# CyKEGGParser User Manual

## Table of Contents

## Introduction

CyKEGGParser v1.0 is an app for Cytoscape which operates on pathways derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. It provides the user with functionality of parsing and visualization of KEGG pathway maps in Cytoscape. Along with this, it provides an option for semi-automatic correction of inconsistencies between KEGG static pathway images and accompanying KGML files. The two main features of the app are the functionality of deriving tissue-specific pathways and for protein-protein interaction (PPI) based drill down of the pathways.

## Development

The app has been developed by the members of the Bioinformatics Group at the Institute of Molecular Biology of the National Academy of Sciences of the Republic of Armenia (IMB NAS RA). You can visit the group's webpage at the following link.

## Citation

When using the app in your research, please refer to the app webpage.

## License

# Getting started

*Installing the app from web*

In Cytoscape, go to Apps -> “App Manager”, choose CyKEGGParser under “Network generation, Network analysis, pathway database” categories and click on the install button. In case of successful installation, the plugin menu “KEGGParser” should appear under “Apps” menu.

# Pathway loading

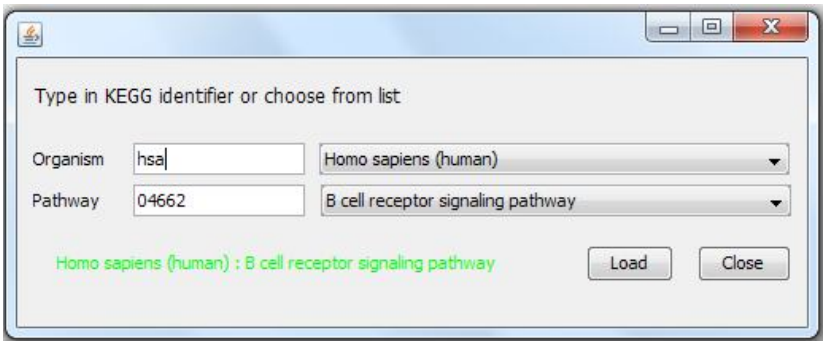
## Laoding KEGG pathways from local KGML files

Go to “Apps-> KEGGParser -> Load KGML -> Load local KGML” and choose the KGML file to be imported.

## Importing KEGG pathways from the web

Go to “Apps -> KEGGParser -> Load KGML -> load KGML from web“. Note that the window for web import will take a while to load organism and pathway lists.

From the drop down menu choose the organism and pathway names, or alternatively type in the organism identifier or KEGG pathway ID in the left-hand boxes. Press the “**Enter**” key after typing in order for the selection to take place. Press the “**Load**” button after choosing a pathway to load.




## Pathway parsing

After specifying the pathway to load it is parsed with current automatic correction options applied, and visualized in Cytoscape environment. Note that a visual style named “kegg_vs” is created. If the pathway is visualized with another style, simply go to Style panel and choose “kegg_vs” visual style. Because of concurrency issues, the visual style may fail to be set or may be set incorrectly. In this case the user will have to switch (or re-switch) to “kegg_vs” visual style manually and, if necessary, uncheck the “Lock node width and height” option to achieve KEGG-specific visual style.

Along with pathway parsing, all the meta-data contained in KGML files is mapped onto nodes and edges through creation of corresponding attributes in Cytoscape. If corrections are applied to nodes and edges, these will be mentioned with the attribute “Comments”.

