# Chapter 9

# Computing and Visualizing Gene Function Similarity and Coherence with NaviGO

Ziyun Ding, Qing Wei, and Daisuke Kihara

## Abstract

Gene ontology (GO) is a controlled vocabulary of gene functions across all species, which is widely used for functional analyses of individual genes and large-scale proteomic studies. NaviGO is a webserver for visualizing and quantifying the relationship and similarity of GO annotations. Here, we walk through functionality of the NaviGO webserver (http://kiharalab.org/web/navigo/) using an example input and explain what can be learned from analysis results. NaviGO has four main functions, accessed from each page of the webserver: "GO Parents," "GO Set", "GO Enrichment", and "Protein Set." For a given list of GO terms, the "GO Parents" tab visualizes the hierarchical relationship of GO terms, and the "GO Set" tab calculates six functional similarity and association scores and presents results in a network and a multidimensional scaling plot. For a set of proteins and their associated GO terms, the "GO Enrichment" tab calculates protein GO functional enrichment, while the "Protein Set" tab calculates functional association between proteins. The NaviGO source code can be also downloaded and used locally or integrated into other software pipelines.

**Key words** NaviGO, Gene ontology, Functional similarity, Visualization, Quantification, Function enrichment analysis, GO association score, Protein functional association score, Proteomic analysis

## 1 Introduction

The gene ontology (GO) is a widely used vocabulary for representing gene functions across all species [1, 2]. It is maintained and updated by the Gene Ontology Consortium. Currently, GO terms are classified into three categories: biological process (BP), which describes pathway information of gene products such as cellular physiological process or signal transduction; molecular function (MF), which describes molecular level activities such as enzymatic activity; and cellular component (CC), which describes cellular localization of gene products. Currently, there are over 46,000 GO terms, which are organized in a hierarchical structure, a directed acyclic graph (DAG). GO is very useful, particularly

Hiroshi Mamitsuka (ed.), *Data Mining for Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 1807, https://doi.org/10.1007/978-1-4939-8561-6_9, 

for computational analysis of gene functions; however, the volume of the vocabulary and the complicated relationships often makes analysis cumbersome.

NaviGO was developed to facilitate easy handling of GO terms, particularly for quantifying and visualizing relationships between GO terms [3]. NaviGO has four main functions: "GO Parents", "GO Set", "GO Enrichment", and "Protein Set." "GO Parents" maps and visualizes the hierarchical relationship of GO terms in an interactive fashion, and "GO Set" calculates six functional similarity and association scores and provides two visualization tools, a network and a multidimensional scaling visualization. For a list of proteins and associated GO terms, "GO Enrichment" performs GO enrichment analysis, while "Protein Set" identifies functionally related proteins. Compared with other related online tools [4, 5], NaviGO server has several advantages: first, it provides multiple similarity scores, which not only compare GO terms in the same GO category but also across GO categories. NaviGO provides biologists an intuitive and interactive tool to visualize parental relationships between GO terms. NaviGO is also integrated into the popular gene function prediction webservers PFP [6, 7] and ESG [8, 9].

## 2 Materials

NaviGO can be freely accessed at http://kiharalab.org/web/navigo/. It is a web application and does not require any platform other than a web browser. The source codes of NaviGO and GO Visualizer, a tool for visualizing the hierarchy of GO terms, can be downloaded from GitHub at https://github.com/kiharalab/NaviGO and https://github.com/kiharalab/GOVisualizer under the terms of the GNU Lesser General Public License Ver. 2.1.

In order to use NaviGO, users need to provide a set of GO terms or a set of UniProt IDs of proteins and associated GO terms to be analyzed. These can be retrieved from the Gene Ontology Consortium website (http://www.geneontology.org/) [1] or from the UniProt database (http://www.uniprot.org/) [10], respectively.

## 3 Methods

### 3.1 Overview of NaviGO

The NaviGO server has four main functions (Fig. 1). Either a set of GO terms or a set of proteins (with their GO terms) can be analyzed. For a set of GO terms, the "GO Parents" tab visualizes input GO terms in the GO DAG, and the "GO Set" tab calculates the functional similarity and association scores and visualizes them. On the other hand, for a list of proteins with their GO terms, the

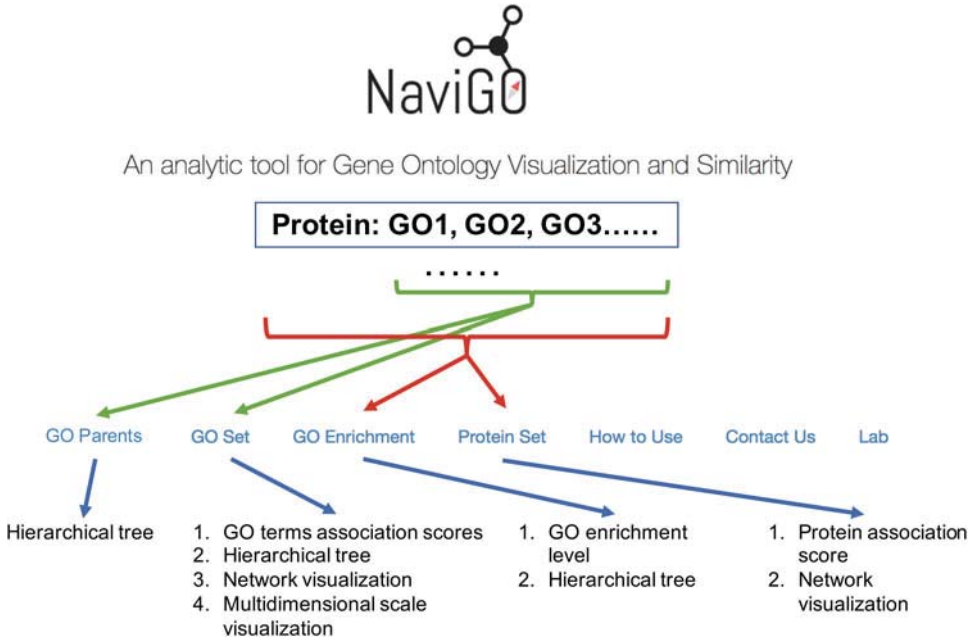



**Fig. 1** Overview of NaviGO functionality. Input to be analyzed can be either a set of GO terms or a set of proteins

"GO Enrichment" tab performs GO enrichment analysis, and the "Protein Set" tab calculates functional similarity and association scores between proteins (Fig. 1).

