BioGateway Cytoscape App Manual

1 Installation

Download the BioGateway plugin .jar file, and save it to disk.

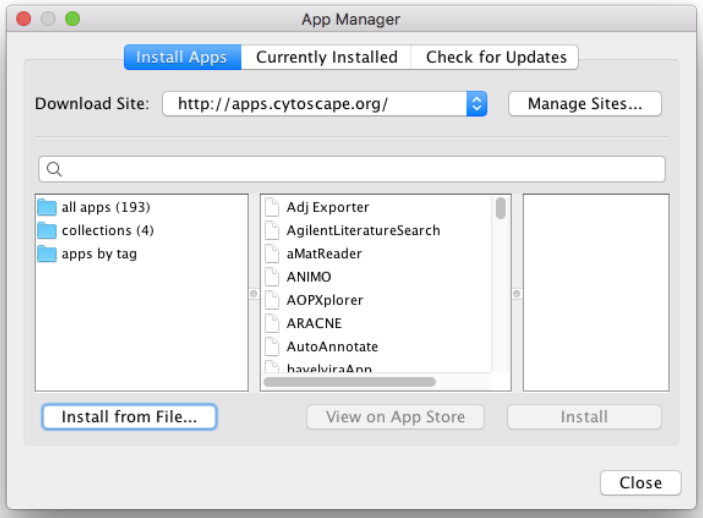


Fig. 1: Installing through the Cytoscape App Manager

Open the App Manager by choosing *"App Manager..."* under the *"Apps"* menu in Cytoscape. Press *"Install from File..."* and locate the .jar file that was downloaded.

2 The Query Builder

The Query Builder is a powerful search tool that lets the user construct complex queries with several variables to search for. By combining specific biological entities, which we will name *"nodes"*, or by specifying variables, the user is able to find results that depend on several factors at once. To open the Query Builder, click the *"Apps"* menu, and select *"BioGateway"* → *"Create query"*.

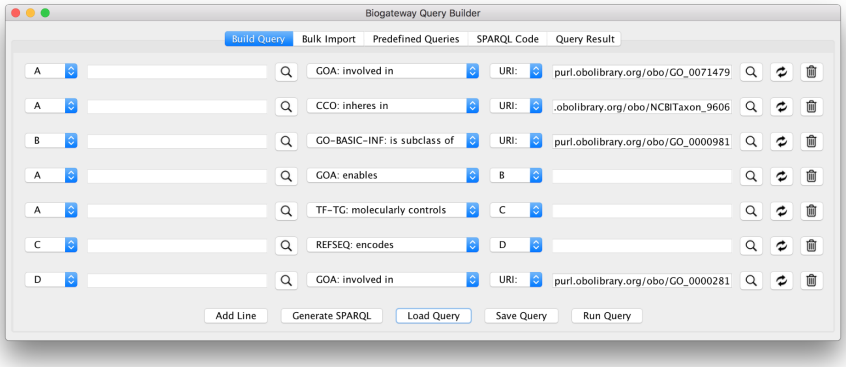


Fig. 2: The BioGateway Query Builder

2.1 Constructing Queries

A query consists of one or more rows, each row representing a relation between two entities. The nodes can be defined in one of two ways; as a URI representing an actual entity in the BioGateway database, such as a protein or a gene, or as a variable represented by a letter such as A, B, C, etc. The pulldown menus to the left of the text fields lets the user pick between *"URI:"* or assign the node to a variable letter.

The ordering of the nodes are important; each relation goes from the node left of it, to the node that is on the right side. The the 4th line of figure 2 specifies that *A enables B*, not the other way around.

To get additional lines, click the *"Add Line"* button in the lower left corner. To remove a line, click the garbage can icon on the far right side. To swap the node parameters of the left and right side of the relation to change the directionality of the relation, click the swap icon next to the garbage icon.

Fig. 3: A query line

2.1.1 Node URIs

When a node is set to be represented by an URI, the value in the text field will be used. All URIs starts with *"http://"*, and are valid web locations which describe the biological entity represented. The URI must be in the BioGateway database to work correctly. In order to get the correct URI for an entity, the Node Lookup view can be used by clicking the magnifying glass next to the URI text field. The URIs are also available in *"identifier uri"* column of node data tables created by the BioGateway plugin.

To see the name of the entity represented by an URI, hover the mouse over the

URI field. If the entity has been loaded from the server, it will be shown in the mouseover tooltip.