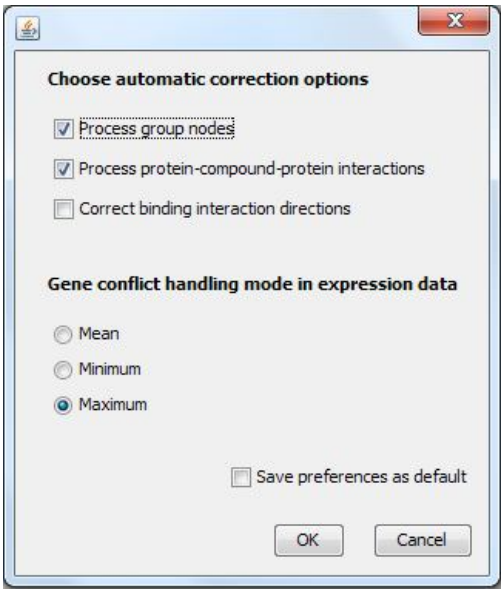
## Automatic correction options

In most cases KGML files do not fully correspond to the static pathway images. Inconsistencies may include absence of event or entity labels, reversed directions for some associations, absence of some interactions, ambiguous definitions of group nodes, compounds and their interactions.

CyKEGGParser is able to automatically correct for *some* of those inconsistencies based on meta-information contained in KGML files or on pathway topology. In other cases, though, the user will have to perform corrections manually, using Cytoscape functionality of adding, removing and modifying nodes and edges.

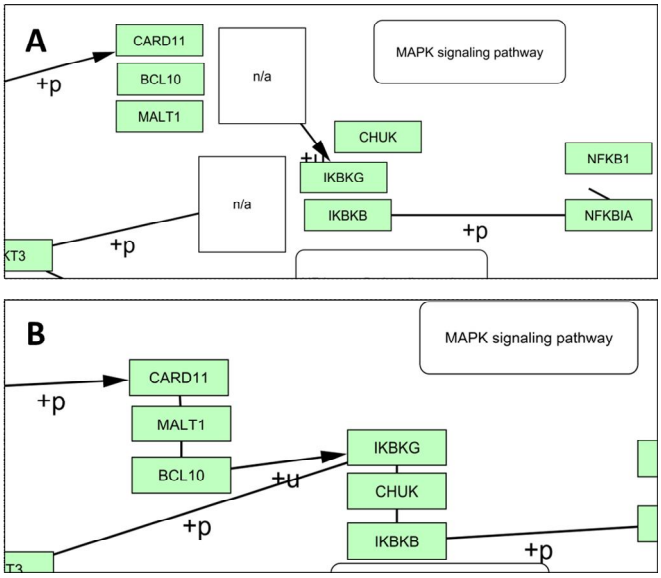
To adjust automatic correction options of go to Apps -> KEGGParser -> Preferences and check/uncheck the automatic corrections option boxes.

Those are explained in detail below.




## Group node processing

In KEGG pathways nodes are combined into groups, if protein complex formation is a necessary step for downstream events to take place. In KGML files, there are separate nodes for groups and for each of the group members, and after parsing with **"Process group nodes"** box unchecked they will appear in Cytoscape as an empty node (the group node) and member nodes separately. The edge distribution for these nodes varies and depends on the KGML file.



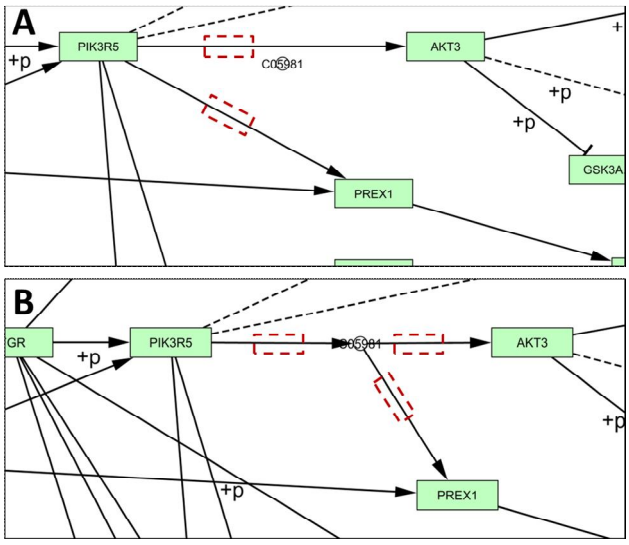

**(A) CARD11, BCL10, MALT1 and CHUK, IKBKG, IKBKB form two groups, which, according to the KGML file, appear separately from their group nodes (n/a). (B) After corrections, group nodes disappear, and the member nodes are connected.**

If "**Process group nodes**" option is chosen, during parsing the group node will disappear and its member nodes will be connected with binding edges and arranged in a sequence. The order of node arrangement depends on the incoming and outgoing relations of those nodes. This processing is important for preserving pathway flows, which may have significant effects on pathway-flow based algorithm results.

