to access the platform or service at any device with an internet connection. The limitations of a web platform include the requirement of a stable and constant internet connection and that the resource is dependent on the server host. At any time, the web resource may become unavailable to the user through the actions of the web hosting service.

With regards to pre-constructed databases, the benefit of their use is that queries are fast as no dataset integration steps need to be processed; the data is immediately available for use. The cost to this increased speed is the loss of user customisation as the user is not able to specify which datasets they would like to be included in the database, though UniHI (Kalathur et al., 2014) attempts to mitigate this by informing users of the original resource from which an interaction was extracted. By using pre-constructed databases, the user is also required to trust the curation process of the database architects.

As an attempt to avoid pre-constructed databases, and the issues resultant from PPI datasets being regularly updated, we present the Integrated Human PPI Generator tool (IHP-PING). IHP-PING is a software tool which generates a human protein-protein interaction (PPI) network from freely available online resources in real-time. This tool is considered to be user-friendly and accessible to biologists so that it may be used without extensive training in software applications. The tool downloads and integrates the PPI data of multiple sources to generate a PPI network which is stored locally for the user. While an internet is required for the download of the datasets specified by the user, any downstream analysis of the network produced can be run locally without an internet connection.

3.4 Implementing IHP-PING

IHP-PING is implemented as a Python package which can be run through a command-line terminal (see **Appendix-1**). In order to use IHP-PING, the user invokes the script through Python alongside command arguments. The main argument to be provided is the list of datasets (see Table 1) to be incorporated into the final unified network, with other parameters changing the operation of the program according to user preferences, such as the format of the output file (see **Appendix-1**). Each requested database is downloaded from specific Uniform Resource Locations (URLs) and there are different types of resources which may be integrated by IHP-PING, including PPI databases and PPIs predicted from protein sequences and signatures. The PPI datasets which are requested in the current IHP-PING version are presented in Table 1 and different main steps shown in Figure 1.

PPI datasets are retrieved sequentially, stored in memory with each interaction being extracted from the downloaded files, cleaning the memory space for each source, once PPI extraction process is done. IHP-PING stores these interactions in the output file alongside a score for each interaction,

which is calculated differently depending on the dataset from which the interaction was obtained. In the case of MINT and STRING, there is an interaction score within the dataset which is extracted by IHPPIG and entered into the output directly. The HPRD dataset does not contain interaction scores but, for each interaction, it lists the publications and evidence sources that support the interaction, which are then used to estimate the reliabibility or confidence score. Given proteins p and q, this score is calculated as follows:

$$s_{pq}(n) = 1 - \frac{1}{n} \tag{1}$$

where n is the total number of confidence sources and publications.

BioGRID, DIP, IntAct, and MPPI-MIPS datasets do not contain interaction scores and thus a default score of 0.7 for DIP and 0.6 for others is assigned by IHP-PING, which the authors set based on the confidence level about the database under consoderation. Additionally, the interactions predicted from protein sequences receive a score according to sequence similarity and shared protein signatures computed using an information theory-based scheme described in [6], calculating the cumulative standard normal distribution function, $\phi(x)$, as:

$$\phi(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{2}$$

with $\text{erf}(z) = \frac{1}{\sqrt{\pi}} \int_{-z}^{z} \exp\left(-t^2\right) dt$, the Gauss error function implemented in the Python math library.

After calculating the reliability or confidence score for each functional association protein pair, the combined confidence score $\boldsymbol{s_{pq}}$ for interacting proteins p and q using the followinf formula [17, 24]:

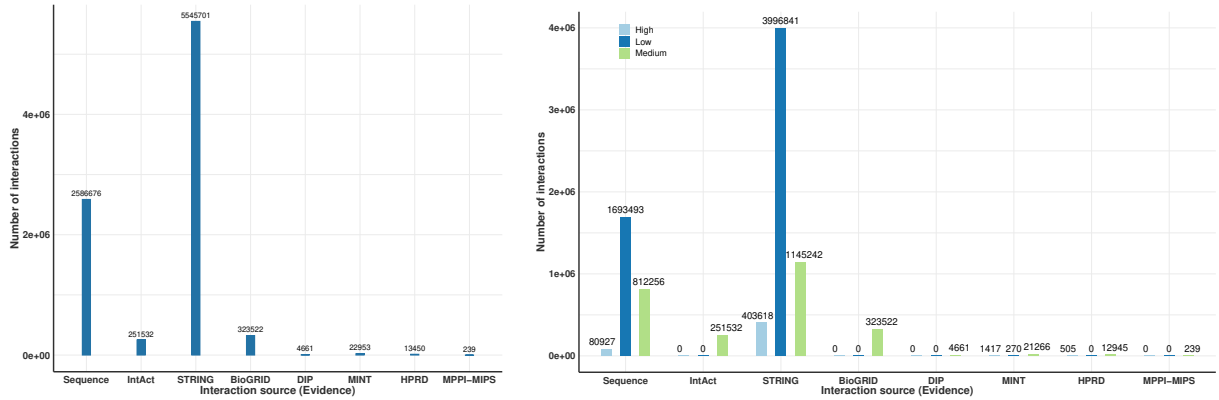
$$\boldsymbol{s_{pq}} = 1 - \prod_{d=1}^{r} (1 - s_{pq}^{d}), \tag{3}$$

