VISTA README

Samvida S. Venkatesh

VISTA - Visualization of Interactions in Space and Time Analysis tool

Introduction

VISTA (Visualization of Interactions in Space and Time Analysis tool) is an interactive program for visualization and exploratory data analysis of dynamic protein-protein interaction (PPI) datasets. VISTA takes advantage of cutting-edge graph visualization algorithms and automatic programmatic access to protein databases to deliver an intuitive and user-friendly data visualization platform. The tool enables users to discover underlying patterns in multiple interaction networks across conditions using an interactive interface with dynamic network visuals and exploratory quantitative analysis. Users can integrate interaction networks of one or many baits, include subcellular localization information and functional annotations to modify networks, analyze dynamic quantitative properties of the network, and identify properties of proteins shared between multiple baits - all with their own data, with few manual steps.

Motivation

There is a growing need for integrated visualization and analysis of quantitative interaction data across different conditions and from multiple datasets. Comparing large-scale interactome changes and investigating trends in quantitative measurements of PPIs across conditions can drive novel hypotheses on biological functions. The inclusion of spatial proteomic information, such as subcellular localization, and functional classification via gene ontology enrichment and protein complex identification further increases the power of such assessments.

We built VISTA to specifically address the spatially and temporally heterogenous proteome of the cell during viral infection with human cytomegalovirus (HCMV). Many of the example datasets packaged with the program draw on that data. However, VISTA can be applied to any PPI datasets with measurements across multiple conditions.

Select Screenshots (see next page)