## Protein-compound-protein (PCP) interaction processing

Interactions of this type are represented as protein-protein interactions with “compound” interaction subtype. For example, in the KGML file of “Chemokine signaling pathway”, the interaction between nodes PIK3R5 and AKT3, and PIK3R5 and PREX1 is presented as:

```
<relation entry1="10" entry2="57" type="PPrel">
      <subtype name="compound" value="62"/>
      <subtype name="activation" value="--&gt;"/>
   </relation>
<relation entry1="10" entry2="45" type="PPrel">
      <subtype name="compound" value="62"/>
      <subtype name="activation" value="--&gt;"/>
   </relation>
```




**A. Direct protein-protein interactions between PIK3R5 and AKT3, and PIK3R5 and PREX1. The compound node is left out from the network. B. Protein-compound-protein interactions between those nodes recovered, after automatic processing.**

where the compound “62” is PIP3. While in the original KGML, PIP3 is not connected to any node, PCP fixing allows restoring the right order of connections in the final network.

Note that, PCP interaction processing is of interest for pathway flow recovery and for better visualization. If, instead, direct protein-protein interactions are of particular concern (e.g. for further network drill-down based on protein-protein interaction data), processing of PCP interactions should be turned off. This specifically matters in case of metabolic networks.

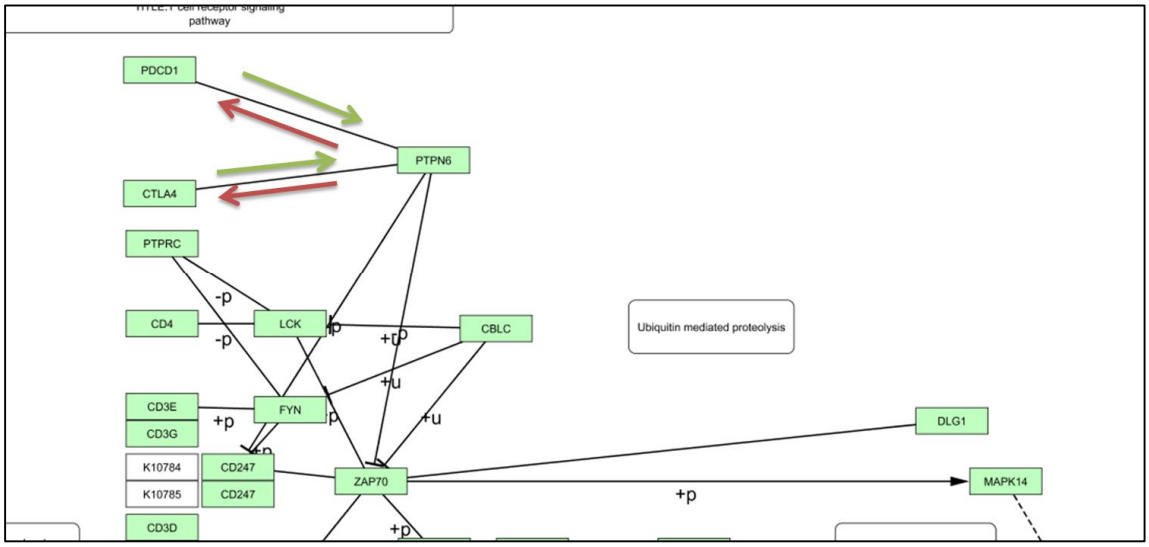
## Correction of binding interaction directions

Although a binding event is non-directional in terms of molecular interactions, the direction of an event makes sense in the context of information flows in a graph: reverse directions may lead to flow disruptions. Direction fix is based on relative positions of two interconnected nodes. The node upstream in the pathway is considered as the source, while the other is considered as the target.