2.1.2 Variable letters

To denote any matching node, variable letters should be used. The variable letter *"A"* would mean *"any node in the database which satisfies the conditions..."*, where the conditions are the relations in the rest of the query. In figure 3, the variable *"A"* would represent all nodes which have the relation *"molecularly controls"* pointed to the node with the URI *"http://identifiers.org/ncbigene/7157"*, in this case the URI for the *TP53* gene.

2.1.3 Combining multiple lines

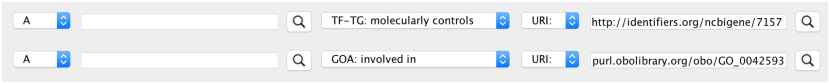


Fig. 4: Variables over multiple lines

While getting all the nodes which a specific relation to a specific node is useful, the real functionality in the Query Builder is the ability to combine several lines to get more specific results. By involving the same variable in several relations, only the nodes which satisfy all conditions are returned. In figure 4, the query from figure 3 has been expanded with a new relation, constraining the results to the entities which also are involved in a GO term, in this case *"glucose homeostasis"*, so the results would be all transcription factors for the *TP53* gene which are also involved in *"glucose homeostasis"*.

Queries can span several lines, and have variables on both sides of the relations. For instance, the query in figure 2 looks for the following:

All nodes *A* which:

1. Are involved in the *GO:0071479* term.
2. Inheres in *"homo sapiens"*.
3. Enables the GO term *B*.
4. Molecularly controls the gene *C*.

Where:

1. *B* is a subclass of the *GO:0000981* term.
2. *C* encodes the protein *D*.
3. *D* is involved in the *GO:0000281* term.

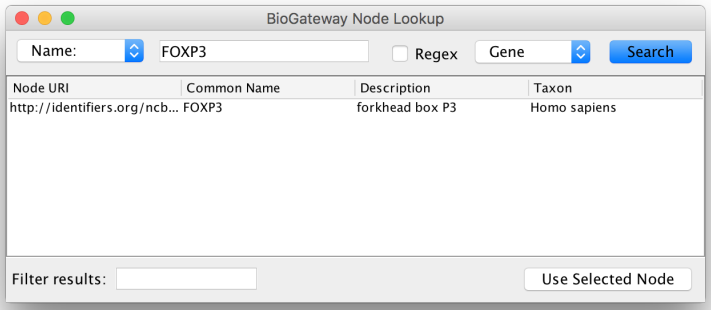


Fig. 5: Looking up the FOXP3 gene

2.2 Using the Node Lookup View

To easily look up a node URI, click the magnifying glass in figure 4. next to the URI text field. The Node Lookup view has several ways of searching for a node, which can be selected in the drop-down menu in the upper left corner. To use one of the resulting nodes, select it and click "Use Selected Node". If many results are found, the filter text field in the lower left corner can be used to limit the results.

2.2.1 Search by Name

In this mode, the BioGateway will search for nodes matching the name in the text field.

2.2.1.1 Regex The *"Regex"* checkbox indicates if BioGateway should search for partial matches to the search text. The regex search will match with anything that contains the search text. Normal regex symbols such as are usable here. The most notable symbols are . and *, meaning *"any character"* and *"zero or more occurrences of the previous character"*. Combining them to *".*"* means *"zero or more of any character"*, which can be useful when searching. For example, searching for GO terms named *"response radiation"* will not match anything, while *"response.*radiation"* will return all terms with *"response"* and *"radiation"* in their name.

Regex searches are much more resource-intensive than exact matches, and takes much longer time. Searches that are too broad might have too many results, and fail to complete.

2.2.1.2 Entity Type The type of biological entity to search for must be specified. The available types are:

- Protein

- Gene
- GO Term
- Taxon

Note that when searching for taxon, searches are with regex by default.

2.2.1.3 Search by URI This allows the user to verify the existence and name of a URI, and will try to fetch the node with this specific URI. There will only be one result. This mode is also useful when using the Node Lookup view from within a Cytoscape network to add new nodes.

2.2.1.4 Search by UniProt ID A URI will be generated by appending the UniProt ID to *"http://identifiers.org/uniprot/"*. For example, when looking up the UniProt ID *"P04637"*, it will generate the URI *"http://identifiers.org/uniprot/P04637"*. The URI lookup is the same as the URI search above.