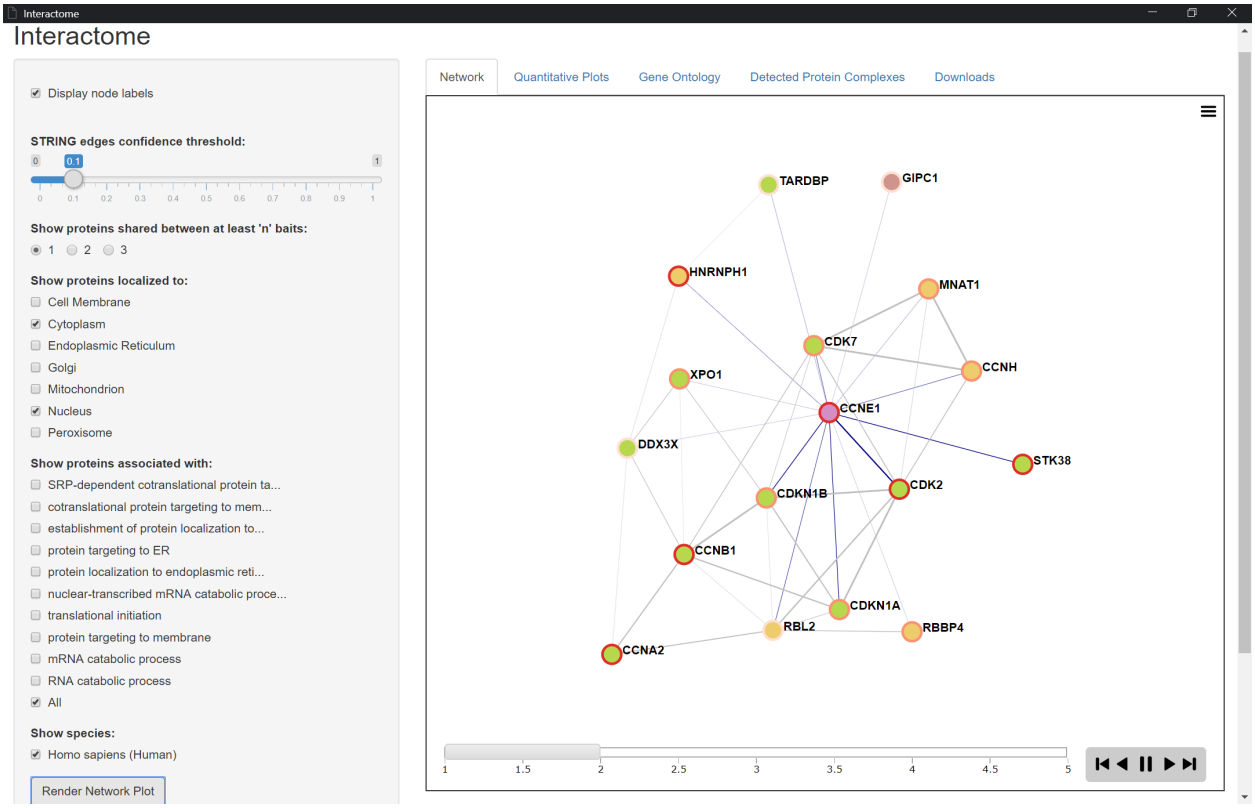

Figure 1: Network Plot

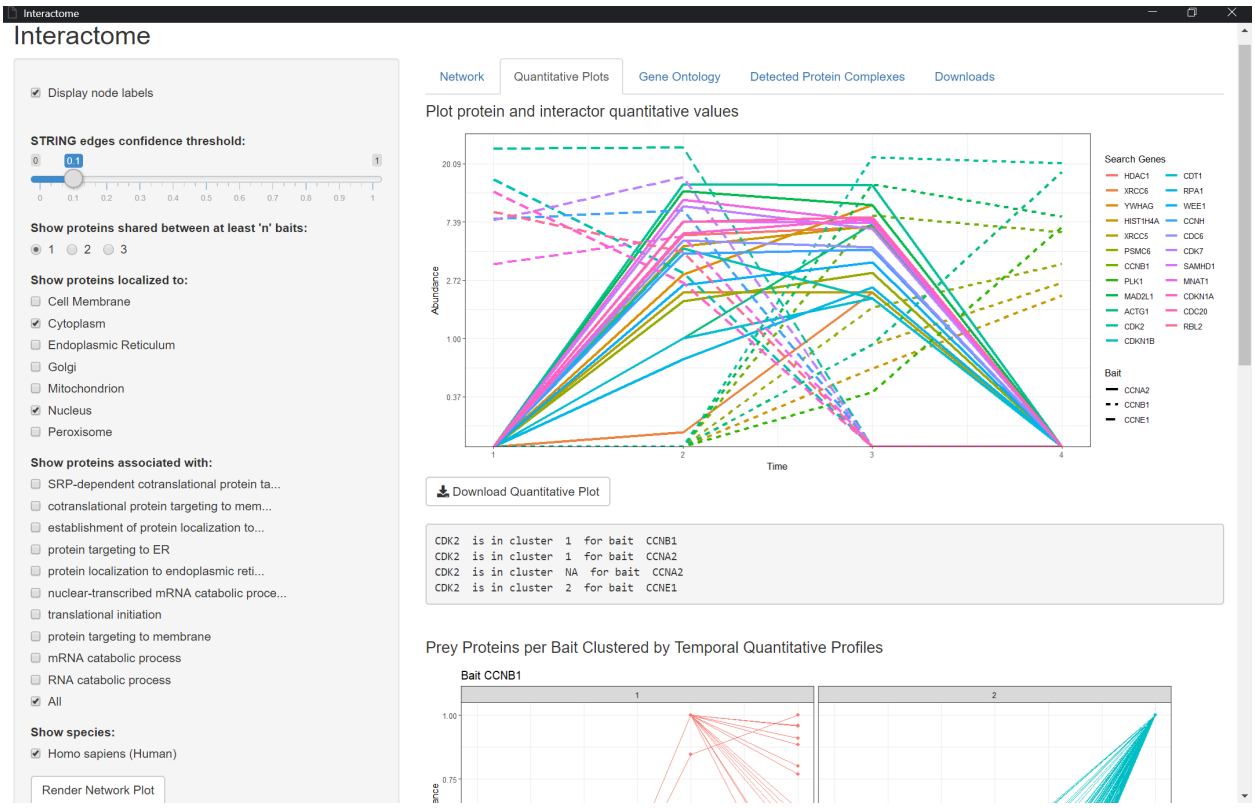

Figure 2: Quantitative Plot

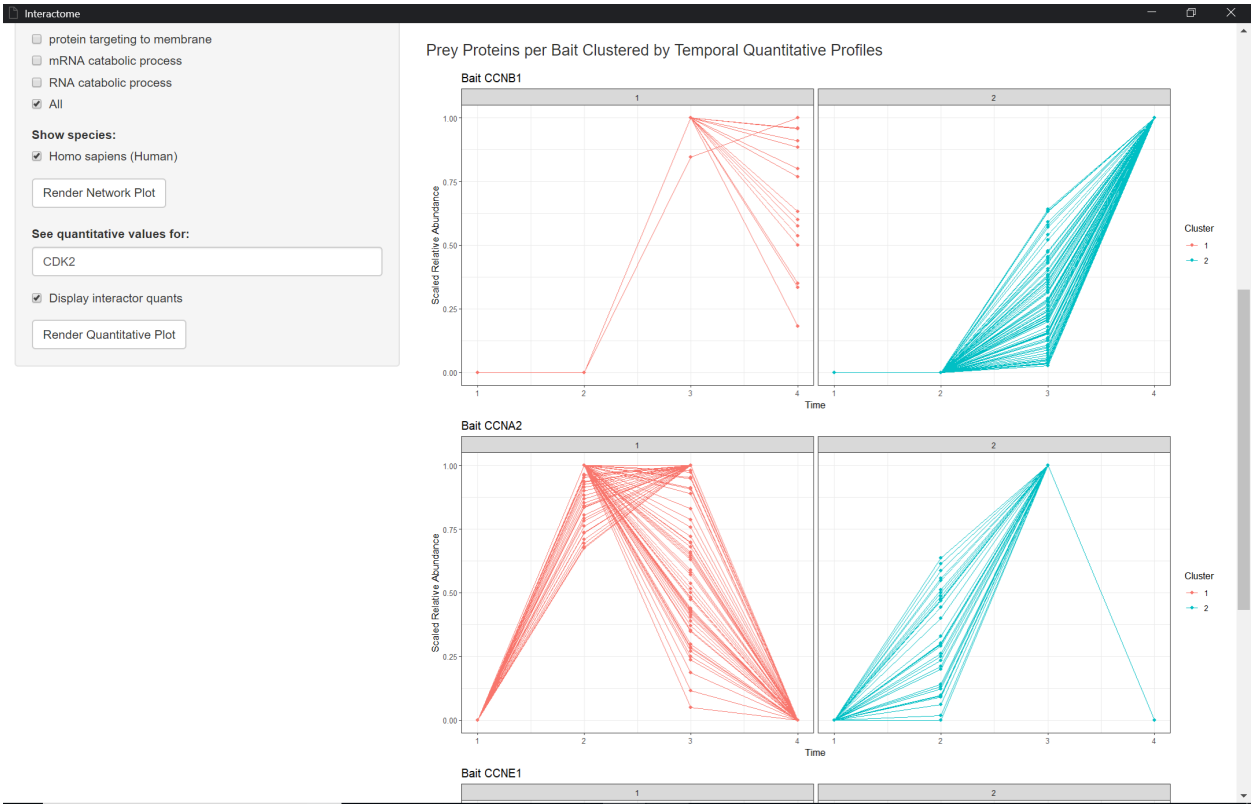

Figure 3: Cluster Plots

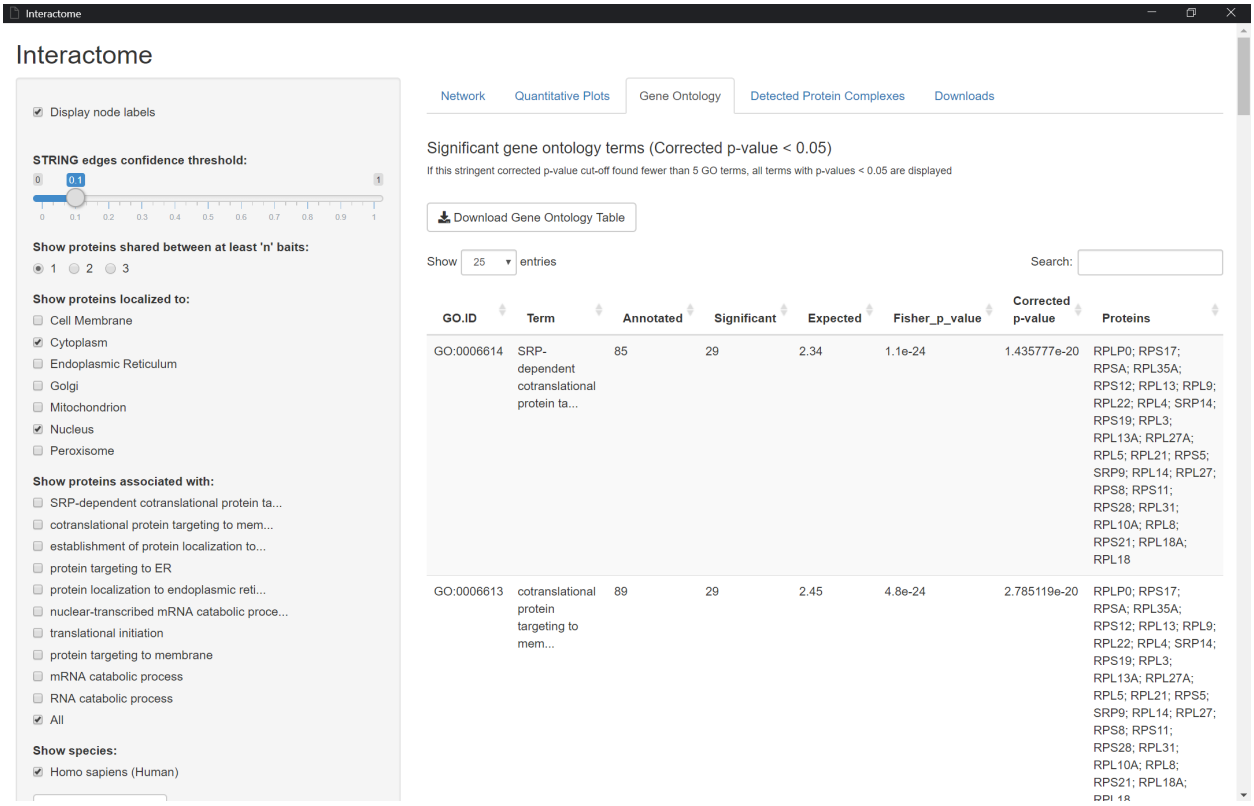

Figure 4: Gene Ontology Table

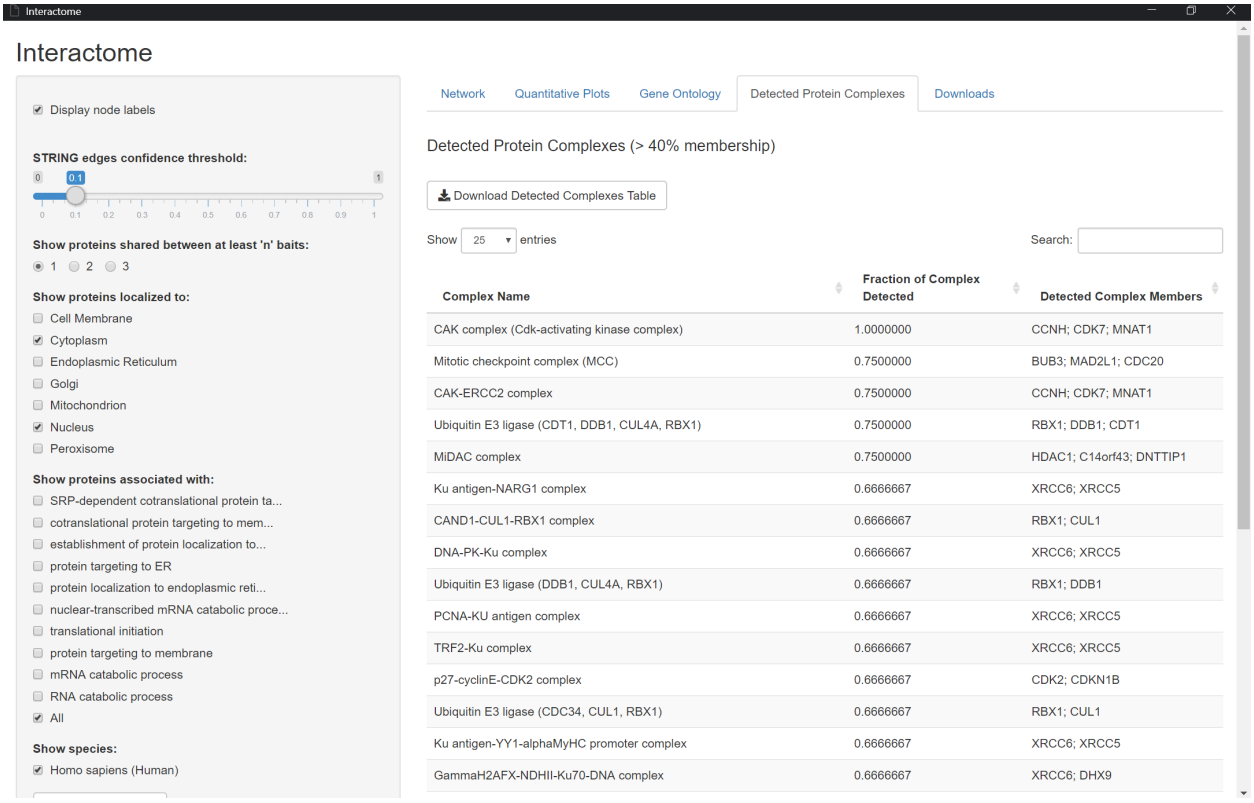

Figure 5: Detected Protein Complexes

Instructions

Set-up

Ensure you have Python v3.5+ and R v3.5+ installed on your device. You can download Python here and download R here.

Generate nodes and edges files for VISTA