In NOD-like receptor signaling pathway, the binding interactions “PDCD1 to PTPN6” and “CTLA4 to PTPN6” are represented in the KGML file as follows:

```
<relation entry1="61" entry2="63" type="PPrel">
      <subtype name="binding/association" value="---"/>
</relation>
<relation entry1="61" entry2="62" type="PPrel">
      <subtype name="binding/association" value="---"/>
</relation>
```

These directions are, in fact, opposite to the real picture of signal flow. Automatic corrections of binding interaction directions allow for reversing these edges. In some cases, however, the incorrect edges might still remain in the network, thus the user should be careful when directionality matters.




**Results of corrections of protein-compound-protein interaction directions. Red arrows indicate directions in KGML files, green arrows indicate reversed directions.**

## Parsing result logging

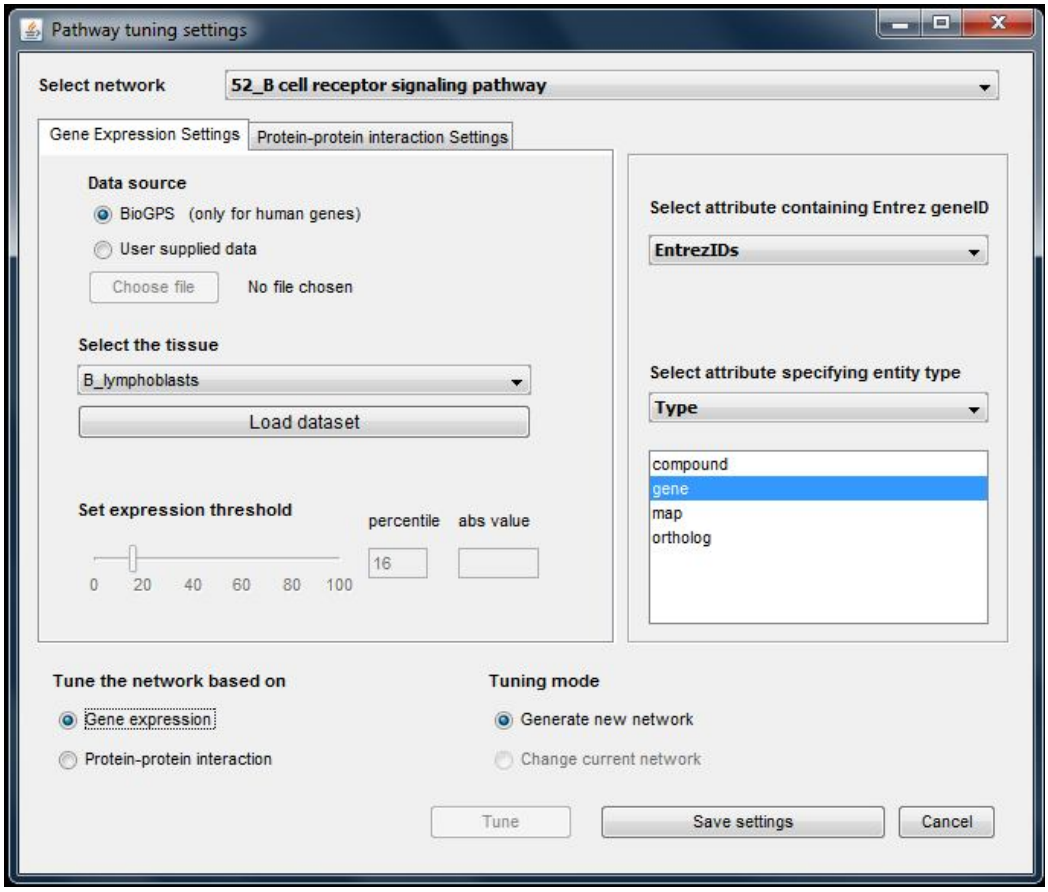
The report of the results is kept in “logs/parsing.log” file, in the plugin’s directory (by default, this is $USER_HOME/CytoscapeConfiguration/app-data /CyKEGGParser/).