Throughout this tutorial, we use the following six proteins, which are involved in the light signaling pathway, as examples: phytochrome A (PHYA, UniProt ID: P14712), phytochrome B (PHYB, UniProt ID: P14713), phytochrome D (PHYD, UniProt ID: P42497), phytochrome-interacting factor 3 (PIF3, UniProt ID: O80536), pseudo-response regulator 7 (PRR7, UniProt ID: A0A1P8BCB0), and histone deacetylase 15 (HDA15, UniProt ID: Q8GXJ1) (Table 1). PHYA, PHYB, and PHYD are from the phytochrome family and mainly function as red and far-red photoreceptors. They have been experimentally verified to interact with each other [11]. Interaction of the transcription factor PIF3 with the phytochrome family causes phosphorylation and degradation of phytochrome [12, 13]. Interaction of PIF3 with HDA15, a transcriptional repressor, represses the chlorophyll biosynthesis and the photosynthesis [14]. PRR7 is one of the key components of molecular clock in *Arabidopsis* and involved in the phytochrome-mediated red light signal transduction pathway [15]. PRR7 interacts with phytochrome and PIF to regulate the red light signal transduction. Known physical interactions of the six proteins are summarized in Fig. 2.

**Table 1**
**List of the six example proteins and their associated GO terms**

| Protein name | UniProt ID | CC | MF | BP |
|---|---|---|---|---|
| PHYA | P14712 | GO:0005737, GO:0016604, GO:0016607, GO:0005634 | GO:0031516, GO:0042802, GO:0003729, GO:0000155, GO:0042803, GO:0004672, GO:0009883 | GO:0009584, GO:0009630, GO:0017148, GO:0009640, GO:0009638, GO:0018298, GO:0017006, GO:0010161, GO:0006355, GO:0046685, GO:0010201, GO:0010218, GO:0010203, GO:0006351 |
| PHYB | P14713 | GO:0005829, GO:0016604, GO:0016607, GO:0005634 | GO:0031516, GO:0042802, GO:0000155, GO:1990841, GO:0042803, GO:0031517, GO:0009883, GO:0043565 | GO:0009687, GO:0006325, GO:0010617, GO:0009584, GO:0009649, GO:0009630, GO:0009867, GO:0045892, GO:0009640, GO:0015979, GO:0009638, GO:0018298, GO:0017012, GO:0010161, GO:0031347, GO:2000028, GO:0010029, GO:0009409, GO:0010218, GO:0010244, GO:0010202, GO:0009266, GO:0010374, GO:0006351, GO:0010148 |
| PRR7 | A0A1P8BCB0 | GO:0005634 | NA | GO:0000160 |
| PIF3 | O80536 | GO:0005634 | GO:0003677, GO:0042802, GO:0046983, GO:0003700 | GO:0009704, GO:0009740, GO:0031539, GO:0010017, GO:0009585, GO:0006355, GO:0009639, GO:0006351 |
| PHYD | P42497 | GO:0005634 | GO:0042802, GO:0000155, GO:0009881, GO:0042803 | GO:0018298, GO:0017006, GO:0009585, GO:0006355, GO:0006351 |
| HDA15 | Q8GXJ1 | GO:0005634 | GO:0046872, GO:0032041 | GO:0006355, GO:0006351 |

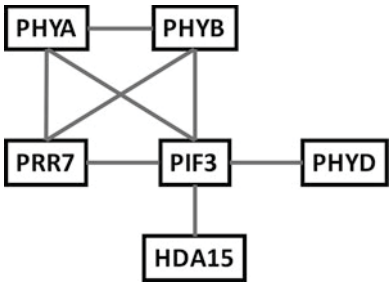



**Fig. 2** The interaction relationship of the six example proteins. These six proteins are involved in the light signaling pathway

### *3.2 Quantification and Visualization of GO Term Association and Similarity*

The "GO Set" tab computes six GO term similarity and association scores for all the pairs of input GO terms. The scores are Resnik's, Lin's, relevance similarity score (RSS), the Interaction Association Score (IAS), the PubMed Association Score (PAS), and the Co-occurrence Association Score (CAS). The first three scores quantify similarity of a GO term pair of the same category. They are calculated based on the frequencies of two GO terms in the gene annotation database and their location in the GO DAG [16–18]. Among the three scores, RSS not only considers the relative depth of the common ancestor between the two GO terms but also considers how rare the query GO terms are to identify the common ancestor. The last three semantic-based functional similarity scores were developed by our group. IAS quantifies the probability that two GO terms appear in physically interacting protein pairs [19]. PAS and CAS quantify the frequency with which two GO terms appear in the same PubMed abstract and in a single gene annotation, respectively [20].

To use the "GO Set" tab, please follow the steps described below:

1. Enter your input in the box. The input format of the "GO set" tab is a list of GO terms separated by comma. Users can upload a formatted file or type in the GO term ID. As a GO term ID is being typed, NaviGO will automatically recognize the GO term with the number and show candidates in a pull-down list. Thus, users can choose one from the list. For example, "GO:0005737" can be retrieved after typing "5737" and clicking the first GO term in the pull-down list (Fig. 3).
2. To empty inputs, click the "Reset button" located above the input box. To delete a single GO term in the input box, click the "X" sign at the GO term.
3. Clicking the "Submit" button below the input box will start the analysis and show a result page when done.

At the top of the results page, query GO term scores are listed in colors that indicate categories: BP terms are in red, MF in blue,