VISTA requires two input files. The first, labeled "nodes.csv", must contain the following information: UniProt accession, gene name, taxonomic ID, and (optionally) localization information for each protein you wish to analyze. If providing localization information, please ensure that multiple localizations for a single protein are formatted as "Loc1; Loc2; Loc3" (example: "Cytoplasm; Mitochondria; Golgi").

The second file contains the list of interactions. It must be labeled "edges_x.csv", where 'x' is a number from 0-1 that specifies the threshold confidence level above which edges will be incorporated into the interactome. The interactions file must contain the following information: bait protein UniProt accession, prey protein UniProt accession, edge confidence at each condition, and (optionally) quantitative information for prey proteins (ex: spectral counts, MS1 abundance) at each condition.

The user can either provide these files in the specified format (see the "edges_x.csv" and "nodes.csv" files in the 'Example' folders), or use the provided program to generate files in the required format. Navigate to the 'Generate Files' folder and double-click **main**.py. Follow the instructions in the pop-up windows until the required files are produced in the 'Generate Files' folder.

Supply background gene list for gene ontology enrichment

The final piece of user-provided information is a background gene list to perform gene ontology (GO) enrichment. This is a list of UniProt accessions and taxonomic IDs for all proteins you wish to include as background. The provided 'Example' folders have sample "background_gene_list.txt" files. A set of reference gene lists are also provided in the "Background Gene Lists" folder, including proteins expressed in: human fibroblast cells, human HeLa cells, and differentiated PC12 rat neurons.