## Pathway tuning

KEGG pathways are generalized pathways, with nodes representing collections of gene products, with abstract interactions between them, not always occurring together in the same biological context. Specifically, one of the abstractions is that paralogous genes/gene products are grouped into one node. Another abstraction is that pathways are drawn under the assumption of a generalized cell, in which all the genes are assumed to be expressed. Additionally, it's important to take into account the source of information on interactions depicted in pathways, and be able to filter among real physical and indirect interactions, as well as experimental and prediction-based sources.

All of these may have crucial effects on various types of automated pathway-based analysis algorithms, and therefore should be handled appropriately. Therefore, we have implemented functionality for customizing pathways according to specific biological context: particular tissue or cell type and experimentally confirmed physical interactions.

To perform pathway tuning, go to "Apps -> KEGGParser -> Pathway tuning". From the top drop-down menu ("**Select network**"), choose the pathway to be tuned. In the bottom of the panel, choose the tuning type ("**gene expression"** or "**protein-protein interaction**"), as well as choose the tuning mode ("**generate new network"** or "**modify current network"**). Specify corresponding tuning settings and press the Tune button.



## Tissue-specific pathway tuning

In the tuning window navigate to "**Gene Expression Settings**" tab and choose the source of gene expression data. For human pathways, the user can choose to tune based on the dataset provided by GeneCards (http://www.genecards.org/), which contains BioGPS data, representing log-transformed and normalized data for human normal tissue and cancer cell lines (http://biogps.org/#goto=welcome). Alternatively, the user can supply gene expression data in the specified format.

### User supplied gene expression data format

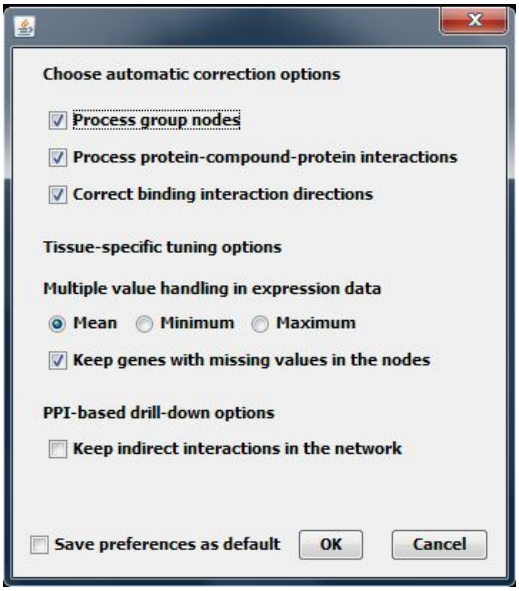
The data should be presented in tab-delimited format, where the first column represents gene identifiers (Entrez Gene IDs) and the remaining columns are for gene expression values for different tissues or cell-types (one column per each tissue). The first line of the first column should be named "ID". The first line of each tissue column should contain the tissue header, and the rest of the rows should represent gene expression values. Missing values should have value 0.0.