where r is the total number of PPI data sources and s_{pq}^{d} is the confidence score of an interaction between p and q retrieved from the PPI data source d. Thus, for minimizing the likelihood of false positive interactions, a reliability cutoff can be applied, which may lead to highly reliable PPI network.

Naturally, there are some issues when integrating interactions from multiple sources as datasets often use different protein identifier (ID) systems. For example, STRING [8] uses a unique ID system while DIP [10] includes its own protein ID alongside the corresponding UniProt protein ID [5]. Different protein identifiers are mapped to reviewed proteins only from Swiss-Prot under the non-redundant UniProt identifier system for harmonization before integration. Once all IDs have been mapped successfully, the interactions from each database are integrated into a single data frame. The final output of IHP-PING is a local file stored in the desired directory which contains all PPI information extracted from the datasets, namely the two protein IDs and the score from each source with a combined score in the last column for each interaction. This file may then be used to perform downstream PPI network analysis.

3.5 IHP-PING Results and Basic Network Topological Properties

IHP-PING retrieved PPIs from all the sources and generated an output file in a tabular format with 11 columns with each row representing a unique PPI. The first two fields contained the IDs



(a) Total number of interactions per source.

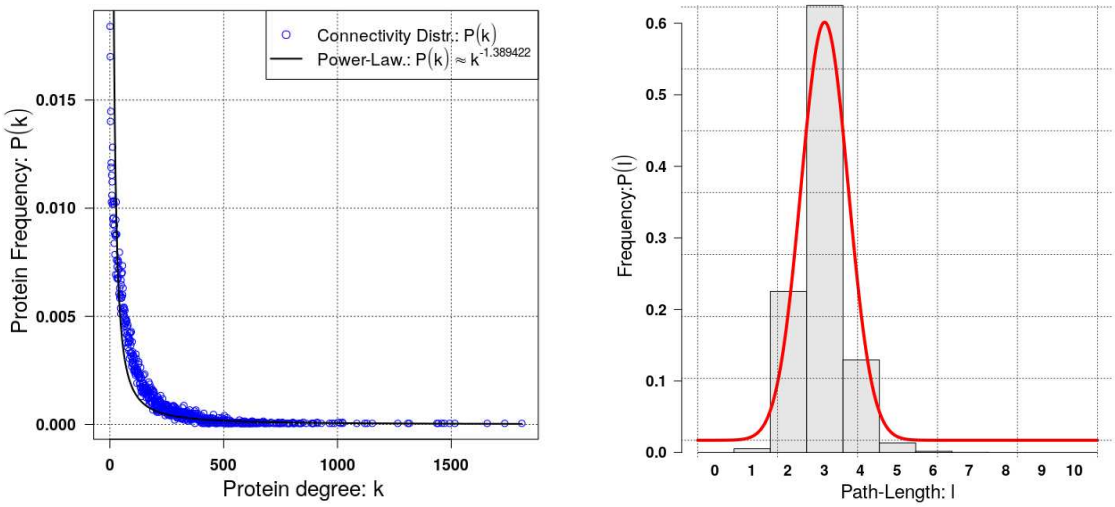
(b) Confidence levels based interaction distributions.

Figure 2: Distribution of interactions obtained from different resources contributing to a unified human PPI network–All interactions per source in (a) and high-low-medium confidence levels interaction frequencies in (b).

of the two proteins involved in the interaction, with columns 3 through 10 showing the scores for the interaction from each source. The last value in the row contained the combined score of the interaction. The total number of interactions obtained from each source is shown in Figure 2(a), distribued in low, medium and high confidence levels, with score less than 0.3, ranging between 0.3 and 0.7, and greater than 0.7, respectively, in Figure 2(b).

The human PPI network generated in this instance contained 8 017 087 interactions connecting 19 957 proteins out of 20 366 reviewed human proteins with a total of 5 276 025 interactions with low confidence, 2 045 319 with medium confidence and 695 743 with high confidence level. In analysis of these interactions, 51 466 interactions with low confidence (interaction score less than 0.3) were predicted by at least two different datasets. As pointed out previously, due to relatively high noise related to high-throughput data or experiments from which interactions are inferred, the protein-protein interaction network generated may contain incorrectly classified interactions, i.e., failing to detect interactions (false negatives) or wrongly identifying some other interactions (false positives). To minimized the likelihood of incorrectly classified interaction computationally by: (1) using a data integration model, combining information from multiple interacting data sources into one unified network, and (2) applying a strict interaction reliability or confidence score cutoff. These techniques are expected to significantly reduce the false negative and positive rate of the network produced, leading to a PPI network of high confidence interactions with an increased coverage [7].