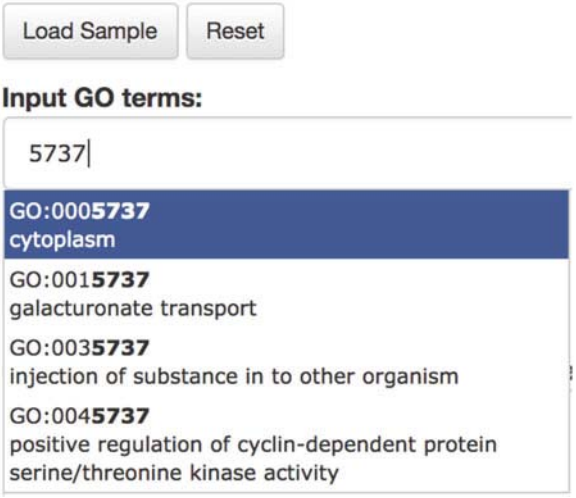
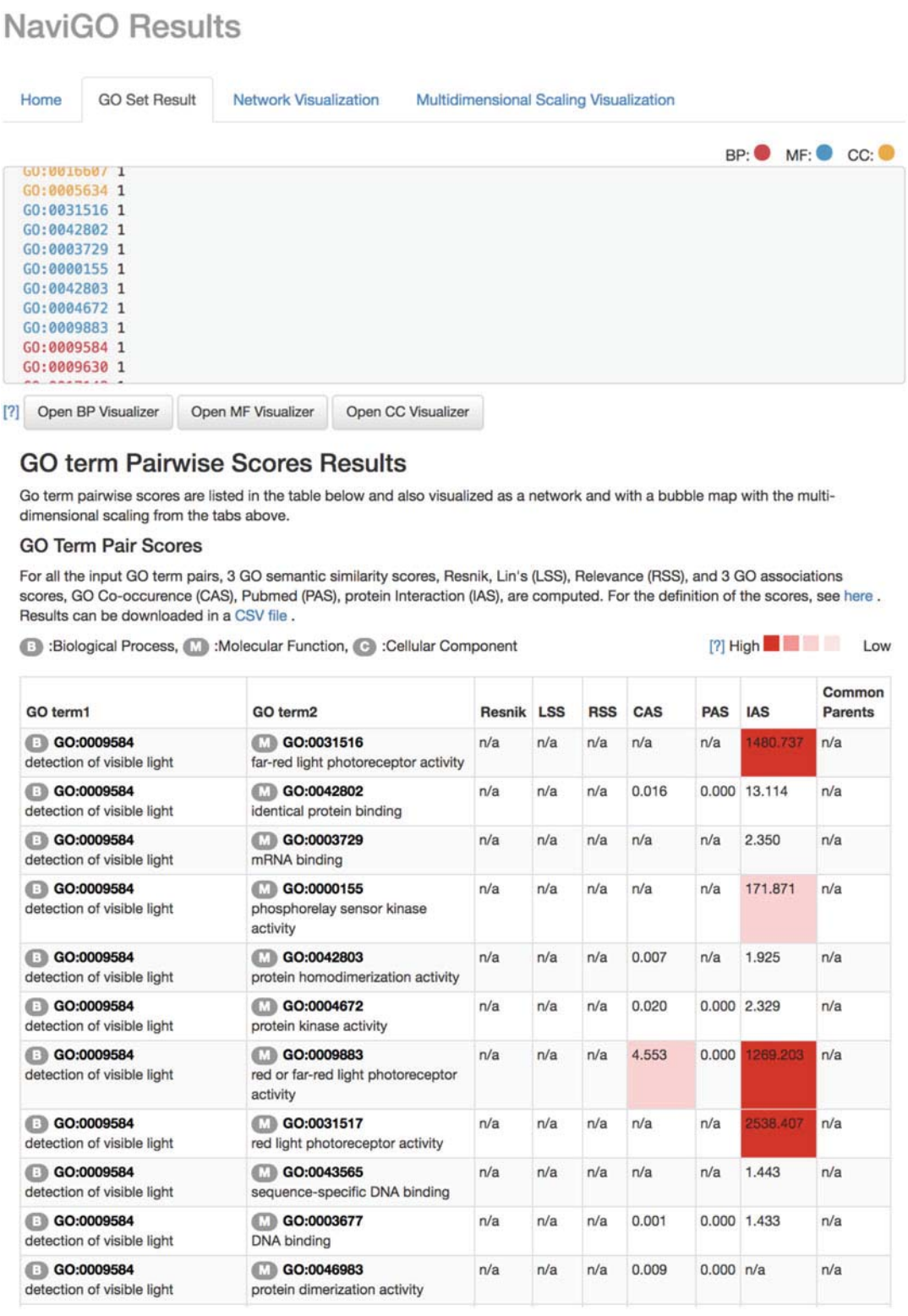



**Fig. 3** Example of inputting GO terms

and CC in yellow (Fig. 4, top). The numbers on the right side of GO terms are the counts of each GO term in the input. Clicking the BP/MF/CC Visualizer button below the query GO term list will open a new page that shows the GO terms of the category in the GO DAG (Fig. 5). The color legends of GO terms are listed on the right side, and colors of GO term relationships are shown in the left upper corner of the page. The query GO terms are shown in a larger font in the DAG. Clicking a GO term in a graph will expand links to all the children GO terms. In the example shown in Fig. 5, seven molecular function GO terms are mapped (Fig. 5). We can see that the GO terms locate in two branches, photoreceptor activity and protein-binding activity.

Pairwise GO term scores are calculated and listed in the table below the input GO term list (Fig. 4, bottom). GO pairs in the table can be sorted by a score by clicking the title of the score column. If the members of a pair of GO terms do not belong to the same category or the score of the pair is not available, "n/a" is shown. The significance level of scores in each column is indicated in a color scale, from light pink to red as the significance level increases. Clicking the "+" in the "common parents" column expands the list of all common parents of the GO pair in the GO DAG. Clicking a GO term will take users to the AmiGO website, which provides more detailed information of the term.

In Fig. 4, a part of the result page for the example input in Table 1 is shown. IAS of the BP term GO:0009584 (related to detection of visible light) and the MF term GO:00031516 (far-red light photoreceptor activity) is highlighted in red, because the score 1480.737 is within the top 1% of scores relative to the score background distribution. Both GO terms annotate the




| GO term1 | GO term2 | Resnik | LSS | RSS | CAS | PAS | IAS | Common Parents |
|---|---|---|---|---|---|---|---|---|
| B GO:0009584 detection of visible light | M GO:0031516 far-red light photoreceptor activity | n/a | n/a | n/a | n/a | n/a | 1480.737 | n/a |
| B GO:0009584 detection of visible light | M GO:0042802 identical protein binding | n/a | n/a | n/a | 0.016 | 0.000 | 13.114 | n/a |
| B GO:0009584 detection of visible light | M GO:0003729 mRNA binding | n/a | n/a | n/a | n/a | n/a | 2.350 | n/a |
| B GO:0009584 detection of visible light | M GO:0000155 phosphorelay sensor kinase activity | n/a | n/a | n/a | n/a | n/a | 171.871 | n/a |
| B GO:0009584 detection of visible light | M GO:0042803 protein homodimerization activity | n/a | n/a | n/a | 0.007 | n/a | 1.925 | n/a |
| B GO:0009584 detection of visible light | M GO:0004672 protein kinase activity | n/a | n/a | n/a | 0.020 | 0.000 | 2.329 | n/a |
| B GO:0009584 detection of visible light | M GO:0009883 red or far-red light photoreceptor activity | n/a | n/a | n/a | 4.553 | 0.000 | 1269.203 | n/a |
| B GO:0009584 detection of visible light | M GO:0031517 red light photoreceptor activity | n/a | n/a | n/a | n/a | n/a | 2538.407 | n/a |
| B GO:0009584 detection of visible light | M GO:0043565 sequence-specific DNA binding | n/a | n/a | n/a | n/a | n/a | 1.443 | n/a |
| B GO:0009584 detection of visible light | M GO:0003677 DNA binding | n/a | n/a | n/a | 0.001 | 0.000 | 1.433 | n/a |
| B GO:0009584 detection of visible light | M GO:0046983 protein dimerization activity | n/a | n/a | n/a | 0.009 | 0.000 | n/a | n/a |