| ID | B_lymphoblasts | Adipocyte2 | AdrenalCortex | Adrenalgland |
|---|---|---|---|---|
| 1 | 4.0000 | 4.1000 | 4.1500 | 4.1000 |
| 2 | 6.1100 | 120.5200 | 165.3700 | 0.0000 |
| 9 | 87.7222 | 23.4444 | 19.5611 | 32.0944 |
| 10 | 5.1500 | 4.2500 | 4.5667 | 3.6167 |
| 12 | 4.6500 | 56.6000 | 7.7000 | 13.8500 |
| 13 | 7.5333 | 5.8333 | 23.3000 | 21.2000 |
| 14 | 132.7500 | 20.9000 | 19.2500 | 16.0500 |
| 15 | 4.4000 | 3.4500 | 3.8000 | 2.9500 |
| 16 | 769.7833 | 217.6833 | 169.8833 | 167.7333 |
| 18 | 12.0000 | 15.3333 | 22.6167 | 14.7500 |
| 19 | 15.6929 | 20.3238 | 20.8381 | 19.0262 |
| 20 | 7.3667 | 6.5667 | 6.5500 | 5.3667 |

### Tissue-specific tuning options

If gene identifiers are repeated in the dataset they are handled in the manner specified by the user in "**Preferences**". In order to change gene conflict handling settings go to "Apps -> KEGGParser -> Preferences" and choose one of the three modes: mean, minimum or maximum. If the mean mode is chosen, gene expression values will be averaged for the same gene. In case of minimum or maximum mode the minimum or maximum values will be chosen among single gene expression values for a particular tissue.

Those genes whose expression values are not found in the dataset may either remain untouched in their nodes or removed from the network, upon user's choice. This option can be adjusted by going to "Apps -> KEGGParser -> Preferences" and checking (unchecking) the "Keep the genes with missing values in the nodes" checkbox.




## Attribute specification panel

After CyKEGGParser parses KEGG pathways, it creates a node attribute in Cytoscape, named "EntrezIDs", where the set of Entrez IDs for each node is kept. After performing protein-protein interaction based tuning and drill-down, the Entrez ID for each gene is kept in "entrezId" attribute. Alternatively, the user may load another attribute data, where the identifiers are kept. The attribute should be selected and set from the "**Select attribute containing Entrez gene ID**" drop-down menu. "EntrezIDs" attribute is set as default.

Additionally, in "**Select attribute specifying entity type**" menu, the user should also select the entity type to be processed by gene expression based pathway tuning. For this, they should choose the attribute name and its value(s): e.g. if "gene" entity type is chosen, only the nodes having "gene" value of the attribute "Type" will be considered for removal during the tuning process.

## Tissue and threshold specification

After specifying the gene expression data source, the tissues contained in the dataset are listed in the "**Select the tissue**" drop-down menu. After choosing the tissue, the "**Load dataset**" button is activated, clicking which will result in loading the gene expression data for the specified tissue and the nodes in the network, according to chosen attributes. After a while the data will be loaded and the user will be able to set gene expression threshold. The latter is based on percentiles of the distribution of gene expression values for the specified tissue. Along with those, the absolute value corresponding to each percentile is displayed in the "**abs value**" box.