2.2.1.5 Search by GO Term The term must be on the form *"GO: ..."*. A URI will be generated from the GO term. As an example, the term *"GO:0003674"* will result in a search for the URI *"http://purl.obolibrary.org/obo/GO_0003674"*. The URI lookup is the same as above.

2.3 Running queries

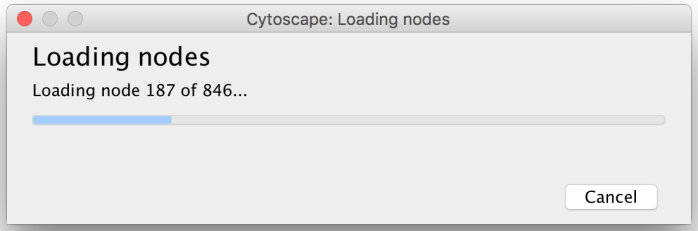


Fig. 6: In progress of loading nodes from the server

After clicking the *"Run Query"* button in the Query Builder, the query will be generated and sent to BioGateway. When executing a query, the results are loaded in several steps. If a result set is particularly huge, it might return thousands of nodes. The BioGateway app will try to reuse previously loaded nodes to avoid additional loading times, but will warn the user if more than 1000 nodes needs to be loaded, as shown in figure 6.

The warning will let the user cancel the query, proceed with loading all the nodes, or just show the nodes as URIs without loading their names and descriptions. This can be useful when only the nodes already loaded are needed.

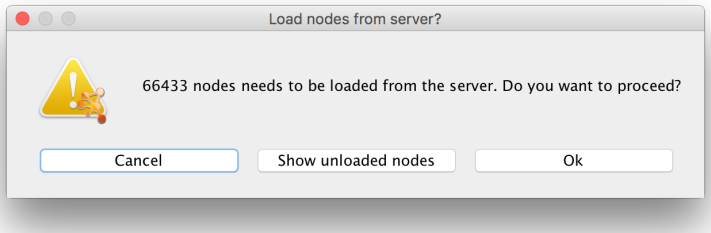


Fig. 7: A warning about a large result set that will take time to load

When the results are fully loaded, or if the user decides to not load all nodes, the Query Result window appears. This is similar to the Query Result tab of the Query Builder tool, but with a few differences.

2.4 The relation list

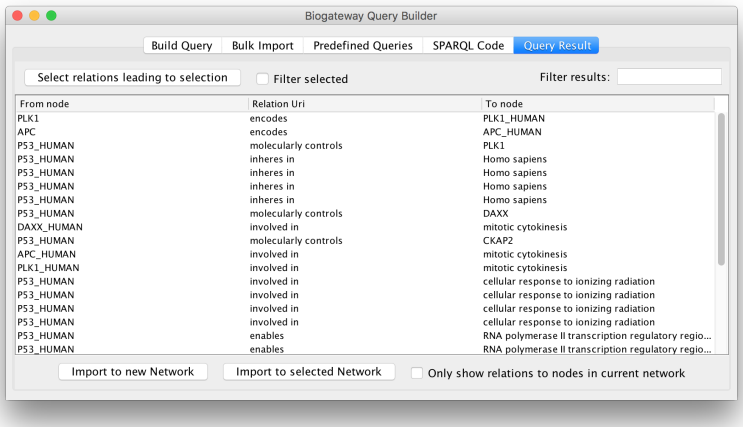


Fig. 8: The results after running the query in figure 2

The results from the query will be a set of all relevant relationships found between the matching nodes. If the query was on the form of $A \rightarrow B \rightarrow C \rightarrow D$, the relations between $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow D$ will be shown individually.

In the cases where a large amount of results are found, the *"Filter results"* text field can be used to limit the result set.

2.4.1 Select upstream relations

This button will select all relations that are pointing to the *"From node"* in one of the selected relations, recursively. This feature is useful for finding the rele-

vant path of relations leading to the relations selected, so they can be included in the network.

2.4.2 Filter selected

This checkbox toggles the filtering to only show the selected nodes, instead of filtering by text search. This can be useful to see the set of relations that are about to be imported.

2.4.3 Only show relations to nodes in current network

This checkbox will toggle a filter that will only show the results that include a node that exists in the current network. This is useful for comparing query results with the network that the user is actively working with.

2.4.4 Import to new/selected Network

These buttons will import the selected relations to a new Cytoscape network, or the currently selected one. New networks created with BioGateway will automatically apply the BioGateway visual style.