**Fig. 4** GO Set result page

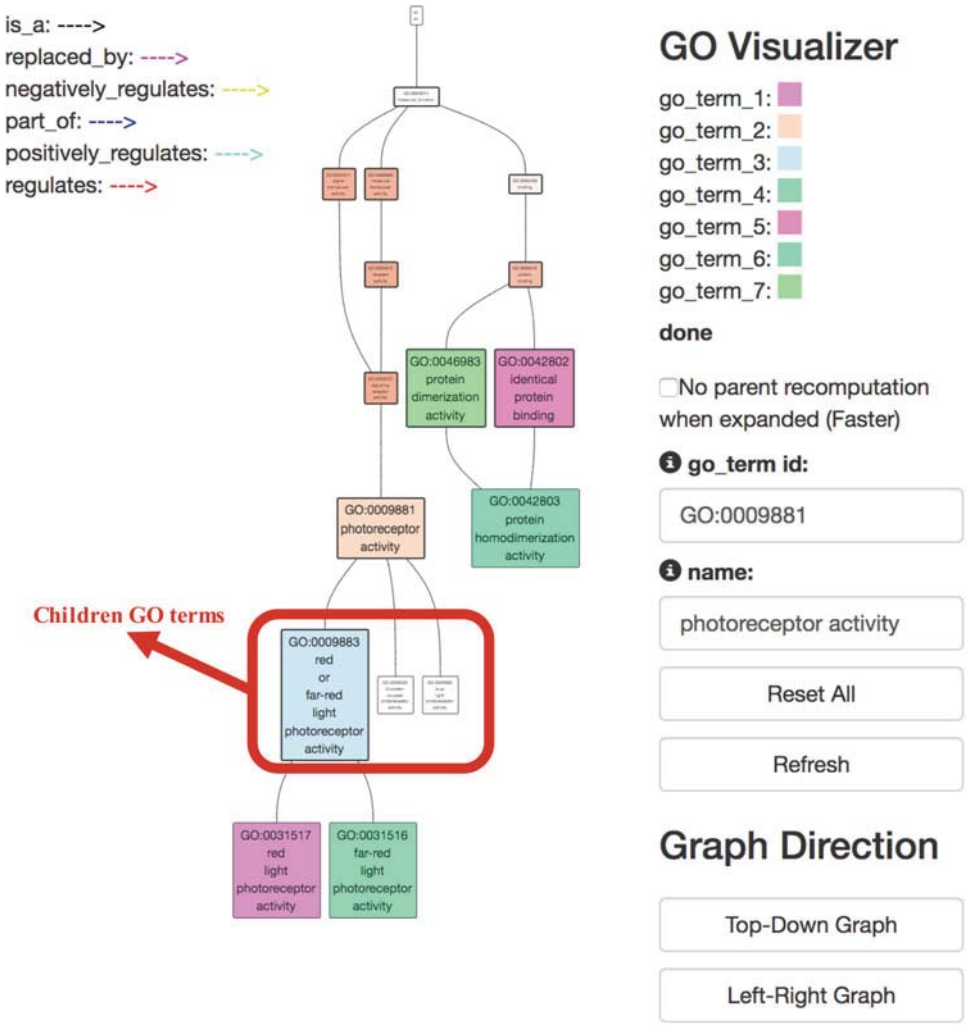



**Fig. 5** Hierarchical graph representation using GO Visualizer. Seven molecular function GO terms are visualized here, GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803, and GO:0046983. Clicking a GO term expands edges to all the children GO terms

phytochrome proteins, which are known to form heterodimers [11], so it is reasonable that the two GO terms annotating these interacting proteins have a very high IAS.

The results table can be also downloaded in a comma separated data file (a CSV format file) by clicking the "CSV file" button.

NaviGO provides two types of visualizations for GO pairwise score results. One is a network visualization available under the "Network Visualization" tab (Fig. 6). In the network, functionally related GO terms are connected by edges. The score cutoff value to define edges can be controlled by sliding the bar or by typing the value in a text box. The scores to visualize can be chosen at the upper left corner of the page. In the example in Fig. 6, the

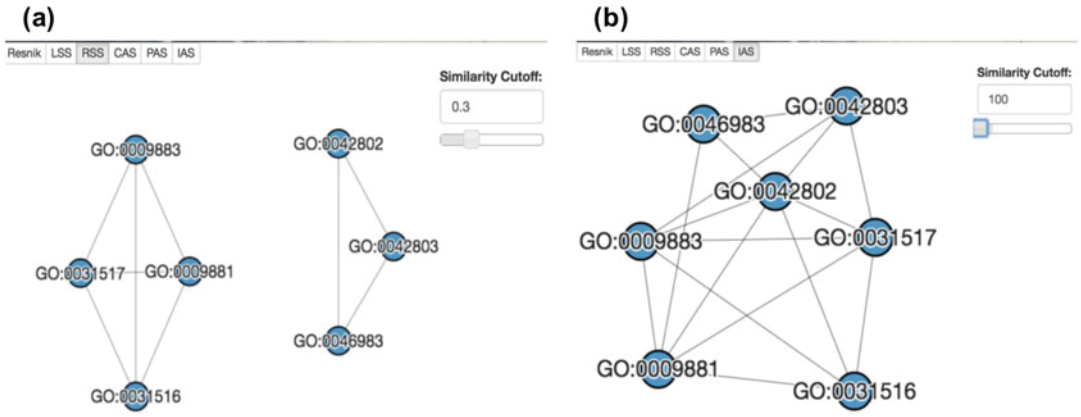



**Fig. 6** The network of functional association score of seven GO terms: GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803, and GO:0046983. (**a**) Network using RSS with a cutoff value of 0.3. (**b**) Network using IAS with a cutoff value of 100

same set of GO terms as Fig. 5 was used. In the network with RSS (Fig. 6, left), the GO terms were clustered into two groups, which is consistent with the two branches in the GO hierarchy shown in Fig. 5. Using IAS, all GO terms are connected (Fig. 6b). This is also reasonable because these terms are associated with light signaling proteins, and they are known to interact with each other as shown in Fig. 2.

The second visualization is available at the "Multidimensional Scaling Visualization" tab. In this two-dimensional (2D) graph, GO terms are classified and mapped onto a 2D space with two scores selected by users. Placing the cursor over a GO term will show the normalized functional score of the GO term. In the example in Fig. 7, the x-axis is RSS and the y-axis is PAS. The GO terms are largely classified in two groups, which are again consistent with the results in Figs. 5 and 6.

### *3.3 GO Enrichment Analysis*

The goal of GO enrichment analysis is to find if any GO term appears more frequently in a set of proteins than would be expected from the background frequency of the term in the genome. The significance of protein is quantified by a *p-value*. A *p-value* of a GO term for a protein set is calculated by considering the number of proteins in the set, the number of proteins annotated with the GO term in the genome, and the total number of proteins in the organism. The smaller the *p-value* is, the more significant the GO term is.

The input format of the "GO Enrichment" tab is a list of UniProt IDs associated with GO terms, e.g., "A0A1P8BCB0 GO:0005634, GO:0000160". NaviGO automatically identifies the organism based on the UniProt ID of the first protein in the input.