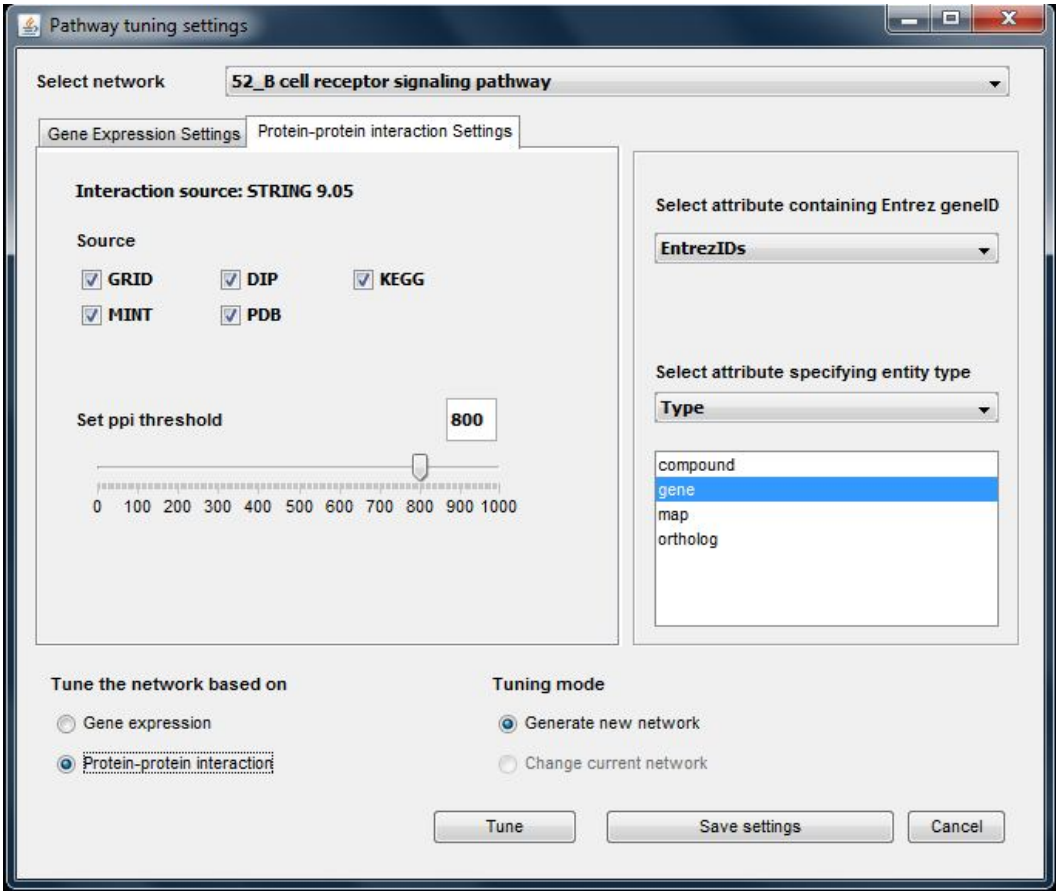
## Tuning results logging

All the genes with expression values higher or equal to the specified threshold will remain in the network after tuning. Since each node contains a number of genes, the gene ids contained in "EntrezIDs" attribute will be removed from the network. If no gene ID remains in a node, the node will entirely be removed from the network. If no other entity type is specified aside from "gene", the other entities, such as compound and labels, will remain as they are.

The report of the results is kept in "logs/tuning.log" file, in the app's directory (by default, this is $USER_HOME/CytoscapeConfiguration/app-data /CyKEGGParser/).

## Protein-protein interaction based tuning

Go to “Apps -> KEGGParser -> Pathway tuning”. From the “**Select network**” drop-down menu, choose the pathway to be tuned. In the tuning window navigate to “**Protein-protein interaction Settings**” tab. Here the user can choose the source of information about interactions and set the confidence score threshold. In the bottom menu choose “**Protein-protein interaction**” tuning type and press the “**Tune**” button.




The PPI tuning performs by initially drilling down the pathway through expanding each node of “gene” type into separate nodes for each gene. The gene-node label equals to Entrez ID of the gene. Furthermore, the algorithm iterates on all the pairs of interacting nodes, and connects those member gene-nodes for which there is real physical interaction. The existence of a physical protein-protein interaction is based on the data provided by String database (http://string-db.org/), which contains data from GRID, DIP, KEGG, MINT and PDB databases. The algorithm queries the database for the interactions between genes in the pathway according to data sources and confidence score provided by the user. In String, the confidence score is derived by combining evidence about protein-protein interactions from various sources, adjusted for probability of randomly observing the interaction. More information about confidence score meanings and interaction sources can be found at the Help page of String database: http://string-db.org/newstring_cgi/show_info_page.pl?UserId=z7Cu7ePjQXnI&sessionId=rD_f8EsHGmAQ. The interactions are manually updated in the local My-SQL database and the version of String used is mentioned on the Tuning dialogue.

If the “Keep indirect interactions in the network” checkbox is checked, the indirect interactions will remain in the network if those are not found in the interaction database, and will be removed otherwise.