2.5 SPARQL Code

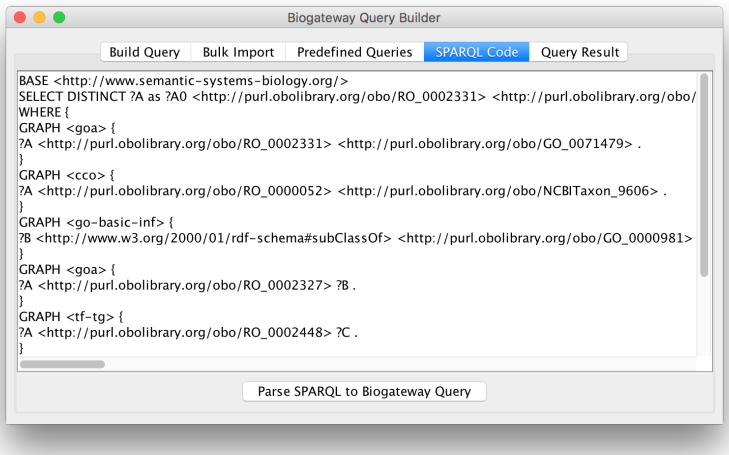


Fig. 9: The SPARQL Code viewer

Clicking the button marked *"Generate SPARQL"* will show the *"SPARQL Code"* tab, which shows the SPARQL query generated by the current query in the Query Builder. This tab allows the user to modify the SPARQL code,

and parse it back to the Query Builder. However, only minor changes are supported, as the SPARQL query must be structured in the exact same way for the Query Builder to represent it.

2.6 Saving and Loading queries

Queries built with the Query Builder can be saved and loaded as text files for later use. The files will be stored with the extension *.bgwsparql*, and stored in the same format as the SPARQL code shown in the *"SPARQL Code"* tab. The app will remember the last used location to save and load queries, so directories can be used to organise the stored queries.

3 Using the Bulk Importer

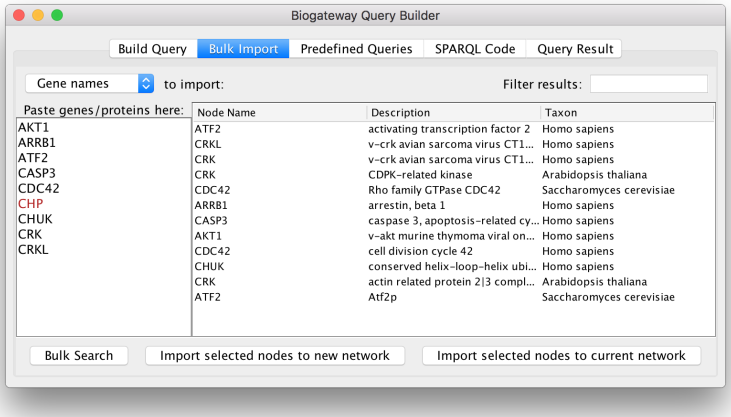


Fig. 10: Importing a list of genes with the Bulk Importer

The Bulk Importer is a part of the BioGateway Query Builder, and can be used to quickly import a set of genes, proteins or GO terms into Cytoscape in a BioGateway compatible network. The results can be filtered to match a specific name, description or taxon, and can be selected and imported into a new network, or the currently active one.

The nodes that were not found are marked in red after the query has been run.

3.1 Searching for gene/protein names

To search for genes or proteins by their names, select either *"Gene names"* or *"Protein names"* in the pull-down menu in the upper left corner. Paste a list of names to search for, with one name on each line. The tool will only search for exact matches of the names, for searching for partial matches, the *"Add BioGateway node"* feature described below must be used.

3.2 Searching for UniProt IDs

To search for a list of UniProt IDs, select *"UniProt IDs"* from the dropdown menu in the upper left corner. The tool will look for nodes matching the URI by appending the UniProt ID to *"http://identifiers.org/uniprot/"*. For example, when looking up the UniProt ID *"P04637"*, it will generate the URI *"http://identifiers.org/uniprot/P04637"*. The tool will then return these nodes if they are found.

3.3 Searching for GO terms

To search for a set of GO term IDs, select *"GO terms"* in the upper left corner. The terms must be on the form *"GO: ..."*. The tool will generate an URI from the GO term. As an example, the term *"GO:0003674"* will result in the URI *"http://purl.obolibrary.org/obo/GO_0003674"*. The tool will then return any nodes with the URIs generated.