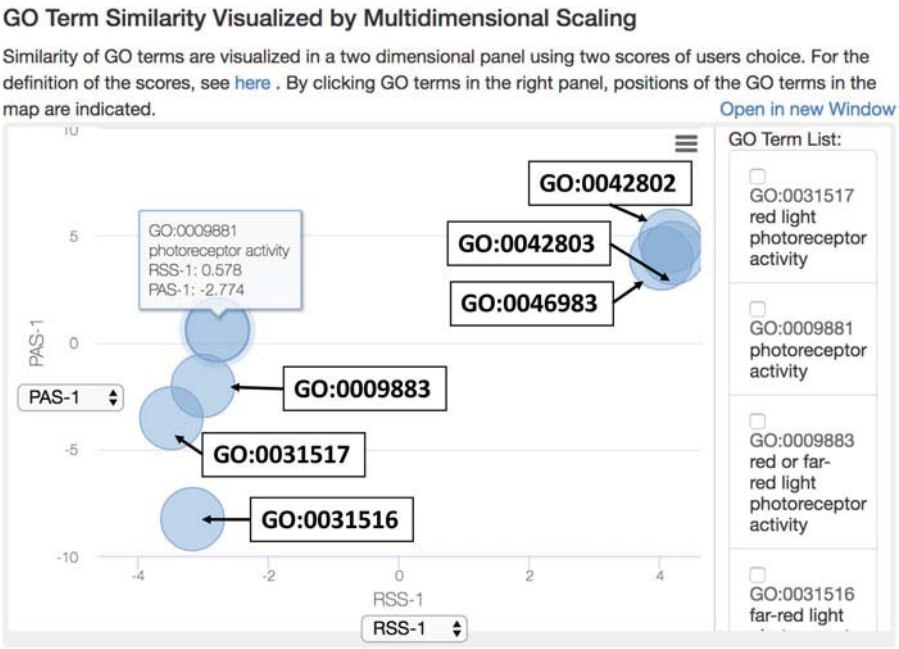
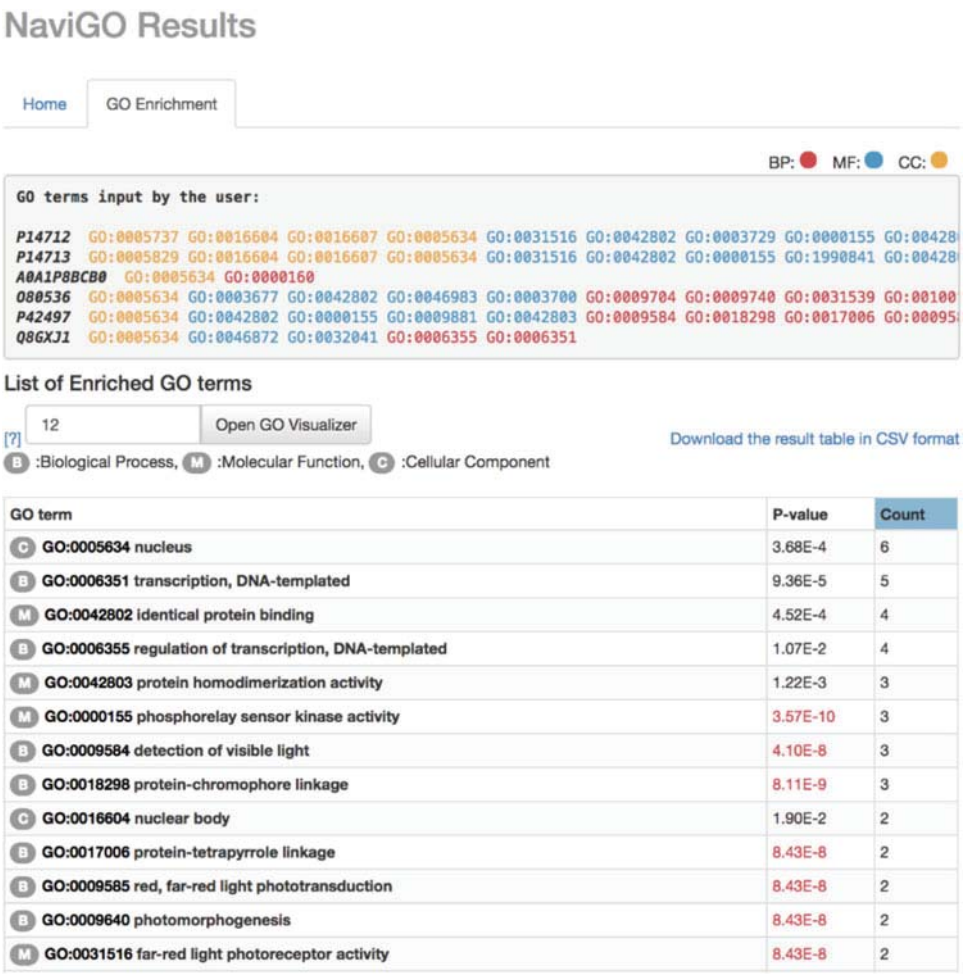



**Fig. 7** The multidimensional scaling visualization of seven GO terms: GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803, and GO:0046983

In the results page, GO terms of input proteins are sorted by their p-value (Fig. 8). The significant p-value (below 0.00005 or top 30) GO terms are highlighted in red. The total number of significantly enriched GO terms is counted in the box on the left from "Open GO Visualizer". The third column shows the number of input proteins that have the GO term. In the example shown in Fig. 8, the most enriched GO term among the proteins from *Arabidopsis* is GO:0000155 *phosphorelay sensor kinase activity* with a p-value of 3.57E-10 and GO:0018298 *protein-chromophore linkage* with a p-value of 8.11E-9.

Enriched GO terms with p-value above 0.00005 (or the top 30 GO terms) can be mapped to the GO hierarchy by clicking "open GO Visualizer". The enriched GO terms are shown in a larger font and colored based on p-value from red to yellow indicating most to least significance. The figure can be downloaded by clicking "Download Figure Here". In the example in Fig. 9, the significantly enriched GO terms are involve in the signal receptor activity such as GO: 0000155, "phosphorelay sensor kinase activity" with p-value of 3.57e-10, and GO:0009883, "red or far-red light photoreceptor" with *p-value* of 8.43e-8. Also, GO terms identified as enriched are involve in red or far-red light signaling pathway such as GO:0010161, "red light signaling pathway" with *p-value* of 8.43e-8, and detection of light stimulus such as GO:0009854, "detection of visible light" with *p-value* of 4.10e-8.




| GO term | P-value | Count |
|---|---|---|
| C GO:0005634 nucleus | 3.68E-4 | 6 |
| B GO:0006351 transcription, DNA-templated | 9.36E-5 | 5 |
| M GO:0042802 identical protein binding | 4.52E-4 | 4 |
| B GO:0006355 regulation of transcription, DNA-templated | 1.07E-2 | 4 |
| M GO:0042803 protein homodimerization activity | 1.22E-3 | 3 |
| M GO:0000155 phosphorelay sensor kinase activity | 3.57E-10 | 3 |
| B GO:0009584 detection of visible light | 4.10E-8 | 3 |
| B GO:0018298 protein-chromophore linkage | 8.11E-9 | 3 |
| C GO:0016604 nuclear body | 1.90E-2 | 2 |
| B GO:0017006 protein-tetrapyrrole linkage | 8.43E-8 | 2 |
| B GO:0009585 red, far-red light phototransduction | 8.43E-8 | 2 |
| B GO:0009640 photomorphogenesis | 8.43E-8 | 2 |
| M GO:0031516 far-red light photoreceptor activity | 8.43E-8 | 2 |

**Fig. 8** The GO enrichment analysis result of six proteins: PHYA (P14712), PHYB (P14713), PRR7 (A0A1P8BCB0), PIF3 (O80536), PHYD (P42497), and HDA15 (Q8GXJ1)

### *3.4 Quantifying Functional Association of Proteins*

This function identifies protein pairs in a query protein set that have functional relevance. Using a GO pair score, functional relevance of a protein pair is evaluated by the funSim score [21], which is in essence the average GO pair scores of GO annotations of the two proteins (*see* **Note 1**). Eight different GO pair scores are used in NaviGO (Fig. 10): "MF", "BP", and "CC" use RSS of the particular GO category, "BP + MF" is the funSim score using BP and MP, while "All" is the funSim using MF, BP, and CC. "PAS", "CAS", and "IAS" use the corresponding functional association scores to compute funSim.