Here, we used high confidence human PPI network, extracted from the unified network generated, considering only interactions with score > 0.7 or predicted by two different sources, to check general topological properties of the bioogical networks, namely power-law and small-world properties. This network consisted of 960514 interactions linking 19345 proteins. The distribution of degree plotted is shown in Figure 3(a) and distribution of path lengths within the network in Figure 3(b). The power exponent, γ, was estimated to 1.38942 with p-value < 0.0001, implying that the network obtained fits perfectly the power-law property. Furthermore, analyzing in terms of the distribution of path lengths within the network shows that with average path length of 2.92607 ≈ 3. These results indicates that the human PPI network conforms to the properties of biological networks.



(a) The network power-law property.

(b) The network small world property.

Figure 3: The unified human PPI network topological properties: (a) Power-law property–visualizing protein degree against connections frequency in the network and (b) Small-world property–the distribution of path length within the interaction network.

3.7 Final Remarks

In the rapidly expanding field of bioinformatics, datasets are dynamic, and the software tools used to analysis these data should be flexible and extensible in order to adequately manage the regular database updates. IHP-PING presents a solution for integrating and generating protein interaction networks, with clear benefits when compared to similar solutions.

Firstly, while IHP-PING requires a stable internet connection at runtime, it is not implemented as a web platform. Web-based tools do have certain advantages such as ease of access but, in the case of large-scale protein interaction networks, a network stored locally can be analysed on local hardware, reducing the user's reliance on server hardware of the platform and allowing the user to make use of any available cluster resources. The network produced by IHP-PING is stored locally and can integrate different datasets into the output network according to the user's preference. Secondly, the modular structure of the code allows for the extension of IHP-PING to support additional datasets. Beyond supporting new PPI databases as they become hosted online, any novel method that predicts protein interactions could theoretically be incorporated into the software tool presented here given that a protein pair with an interaction score can be obtained. In the opposite fashion, datasets that become deprecated can be easily excluded by the user from the network.

In considering the output network, it is clear that the majority of interactions extracted by IHP-PING from PPI datasets come from STRING and predicted from sequence data. This is intuitively due to the additional computational approaches used to predict some of PPIs, which is not the case for other datasets. We believe that the number of interactions that are curated from literature will increase over time as more experiments are performed and published. IHP-PING is uniquely equipped to retrieve the most recent data from its supported resources, thus once a supported PPI database is updated, subsequent runs of IHP-PING will generate a new network based on the updated information. Therefore, the number of interactions obtained by the software is likely to increase over time.

Finally, advances in high-throughput technology have enabled the generation of tissue-specific gene expression information and the inclusion of this information may improve the coverage of the network produced and reduce false negative PPIs. In its current form, IHP-PING does not directly include gene expression information, even though STRING supports gene expression information and retrieving PPIs from STRING implies that this information is implicitly included. There is a need for IHP-PING to support explicitly gene expression information as it is the case for sequence data – This is an area of potential future expansion. In addition, there is a need of visualisation technique which supports drawing of a graph or network from edge and node data, and which is compatible with the output of IHP-PING, supporting the flexibility of the software. This is also an area of future work, where we will assess the dynamic python-networkx and python-matplotlib libraries to implement a graphical user interface (GUI) to support a systematic network visualization.

References

[1] Wu, Z., Liao, Q., and Liu, B. (2019) A comprehensive review and evaluation of computational methods for identifying protein complexes from protein-protein interaction networks. *Brief Bioinform* **pii:bbz085**, bbz085.

[2] Gioutlakis, A., Klapa, M. I., and Moschonas, N. K. (2017) PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology. *PLoS One* **12(10)**, e0186039.

[3] Hu, Y., Vinayagam, A., Nand, A., Comjean, A., Chung, V., Hao, T., Mohr, S. E., and Perrimon, N. (2018) Molecular interaction search tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res.* **46(D1)**, D567–D574.

[4] Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., and Bridge, A. (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360.

[5] UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515.

[6] Mazandu, G. K. and Mulder, N. J. (2011) Scoring protein relationships in functional interaction networks predicted from sequence data. *PLoS One* **6(4)**, e18607.

[7] Mazandu, G. K. and Mulder, N. J. (2011) Generation and analysis of large-scale data-driven mycobacterium tuberculosis functional networks for drug target identification *Advances in Bioinformatics* **2011(Article ID 801478)**, 14 pages.