4 Network right-click menus

Several functions of the BioGateway app are easily accessible by right-clicking in the network view of Cytoscape. Depending on whether none, one or several nodes are selected when right-clicking, different menus will be available. If several nodes are selected, it does not matter if the user right-clicks on the background or on a node. If no node is selected and the user right-clicks on a node, the action will be the same as if that node was selected.

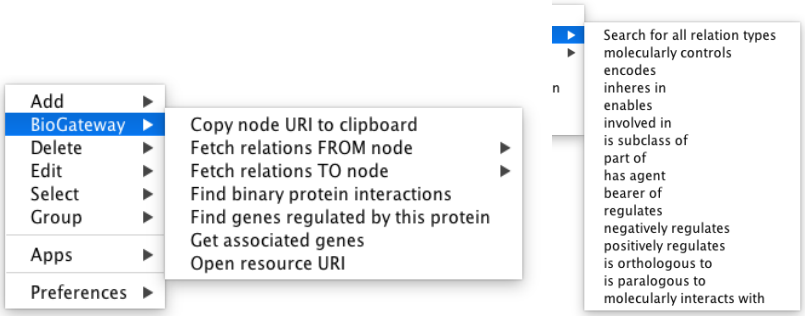
Right-clicking on an edge enables a specific set of actions.

4.1 Adding new nodes to the network

If the user right-clicks on the network background (not on a node or edge) while no node is selected, they will be presented with the option to *"Add BioGateway node"*. This is a handy feature for quickly adding specific nodes to a network, and will open the Node Lookup view described in section 2.2.

4.2 Relation source data

By right clicking on an edge and selecting "View source data", a window appears with additional information about the relation represented by the edge. For relations supported by PubMed IDs, these are available and can be opened in the default web browser for further review. The relation type definition can also be opened in the same way.



(a) The available functions and queries for a single node.

(b) Searching for a specific relation to or from a node.

Relation Type	From Node type	To Node type	Dataset graph
Molecularly controls	Proteins	Genes	TF-TG
Encodes	Genes	Proteins	Refseq
Inheres in	Proteins or Genes	Taxa	Refseq and Refprot
Enables	Proteins	GO Terms	GOA
Inheres in	Proteins	GO Terms	GOA
Has agent	Protein-Protein Interactions	Proteins	IntAct

Tab. 1: Relation directions for the most commonly used relation types.

4.3 Single node queries

4.3.1 Copy node URI to clipboard

This function will simply copy the selected node URI to the clipboard. This is useful for using the URIs of nodes in an existing network to construct a query with the Query Builder tool.

4.3.2 Fetch relations FROM/TO node

The queries *"Fetch relations FROM node"* and *"Fetch relations TO node"* are essentially the same, but searches in different directions. When searching *FROM*, the selected node is the node before the relation, and when searching *TO*, it is set as the node after the relation.

It is important to understand how the datasets are structured in order to know which direction the relations should be search for. Most datasets only have relations in one direction, going from one type of node to another type.

4.3.3 Find binary protein interactions

This menu action is only available if the selected node is a protein.

While *"binary PPIs"* is not a concept supported by the BioGateway server yet, this menu action will generate a query that will achieve the same effect. It will search for any PPI with only two participants where the source protein is one of them, and return a relation of the type *"molecularly interacts with"* pointing to the other participant.

When imported to a Cytoscape network the created edge will be labelled as bi-directional, and the BioGateway visual style will show it with arrows in both ends.

4.3.4 Find genes regulated by this protein

This action is a shortcut to searching for transcription factors or target genes. If the selected node is a gene, it will be shown as "Find proteins regulating this gene" instead.

4.3.5 Get associated genes/proteins

This action is a shortcut to getting the genes encoding a protein, or proteins encoded by a gene, as these are common actions.

4.3.6 Open resource URI

This menu action will open the URI of the selected node as an URL in the default web browser for further details about the biological entity.

4.4 Importing results

When executing a query from a network, the result view shown in figure 12 will be slightly different than the results in figure 2. As the network was initiated from a network's view, the results will always be imported to the same network. The *"Import Selected"* button in will import the relations in the selected rows into the currently active network.

