PiType User Manual

Created By: Florian Goebels & Michael Wrana
Last Updated: Thursday, June 29, 2017

Contents

Introduction

PiType is an open-source, machine learning tool designed to classify protein-protein interactions within a network into two categories: Obligate (O) and Non-Obligate (NO). This manual is designed to help you understand all the features that come with PiType, as well as provide information on what exactly happens behind the scenes while the program is running. If you would like to examine our source code, it can be found on Github.

Usage

First-Time Installation

When starting up PiType, a first-time installation dialog will appear. This is required because the machine learning algorithms needed for classification are platform-dependent.

Automatic

If automatic is selected, PiType will simply handle the Installation by itself, no input required. This requires a 1.5GB download and may take a long time depending on internet speed. If it appears the program has frozen because the loading bar has stopped, it probably hasn't. Some of the installation processes take a long time, and if any error occurs, a dialog menu will appear describing the issue.

The automatic installation works by downloading a program called Miniconda, which can act as a python environment. Miniconda also allows PiType to automatically install all the machine learning libraries and tools it needs. All of this is stored within the PiTypeUtils folder that is created after you first run the installer.

Custom

The custom installer is only recommended for advanced users who understand Python and how it is installed onto a machine. If chosen, a PiTypeUtils folder will still be created to contain the Python scripts associated with PiType, but finding or creating a compatible Python installation is left to the user. It is important to note: PiType REQUIRES PYTHON 2.x NOT PYTHON 3.x

Input and Output

As an input PiType takes a list of protein-protein interactions, and outputs whether it thinks each interaction is Obligate or Non-Obligate. This list is generated automatically from the currently selected network when you run from within Cytoscape. There are three main sources of information that are displayed once PiType has finished running: classification, confidence score, and mappings.

Classification

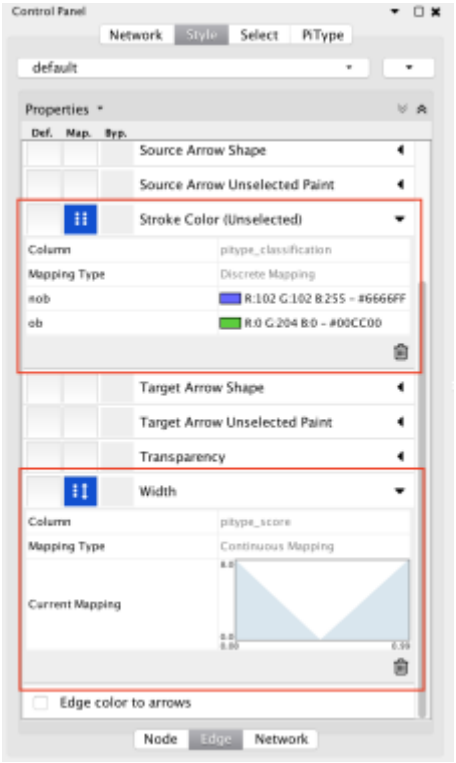
After running, PiType will classify each edge as either obligate or non-obligate. This is added as a column in the edge table of Cytoscape. An interaction which has not been classified is left as "--".

Confidence Score

Next to the classification there will be a decimal value between 0 and 1, listed as PiType Confidence in the edge table. This score represents how sure the machine learning algorithm is that its determination is correct. A value close to 0.5 means it is not very sure as to whether the interaction is O or NO. A value closer to 0.0 means more likely NO, and a value closer to 1.0 means more likely O.

Mappings

PiType will automatically represent its output visually in the current network. The color of each edge represents its classification: blue is NO, green is O. The thickness of each edge represents how confident the classifier is in its determination, where the thicker the edge, the more confident it is. These mappings can be changed in the Control Panel by clicking on the "Style" tab, then selecting "Edge", and modifying the "Stroke Color (Unselected)" and "Width" sub-menus:



Taxonomy

The taxonomy Text Field includes an auto-complete feature for known taxonomies. Either the taxonomy number or name can be entered e.g. "9606" or "Homo sapiens" for Homo sapiens. Currently PiType supports all species supported by String. If an error appears stating the species entered is invalid, try again using the auto-completer to avoid typos or incorrect formatting.

Features

ELMs

It has been suggested that interactions in eukaryotic organisms that are mediated by short linear sequence motifs tend to be non-obligate. To determine the number of ELMs for each protein PiType downloads the ELM database and searches each protein sequence for all occurrences of each ELM. Hence each interaction is characterized by two integer values giving the numbers of ELMs found in both interaction partners.

Degree

The degree of an edge *e* is the number of edges that share at least one node with *e*. In other words, the degree of an edge between nodes v_1 and v_2 is the number of edges that have at least v_1 or v_2 as a node. Proteins that have a very large degree tend to be NO, whereas proteins with a small degree tend to be O. To find the degree of each protein PiType uses the IntAct and BioGrid databases.

String

Rather than analyze physical properties of protein interaction, String DB uses different scoring methods. This database scores each interaction using up to eight different metrics, then combines them into an overall score, which is used by PiType. If you would like more information about each metric, it can be found here.

Functional Similarity

Functional similarity between two proteins is calculated based on their associated Gene Ontology (GO) annotation using this method as implemented in the GOSemSim package. To calculate the functional similarity between a protein A with GO terms GO_1 … GO_i, and a protein B with GO_1 … GO_j, every term in A is compared with every term in B, yielding a matrix with *i* rows and *j* columns corresponding to each GO term of A and B. Functional similarity between A and B is then the mean over the maxima of each row and column of *m*.

Machine Learning Methods

Random Forest

The main machine learning package used for PiType is scikit-learn, and its associated Python API. From all tested classifiers, the decision tree method achieved the best performance in distinguishing between permanent and transient interactions. Moreover, the random forest algorithm has a better accuracy and is more robust than the decision tree approach. PiType uses the random forest classification algorithm with an ensemble of 10 decision trees. These trees are used to create a confidence value for each predicted class *c*, which lies between 0 and 1. This value describes the fraction of decision trees that voted for class *c*. This value is also the PiType Score column produced after running the algorithm.

FAQ

Do I need an Internet connection?

Yes. PiType requires an internet connection to access the online databases which are used to get more information about the proteins you are trying to classify. The first-time installer also requires an internet connection if automatic is selected.

It's taking longer than predicted!

The predictions were made on one test device, and are only meant as guidelines. If your classification time is longer than estimated that is likely because your computer is not as fast as the test device.

The Custom Installation Isn't Working!

Directory Selected

The python installation needed by PiType is NOT the folder where you installed python, but the executable called when running a python script. To determine where this is on your computer try typing a command like `which python` and using that result instead.

Wrong Python Version

The default version on your machine might be python 3, whereas PiType requires python 2. In that case, you must find the command that is used to execute a python 2 script. Additionally, when installing the external modules, ensure you are installing them onto the correct python version.

File Doesn't Exist

If you receive an error saying the given file does not exist, try removing any whitespace from the text box, such as line breaks before or after, as well as tabs and spaces.

Permission Denied Error

Try running Cytoscape with administrator permissions. The reason for this error is that PiType needs to write files into the PiTypeUtils folder, and if it does not have permission an error will be thrown.

File Not Found Error

Try deleting the PiTypeUtils folder which will be in your user folder. This will restart the PiType first time installation and should replace any missing files.

Run Timings

Number of Edges (Protein Interactions)	Approximate Time to Classify (Minutes)
1	10
100	12
500	23
1000	0
2000	0
4000	0
8000	0