[8] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., and Huerta-Cepas, J. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47(D1)**, D607–D613.

[9] Oughtred, R., Stark, C., Breitkreutz, B., Rust, J., Boucher, L., and Chang, C. (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.,* **47(D1)**, D529–D541.

[10] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451.

[11] Keshava, P. T. S., R, G., Kandasamy, K., Keerthikumar, S., Kumar, S., and Mathivanan, S. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–772.

[12] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., and Broackes-Carter, F. (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42(D1)**, D358–363.

[13] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Lannuccelli, M., and Galeota, E. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861.

[14] Mewes, H. W., Ruepp, A., Theis, F., Rattei, T., Walter, M., and Frishman, D. (2006) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220–D224.

[15] Perbal, L. (2015) The case of the gene: Postgenomics between modernity and postmodernity. *EMBO Rep* **16**, 777–781.

[16] Mazandu, G. K., Kyomugisha, I., Geza, E., Seuneu, M., Bah, B., and Chimusa, E. R. (2019) Designing data-driven learning algorithms: A necessity to ensure effective post-genomic medicine and biomedical research. In Artificial Intelligence - Applications in Medicine and Biology IntechOpen Publisher : pp. 3–18.

[17] Mazandu, G. K., Chimusa, E. R., Rutherford, K., Zekeng, E. G., Gebremariam, Z. Z., Onifade, M. Y., and Mulder, N. J. (2018) Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. *Brief Bioinform* **19(6)**, 1141–1152.

[18] Beadle, G. W. and Tatum, E. L. (1941) Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci.* **27**, 499–506.

[19] Altshuler, D., Daly, M. J., and Lander, E. S. Genetic Mapping in Human Disease. *Science* **322(5903)**, 881–888.

[20] Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Steinm, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D., and International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409(6822)**, 928–933.

[21] Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005) Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14(2)**, R171–181.

[22] Mazandu, G. K. and Mulder, N. J. Using the underlying biological organization of the mycobacterium tuberculosis functional network for protein function prediction. *Infection, Genetics and Evolution* **12(5)**, 922–932.

[23] Mazandu, G. K., Opap, K., and Mulder, N. J. Contribution of microarray data to the advancement of knowledge on the mycobacterium tuberculosis interactome: Use of the random partial least squares approach *Infection, Genetics and Evolution* **11(4)**, 725–733.

[24] Akinola, R. O., Mazandu, G. K., and Mulder, N. J. (2016) A quantitative approach to analyzing genome reductive evolution using protein–protein interaction networks: A case study of mycobacterium leprae. *Front Genet* **7**, 39.

[25] Mulder, N. J., Akinola, R. O., Mazandu, G. K., and Rapanoel, H. (2014) Using biological networks to improve our understanding of infectious diseases *Computational and structural biotechnology journal* **11(18)**, 1–10.

[26] Rapanoel, H. A., Mazandu, G. K., and Mulder, N. J. (2013) Predicting and analyzing interactions between mycobacterium tuberculosis and its human host. *PLoS One* **8(7)**, e67472.

[27] Mazandu, G. K. and Mulder, N. J. (2012) Function prediction and analysis of mycobacterium tuberculosis hypothetical proteins. *International journal of molecular sciences* **13(6)**, 7283–7302.

[28] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968.

[29] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. Literature-curated protein interaction datasets. *Nat. Methods* **6(1)**, 39–46.

[30] He, M., Wang, Y., and Li, W. (2009) Plos one *PPI Finder: A Mining Tool for Human Protein-Protein Interactions.* **4(2)**, e4554.

[31] Li, X., Li, W., Zeng, M., Zheng, R., and Li, M. (2019) Network-based methods for predicting essential genes or proteins: a survey. *Brief Bioinform* **pii: bbz017**, bbz017.

[32] Chimusa, E. R., Dalvie, S., Dandara, C., Wonkam, A., and Mazandu, G. K. (2019) Post genome-wide association analysis: dissecting computational pathway/network-based approaches. *Brief Bioinform* **20(2)**, 690–700.

[33] Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., Biankin, A. V., Hautaniemi, S., and Wuet, J. (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res.* **40**, D862–D865.

[34] Carazzolle, M. F., de Carvalho, L. M., Slepicka, H. H., Vidal, R. O., Pereira, G. A., Kobarg, J., and Meirelles, G. V. (2014)) IIS – integrated interactome system: A web-based platform for the annotation, analysis and visualization of protein-metabolite-gene-drug interactions by integrating a variety of data sources and tools. *PLOS One* **9(6)**, e100385.

[35] Kalathur, R. K., Pinto, J. P., Hernández-Prieto, M. A., Machado, R. S., Almeida, D., Chaurasia, G., and Futschik, M. E. (2014) UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. *Nucleic Acids Res.* **42**, D408–D414.