The *"Import relations between existing nodes"* button will import all relations found in the current query, which are between nodes already present in the current network. No new nodes will be added. Note that this does not require any relations to be selected, all relations in the result which are between two existing nodes will be added. This feature is useful finding all relations of a current type in a network, especially if used in conjunction with the multi-node queries described later.

4.5 Multi-node queries

Multi-node queries work similar to the single-node queries, but they are applied to all selected nodes. Such queries can be an efficient way to expand a network.

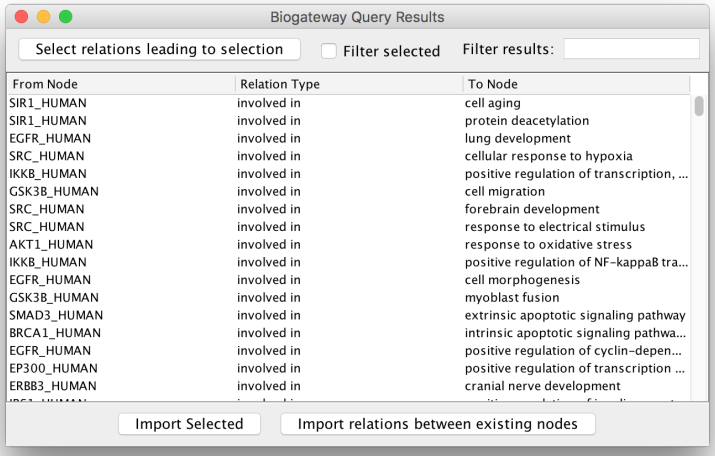


Fig. 12: The Query Result window

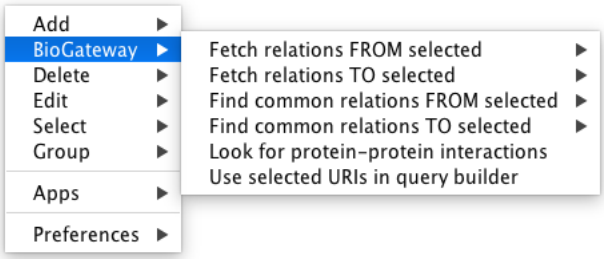


Fig. 13: The right-click context menu when more than one node is selected

4.5.1 Fetch relations FROM/TO selected

This query works just like its single node counterparts. All the resulting relations will be shown together in the Query Result window after completion.

4.5.2 Find common relations FROM/TO selected

This feature lets the user search over a set of nodes and only get the relations pointing to new nodes that have several relations with the ones in the searched set. For instance, if the user want to search for GO terms that a set of proteins are involved in, but filter out the less common terms, the user could set a minimum threshold of common relations.

Example:

- The user is looking for GO terms that at least 5 of the proteins in the search set are involved in.

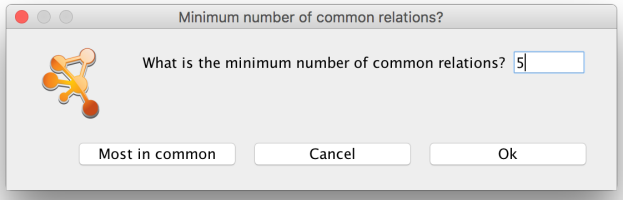


Fig. 14: Selecting the minimum number of common relations

- The user would type in *"5"* as the minimum number of common relations.
- Only the GO terms which at least 5 distinct proteins from the search set are involved in, are shown.
- The results can be sorted by number of common relations.

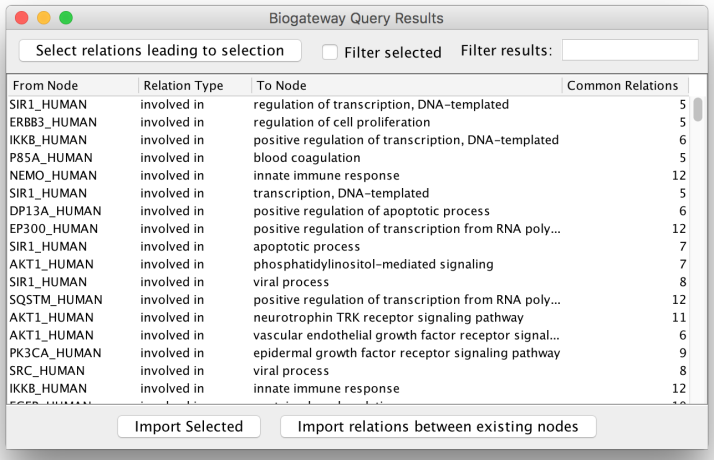


Fig. 15: Results from a common relations search

4.5.3 Look for binary PPIs

This query works in the same way as its corresponding single-node query, and will be executed for all the currently selected nodes. This option will only appear if at least one protein is among the selected nodes.