Optional: Normalize abundances to proteome abundance

Changes in interactions may be driven by either functional or proteomic abundance changes. TO account for the latter, VISTA can normalize the provided interaction abundances to proteome abundances and produce heatmaps of interaction abundances prior to and after normalization to the proteome. To perform this scaling, please provide a file labeled "proteome_abundance.txt" that contains UniProt accession numbers, gene names, and protein abundance at each timepoint or condition *in the same order* (ex: 24 hpi, 48 hpi, 72 hpi) as the interaction edges data. Note 1: VISTA will not automatically use the normalized interaction abundances across the tool. If you wish to use the normalized abundances for further analysis, download these values from the nodes and edges files output by VISTA and replace the provided abundances with the new normalized abundances before running VISTA again. Note 2: A proteome abundance file for HCMV infection from 24 - 96 hpi is provided in the "Proteome Abundance" folder. TO use the file, rename it to "proteom_abundance.txt".

Move files to be analyzed into 'Shiny' folder

Copy the generated "nodes.csv", "edges_x.csv", "background_gene_list.txt", and (optionally) "proteome_abundance.txt" into the 'Shiny' folder. Please ensure that the files are named exactly as specified. Each time you run the program with new data, remove any existing files and replace them with the new data.

Run the program

Double-click "run.vbs". A log file will be generated that you can open with any text editor to check the status of the program. Depending on the size of your input files, VISTA may take anywhere between 5 - 15 minutes to run, so please be patient! When finished, a browser window will open.

Output

VISTA outputs an interactive user interface that enables the user to view a dynamic interaction network, modify the network by toggling proteins and edges on and off based on various properties, and perform exploratory data analysis on quantitative information - including viewing abundance profiles, clustering proteins based on these profiles, etc.

A tutorial video with features of the output is packaged with the program as "VISTA Tutorial.mp4".

The first tab is titled **'Network'**. Use the sidebar on the left to select various properties of the network - STRING edge confidence threshold, shared protein attributes, localization (provided or UniProt), functional annotation, etc. Then click 'Render Network Plot' to build the interactome, which will appear in the main panel. Each node represents a protein, while each edge is an interaction. Nodes are colored by protein localization, with outline colors indicating the duration for which the protein is associated with a bait. Edges are colored by type - user-provided or STRING-inferred, and edge thickness indicates interaction confidence. Click the play and pause buttons to view dynamic changes to PPIs; click nodes or edges to view their properties (localization, duration of activity, confidence, etc.); double-click a node to highlight all its neighbors. To modify the network, choose different parameters from the sidebar and click 'Render Network Plot' again.

The second tab, titled **'Quantitative Plots'**, will appear if quantitative information was provided to VISTA. Use the sidebar on the left to plot quantitative profiles for individual proteins and their neighbors. The main panel also contains the quantitative plots for all proteins clustered by their quantitative profiles over multiple conditions. These are separated by bait, as quantitative profiles are measured for prey proteins per bait. Finally, if the user provides proteome abundance data, heatmaps comparing interaction abundances prior to and after normalization are displayed here. All the plots on this page can be downloaded as pdfs.

The third tab, titled **'Gene Ontology'**, outputs an interactive data table with significant GO terms found in the data. Significance is assessed as having a corrected p-value of < 0.05 after multiple-test correction is applied on Fisher test p-values for enrichment. Data in the table can be sorted, searched, and filtered. This table can be downloaded as a tab-separated file.

The fourth tab, titled **'Detected Protein Complexes'**, outputs an interactive data table with a list of mammalian protein complexes from CORUM with over 40% of members detected in the data. VISTA provides the complex name, fraction of complex detected, and lists the identified proteins in each complex. Data in the table can be sorted, searched, and filtered. This table can be downloaded as a tab-separated file.

The last tab is **'Downloads'** and provides links to download the annotated nodes and edges files produced by VISTA. These annotated files contain information on protein and interaction durations, localization, GO terms, cluster membership, etc. and can be directly exported to Cytoscape. Finally, an interactome "report" that provides exploratory analysis on the duration of activity for various proteins in the network, separated by their localizations and functional annotations, as well as plots of interaction onsets and durations can be downloaded as a pdf.

Examples

Sample files are provided in four 'Example' folders - 'Example pUL13', 'Example pUL37', 'Example Us9', and 'Example Cyclins'. Copy the files in one of these folders into the 'Shiny' folder to run the sample interactomes.