Note that in some pathways the nodes are duplicated in order to keep the pathway view cozy. Drill down combines these nodes into one, redirecting corresponding interactions.

If one of the nodes is not of type “gene” (compounds and entity labels), the interactions between expanded nodes will remain as they are. Finally, all the nodes that appear unconnected will be removed from the network.

Note that in the drilled down network, gene-nodes originating from one node in the original network will appear on top of each other. The user will have to allocate the nodes apart or apply Cytoscape layouts to see all the genes and their interactions in the network.

The report of the results is kept in “logs/tuning.log” file, in the app’s directory (by default, this is $USER_HOME/ CytoscapeConfiguration/app-data /CyKEGGParser/).

## Saving the pathway

In order to save the pathways in KGML format, go to “Apps -> KEGGParser -> Save network -> Save as KGML”. All the modifications done to the network, including tuning based changes and added or removed nodes and edges, are saved in the attributes specific to KGML format. If the necessary attributes values are missing in Cytoscape, those either remain blank, assigned default values or Cytoscape-inherent properties values (such as color and coordinates) in the resulting KGML file (see below).

In addition, CyKEGGParser uses KEGGTranslator in order to convert the pathways in BioPAX_level2 and BioPAX_level3 formats (http://www.biopax.org/documentation.php). During conversion, the network is initially saved as a KGML file, which is then given as input to KEGGtranslator. The latter is called through command line call with the following arguments: “java -jar KEGGtranslator.jar --input [in_file.xml] --output [out_file] --format [out_format]”.

KGML format saving assures that all the attributes required for BioPAX translation are available. For nodes, these are “entry: id” and “entry: type” attributes: these are assigned the default values (the next maximum id in the network and “gene” respectively). Node color and coordinates are not required for format conversion, however, if those are missing in the attributes, they will be assigned the values they have in Cytoscape. For interactions, the required attribute is the “type” attribute. If this attribute is missing in the network, it is assigned a value based on source and target node types, as follows: if both nodes are of type “gene” the interaction is assigned the type “PPrel”, if either of the nodes is of type “compound”, the type “PCrel” is assigned, otherwise the algorithm looks at the required format: in case of KGML and BioPAX_level2, it will assign “maplink” type to those interactions, where one of the nodes is of type “Map”, and will give the default value “PPrel” otherwise; in case of BioPAX_level3 the interaction will be removed from the network, since the latter format does not allow for interactions between non-physical entities.

Be aware, that because of concurrency issues, the pathways may sometimes fail to convert to BioPAX format, in this case, try converting it in another format and/or using Keggtranslator independently.

Saving back to KGML format, allows for using modified and corrected KGML pathways in other applications and analyses. The report of the results is kept in “logs/parsing.log” file, in the app’s directory (by default, this is $USER_HOME/ CytoscapeConfiguration/app-data /CyKEGGParser/).

## Future developments

### Metabolic pathway parsing

In contrast to other pathway types, KGML files of metabolic pathways contain <reaction> tags, which contain information about substrate and product compound transitions occurring with participation of an enzyme. At the moment, CyKEGGParser does not rely on these interaction tags, but uses only <relation>-s. In some cases, this causes disruption of interactions between pathway nodes. These, however, are fixed during pathway drill-down. In future releases, CyKEGGParser will be adjusted for correct handling of <reaction> tags.

## Application of pathway signal flow algorithm

Our group has developed pathway signal flow (PSF) algorithm [*Arakelyan, A., Aslanyan, L. & Boyajyan, A. (2013). High-throughput Gene Expression Analysis Concepts and Applications. Sequence and Genome Analysis II – Bacteria, Viruses and Metabolic Pathways. ISBN: 978-1-480254-14-5. iConcept Press*].

Integration of PSF algorithm in CyKEGGParser will allow for calculation of signal flows in KEGG pathways.

## And of course…bug fixing and improvements!

Keep positive, there are not many of them ;)