For example, when studying whether proteins exist in the same cellular component, it would be interesting to check the CC funSim score. When studying whether proteins are involved in the same pathway or biological process, users would want to check "BP", "MF", or "BP + MF" columns.

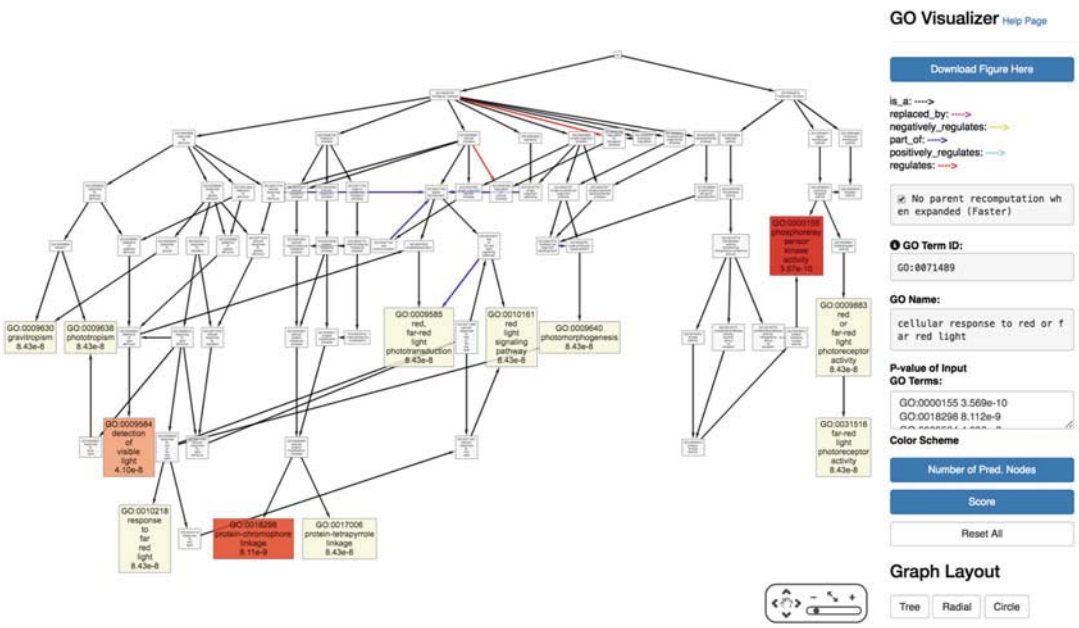
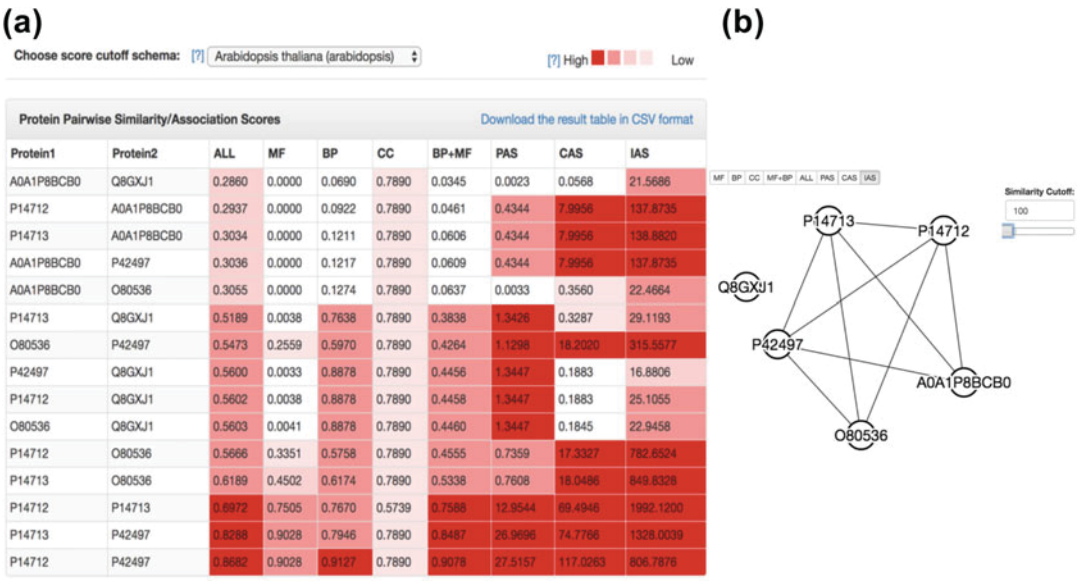



**Fig. 9** Visualization of 12 significantly enriched GO terms




| Protein1 | Protein2 | ALL | MF | BP | CC | BP+MF | PAS | CAS | IAS |
|---|---|---|---|---|---|---|---|---|---|
| A0A1P8BCB0 | Q8GXJ1 | 0.2860 | 0.0000 | 0.0690 | 0.7890 | 0.0345 | 0.0023 | 0.0568 | 21.5686 |
| P14712 | A0A1P8BCB0 | 0.2937 | 0.0000 | 0.0922 | 0.7890 | 0.0461 | 0.4344 | 7.9956 | 137.8735 |
| P14713 | A0A1P8BCB0 | 0.3034 | 0.0000 | 0.1211 | 0.7890 | 0.0606 | 0.4344 | 7.9956 | 138.8820 |
| A0A1P8BCB0 | P42497 | 0.3036 | 0.0000 | 0.1217 | 0.7890 | 0.0609 | 0.4344 | 7.9956 | 137.8735 |
| A0A1P8BCB0 | O80536 | 0.3055 | 0.0000 | 0.1274 | 0.7890 | 0.0637 | 0.0033 | 0.3560 | 22.4664 |
| P14713 | Q8GXJ1 | 0.5189 | 0.0038 | 0.7638 | 0.7890 | 0.3838 | 1.3426 | 0.3287 | 29.1193 |
| O80536 | P42497 | 0.5473 | 0.2559 | 0.5970 | 0.7890 | 0.4264 | 1.1298 | 18.2020 | 315.5577 |
| P42497 | Q8GXJ1 | 0.5600 | 0.0033 | 0.8878 | 0.7890 | 0.4456 | 1.3447 | 0.1883 | 16.8806 |
| P14712 | Q8GXJ1 | 0.5602 | 0.0038 | 0.8878 | 0.7890 | 0.4458 | 1.3447 | 0.1883 | 25.1055 |
| O80536 | Q8GXJ1 | 0.5603 | 0.0041 | 0.8878 | 0.7890 | 0.4460 | 1.3447 | 0.1845 | 22.9458 |
| P14712 | O80536 | 0.5666 | 0.3351 | 0.5758 | 0.7890 | 0.4555 | 0.7359 | 17.3327 | 782.6524 |
| P14713 | O80536 | 0.6189 | 0.4502 | 0.6174 | 0.7890 | 0.5338 | 0.7608 | 18.0486 | 849.8328 |
| P14712 | P14713 | 0.6972 | 0.7505 | 0.7670 | 0.5739 | 0.7588 | 12.9544 | 69.4946 | 1992.1200 |
| P14713 | P42497 | 0.8288 | 0.9028 | 0.7946 | 0.7890 | 0.8487 | 26.9696 | 74.7766 | 1328.0039 |
| P14712 | P42497 | 0.8682 | 0.9028 | 0.9127 | 0.7890 | 0.9078 | 27.5157 | 117.0263 | 806.7876 |