4.5.4 Look for common binary PPIs

This works the same way as finding common relations to/from selected nodes, described above, but searches for binary PPIs instead, as described above.

4.5.5 Use selected URIs in query builder

This feature will open a new Query Builder window with the URIs of the selected nodes already filled out on the left side. By using the swap button, the URIs can be placed on the desired side of the relation if needed.

5 Understanding the Datasets

BioGateway contains several graphs from many different sources. This is a short overview of some of the most relevant graphs, and which datasets they are imported from.

5.1 Gene Ontology

5.1.1 GO-BASIC

This dataset contains the GO terms from the Gene Ontology Consortium[1]. The most relevant relations in this graph are:

- *"subclass of"* - From GO terms to their parent GO terms.
- *"part of"* - A GO term can be involved in, or part of, another GO term.
- *"regulates"* - Regulations between GO terms.
- *"positively regulates"* - From GO terms to GO terms they positively regulate.
- *"negatively regulates"* - From GO terms to GO terms they negatively regulate.

5.2 Gene Ontology Annotations

The GOA dataset contains GO term annotations of proteins. The relations included are:

- *"involved in"* - for biological processes
- *"part of"* - for cellular components
- *"enables"* - for molecular functions

5.3 RefProt

This is a graph containing protein data from the UniProt Reference Proteomes[2]. It includes the relation types:

- *"inheres in"* - A relation from a protein to the organism (taxon) it belongs to.
- *"involved in"* - Relations from proteins to diseases they are involved in.
- *"bearer of"* - Relations from proteins to protein modifications.
- *"encodes"* - Relations from a protein to the gene(s) that encode it.

5.4 RefSeq

This graph contains data from NCBIGene's RefSeqGene[3] dataset. The relation types included are:

- *"encodes"* - Relations from genes to the proteins they encode for.
- *"inheres in"* - Relations from genes to the taxa they inhere in.

5.5 IntAct

The IntAct[4] dataset contains Protein-Protein Interaction (PPI) data. The most important relation in this dataset is the *"has agent"* relation. The graph does not consist of relations between proteins, but rather considers the PPIs as nodes themselves, and thus consists of the PPIs and their participating proteins, annotated as agents. While some PPIs are binary, and only have two participants, others have multiple participating proteins.

This dataset can be used to search for proteins participating in the same PPIs. The BioGateway app also has a helper function to find binary PPIs directly. In this case, the app will hide the PPI nodes and infer the *"molecularly interacts with"* relations between the two proteins. This only applies to PPIs with exactly two participants, to find proteins participating in non-binary PPIs, first search for PPIs with the *"has agent"* relation to the protein of interest, and then search for *"has agent"* relations from these PPIs. This approach should also be used when searching for PPIs in the Query Builder.

5.6 TF-TG

This dataset of transcription factors and target genes (TG-TF) is not currently published, but is available in BioGateway through a collaboration with the authors as a beta dataset. It contains almost 190 000 transcription factor annotations. This dataset should be used for testing purposes only until it is published.

6 The BioGateway Visual Style

The BioGateway app comes with its own visual style to highlight the different types of nodes and relations available through BioGateway. When creating new networks through the Query Builder, it will automatically be applied. If the user does not have any Cytoscape style named *"BioGateway"*, it will create a new one. The user is free to modify the style to their own liking, and modifications will not be overwritten when new networks are created, as long as the style is not deleted. To revert to the default BioGateway style, simply delete it from the Cytoscape visual style manager, and create a new network through the Query Builder.

References

[1] Gene Ontology Consortium. *Gene Ontology*. URL: `http://www.geneontology.org` (visited on 11/30/2017).

[2] UniProt Consortium. *UniProt Reference Proteomes*. URL: `http://www.uniprot.org/help/reference_proteome` (visited on 11/30/2017).

[3] National Center for Biotechnology Information. *RefSeqGene*. URL: `https://www.ncbi.nlm.nih.gov/refseq/rsg/` (visited on 11/30/2017).

[4] EMBL-EBI. *IntAct*. URL: `https://www.ebi.ac.uk/intact/` (visited on 11/30/2017).