**Fig. 10** Example of protein set analysis. (**a**) Results of pairwise protein association scores. (**b**) The protein association network of six proteins PHYA (P14712), PHYB (P14713), PRR7 (A0A1P8BCB0), PIF3 (O80536), PHYD (P42497), and HDA15 (Q8GXJ1) with a cutoff value of 100

The input data is a list of proteins and their GO annotations, the same as described in the GO enrichment analysis. In the results table (Fig. 10), the significance levels of scores are shown in color scale (red to pink for high to low). Since the significant cutoff is defined by the score distribution of a particular organism, there is a

pull-down menu above the table to select the reference organism. In this example of six proteins, they all have the same RSS score of CC (Fig. 10a), reflecting that all proteins are located in the nucleus. PHYA (P14712) and PHYB (P14713) have a significantly high IAS of 1992.12, because they physically interact with each other [22].

Sometimes, it is difficult to see the functional association between proteins by looking at the score numbers in the output table. NaviGO provides a network visualization, which is available at "Open in new Window" (Fig. 10b). In the network, proteins are connected if their association scores are above a cutoff value. In the example, only HDA15 is not connected in the association network with IAS cutoff value set to 100. This is consistent with the STRING database [23], where only HDA15 has low binding scores with all the other proteins in this network.

### 3.5 Downloading Source Codes

The entire source code of NaviGO is available for academic use on GitHub (https://github.com/kiharalab/NaviGO). The code for GO Visualizer is also separately available on GitHub (https://github.com/kiharalab/GOVisualizer). GO Visualizer is the tool integrated in NaviGO that performs real-time rendering of GO DAGs in an interactive way. The code is licensed under the terms of the GNU Lesser General Public License Ver. 2.1. Users can download the package for local use of the software or to integrate it into other software pipelines, for example, a pipeline for proteomic mass spectrometry data with protein function analysis.

To set up the GO Visualizer locally, follow the steps below:

1. Install the software dependencies. The GO Visualizer requires Ruby, Gem, Sinatra, and MySQL. Ruby and Gem are two programming languages which have been installed in most of computers. Sinatra is a free and open-source software web application library. MySQL is one of the most popular open-source relational database management system. The general MySQL installer is available at https://dev.mysql.com/downloads.
2. Download the GO database from the GO Consortium. The size of the GO database is around 12 MB. In a Linux terminal, the GO database can be downloaded by running the following:

```
$ wget http://archive.geneontology.org/latest-full/
        go_monthly-termdb-data.gz
```

3. After downloading a GO database file, the Linux command to uncompress the file is:

```
$ gunzip go_monthly-termdb-data.gz
```

4. The GO database can be created with the following Linux command, where the "database_name" is the name users want to assign for the database, and "db_login" and "db_password" are the username and password defined by users, respectively:

```
$ echo ‘‘create database  database_name’’ |
        mysql --user=db_login --password=db_password
$ mysql --user=db_login --password=db_password
        database_name < go_monthly-termdb-data
```

5. To run GO Visualizer, use the following terminal commands:

```
$ git https://github.com/kiharalab/GOVisualizer.git
$ cd GOVisualizer
$ mv config_template.rb config.rb
```

6. Users need to change the settings in config.rb according to the local MySQL settings using following terminal command:

```
$ ruby server.rb
```

To set up the NaviGO webserver locally, follow the steps below:

1. In order to implement the NaviGO package on the local machine, Perl, Python 2.7, Python 3.4, and MySQL are needed.
2. The pre-calculated pairwise GO IAS, PAS, and CAS scores are also needed to compute protein functional association scores and can be downloaded with the following commands:

```
$ wget http://kiharalab.org/web/navigo/data/PAS.txt
$ wget http://kiharalab.org/web/navigo/data/CAS.txt
$ wget http://kiharalab.org/web/navigo/data/
        BIOGRID-3.2.107_GOpair_IASscores.txt
```

3. To install GO Visualizer, run:

```
$ git https://github.com/kiharalab/NaviGO.git
```

4. Users need to change the directory to “NaviGO/AutoUpdate”:

```
$ cd NaviGO/AutoUpdate
```

5. If there is no temporary folder, users need to make one named “tmp” to save the updated GO database:

```
$ mkdir tmp
```

6. Users need to run the “update.pl” script to generate the database NaviGO uses with the following command. After running the script, there should be a folder named like GO_YYYYMM, for example, GO_201701, and all the files should be inside this folder.

```
$ perl update.pl
```

7. Before running NaviGO, users need to set the paths correctly in “config_template.pl” and then rename the script “config_template.pl” to “config.pl”:

```
$ mv config_template.pl config.pl
```

8. To run the protein set function on NaviGO, users need to create their own folder under the "job" folder with the following commands, where "your_job_name" is the customized job name defined by user:

```
$ cd job
$ mkdir your_job_name
$ cd your_job_name
```

9. Users need to create the input file. The required format of input file can be found on our server NaviGO. For a given list of GO terms, the format should be "GO:xxxxxxx, GO:xxxxxxx . . . . . ." For a given list of UniProt ID and their associated GO terms, the format should be "UniProt_ID: GO:xxxxxxx, GO:xxxxxxx . . . . . ." Currently, NaviGO supports the required format and also the CAFA format (https://github.com/idoerg/cafa-format-check). NaviGO also provides a file checker under the "format_check" folder. The "cafa_go_format_checker.py" script will convert the CAFA format to our required format or check the file you provided. Users can run by the following command, where "file" is the CAFA-formatted file provided by the user, and "input_file" is the required format file converted by the script:

```
$ python ../../format_check/cafa_go_format_checker.py file >
  input_file
```

10. Then, you can run NaviGO by the following command:

```
$ perl ../../run.pl input_file
```

11. To run GO set, users can type the following commands, where "path_to_your_go_file" is the input file provided by users:

```
$ cd job
$ mkdir your_job_name
$ cd your_job_name
$ cp path_to_your_go_file ./input_file
$ perl execute.pl
```

12. To run enrichment analysis, users can type the following commands, where "organism_id" is the organism ID defined by UniProt database:

```
$ cd job
$ mkdir your_job_name
$ cd your_job_name
$ python3.4 ../../Enrich/enrich.py -f input_file -o
        organism_id
```

## 4 Notes

1. Functional relevance of a protein pair is quantified with the *funSim* score of a particular GO pair score [18]. The *funSim* score of a GO pair score for a protein pair is defined as, in

short, the average of the best combination of GO term pairs that annotate the two proteins. For a pair of protein, as shown in Fig. 10, NaviGO provides *funSim* scores for MF, BP, CC, BP + MF, all (i.e., BP + MF + CC), PAS, CAS, and IAS. For the first five scores, RSS of GO term pairs is used. From mathematical standpoint, the *funSim* score of a protein pair is defined as follows [18]:

First, RSS of two GO terms $c1$ and $c2$ is computed as

$$\mathrm{sim}\,(c1, c2) = \max_{c \in \mathrm{Ancestor}(c1,c2)} \left( \frac{2\log\,(p(c))}{\log p(c1) + \log p(c2)} \cdot (1 - p(c)) \right) \tag{1}$$

where $c$ represents a set of their common ancestors and $p(c)$ is defined as the fraction of proteins in the GOA database annotated with GO term $c$.

Then, the *funSim* score of a GO category, $\mathrm{GOscore}_{\mathrm{GOcategory}}\,(X, Y)$, is computed by averaging the *sim* scores between GO annotations of two proteins, $X$ and $Y$, in the given category as

$$\begin{aligned} &\mathrm{GOscore}_{\mathrm{GOcategory}}\,(X, Y) \\ &= \max \left\{ \left( \frac{1}{A_x} \sum_{i=1}^{A_x} \max_{1 \le j \le A_y} sim\,(Pxi, Pyj) \right), \right. \\ &\qquad \left. \left( \frac{1}{A_y} \sum_{j=1}^{A_y} \max_{1 \le i \le A_x} sim\,(Pxi, Pyj) \right) \right\} \end{aligned} \tag{2}$$

where $A_x$ and $A_y$ are the number of annotations for proteins $X$ and $Y$, respectively, in that category, and *Pxi* is ith annotation for protein X, and *Pyj* is jth annotation for protein Y.

For computing the *funSim* score for three categories,

$$\begin{aligned} &funSim\,(X, Y) \\ &= \frac{1}{3} \left( (\mathrm{GOscore}_{BP}\,(X, Y))^2 + (\mathrm{GOscore}_{MF}\,(X, Y))^2 \right. \\ &\quad \left. + (\mathrm{GOscore}_{CC}\,(X, Y))^2 \right) \end{aligned} \tag{3}$$

This is used for "All" in a NaviGO result page and for "BP + MF" is the average is computed for BP and MF.

## Acknowledgments

We thank Charles Christoffer for proofreading the manuscript. This work was partly supported by the National Institute of General Medical Sciences of the NIH (R01GM123055) and the National Science Foundation (DBI1262189, IOS1127027, DMS1614777).

## References

1. Consortium GO (2013) Gene ontology annotations and resources. Nucleic Acids Res 41(D1):D530–D535
2. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29. https://doi.org/10.1038/75556
3. Wei Q, Khan IK, Ding Z, Yerneni S, Kihara D (2017) NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology. BMC Bioinformatics 18(1):177. https://doi.org/10.1186/s12859-017-1600-5
4. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288–289. https://doi.org/10.1093/bioinformatics/btn615
5. Binns D, Dimmer E, Huntley R, Barrell D, O'donovan C, Apweiler R (2009) QuickGO: a web-based tool for gene ontology searching. Bioinformatics 25(22):3045–3046
6. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15(6):1550–1556. https://doi.org/10.1110/ps.062153506
7. Hawkins T, Chitale M, Luban S, Kihara D (2009) PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74(3):566–582. https://doi.org/10.1002/prot.22172
8. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25(14):1739–1745. https://doi.org/10.1093/bioinformatics/btp309
9. Khan IK, Qing W, Kihara D (2015) PFP/ESG: automated protein function prediction servers enhanced with gene ontology visualization tool. Bioinformatics 31(2):271–272. https://doi.org/10.1093/bioinformatics/btu646
10. Pundir S, Martin MJ, O'Donovan C (2017) UniProt protein knowledgebase. Methods Mol Biol 1558:41–55. https://doi.org/10.1007/978-1-4939-6783-4_2
11. Dieterle M, Bauer D, Büche C, Krenz M, Schäfer E, Kretsch T (2005) A new type of mutation in phytochrome A causes enhanced light sensitivity and alters the degradation and subcellular partitioning of the photoreceptor. Plant J 41(1):146–161
12. Nito K, Wong CC, Yates JR, Chory J (2013) Tyrosine phosphorylation regulates the activity of phytochrome photoreceptors. Cell Rep 3(6):1970–1979
13. Al-Sady B, Ni W, Kircher S, Schäfer E, Quail PH (2006) Photoactivated phytochrome induces rapid PIF3 phosphorylation prior to proteasome-mediated degradation. Mol Cell 23(3):439–446
14. Liu X, Chen C-Y, Wang K-C, Luo M, Tai R, Yuan L, Zhao M, Yang S, Tian G, Cui Y (2013) PHYTOCHROME INTERACTING FACTOR3 associates with the histone deacetylase HDA15 in repression of chlorophyll biosynthesis and photosynthesis in etiolated Arabidopsis seedlings. Plant Cell 25(4):1258–1273
15. Ito S, Nakamichi N, Nakamura Y, Niwa Y, Kato T, Murakami M, Kita M, Mizoguchi T, Niinuma K, Yamashino T (2007) Genetic linkages between circadian clock-associated components and phytochrome-dependent red light signal transduction in `Arabidopsis thaliana`. Plant Cell Physiol 48(7):971–983
16. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007
17. Lin D (1998) An information-theoretic definition of similarity. In: ICML, vol 1998. Citeseer, pp 296–304

18. Schlicker A, Domingues F, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics 7:302. https://doi.org/10.1186/1471-2105-7-302
19. Yerneni S, Khan I, Wei Q, Kihara D (2015) IAS: interaction specific GO term associations for predicting protein-protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2015.2476809
20. Chitale M, Palakodety S, Kihara D (2011) Quantification of protein group coherence and pathway assignment using functional association. BMC Bioinformatics 12(1):373
21. Hawkins T, Chitale M, Kihara D (2010) Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP. Bmc Bioinformatics 11(1):265
22. Clack T, Shokry A, Moffet M, Liu P, Faul M, Sharrock RA (2009) Obligate heterodimerization of Arabidopsis phytochromes C and E and interaction with the PIF3 basic helix-loop-helix transcription factor. Plant Cell 21(3):786–799
23. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45(D1):D362–D368. https://doi.org/10.1093/nar/gkw937