A Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks

USER GUIDE
Copyright (C) 2010

Anna Bauer-Mehren, Michael Rautschka, Ferran Sanz, Laura I. Furlong
Integrative Biomedical Informatics Laboratory, Research Group on Biomedical Informatics (GRIB) - IMIM/UPF
PRBB Dr. Aiguader 88, E-08003 Barcelona (SPAIN)

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, with the Front-Cover Text being DisGeNET user guide, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".
DisGeNET user guide

1. Installation guide ........................................................................................................... 3
   1.1. Download and install DisGeNET .............................................................................. 3
   1.2. Troubleshooting ....................................................................................................... 5
   1.2.1. Allocating more memory ................................................................................. 5
   1.2.2. Download and installation problems ................................................................. 5

2. DisGeNET database ........................................................................................................... 6
   2.1. Original data sources ............................................................................................ 6
   2.2. Generation of gene-disease networks .................................................................... 7
   2.3. Mapping of disease vocabularies ........................................................................... 8
   2.4. Gene-disease association ontology ......................................................................... 8

3. Tutorial ............................................................................................................................. 10
   3.1. Basic functions ...................................................................................................... 10
       3.1.1. Generate gene-disease association network ...................................................... 10
       3.1.2. Generate gene or disease projection network .................................................. 12
       3.1.3. Restrict the network to a certain association type ......................................... 12
       3.1.4. Restrict the network to a certain disease class .............................................. 12
       3.1.5. Search for a particular gene/disease or set of genes/diseases ...................... 13
       3.1.6. DisGeNET LinkOut ..................................................................................... 13
       3.1.7. DisGeNET Expand ...................................................................................... 14
       3.1.7.1. Expand DisGeNET networks .................................................................. 14
       3.1.7.2. Expand foreign networks ....................................................................... 15
       3.2. Specific use cases ............................................................................................... 20
       3.2.1. Which are the genes annotated to breast cancer in expert curated databases? 20
       3.2.2. Do comorbidities observed in patients reflect a common genetic origin of the diseases? .................................................................................................................... 22
       3.2.3. Which are the diseases that are associated to post-translational modifications such as phosphorylation? ........................................................................................................... 22
       3.3. Analyzing DisGeNET data using external tools ....................................................... 25
       3.3.1. Extract data from DisGeNET database .......................................................... 26
       3.3.2. Build networks using igraph library ............................................................... 26

4. Contact .............................................................................................................................. 27
   4.1. Biomedical Informatics group ................................................................................ 27
   4.2. Citation ..................................................................................................................... 27
   4.3. Acknowledgements .................................................................................................. 28
   4.4. Contact .................................................................................................................... 28

5. Attribute tables .................................................................................................................. 29

6. References .......................................................................................................................... 30

7. GNU Free Documentation License ................................................................................... 31
1. Installation guide

1.1. Download and install DisGeNET

- Download DisGeNET.jar from http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download

- Put the jar (DisGeNET.jar) in the Cytoscape "plugins" folder. (The default location in Windows is C:\Program Files\Cytoscape-v2.x\plugins). The plugin will be automatically loaded the next time Cytoscape is started, and will appear as a menu item in the plugins menu. You can start the plugin by clicking on Start DisGeNET.

- The first time you start the plugin it will automatically download and unpack the gene-disease database (DisGeNET.db ~326,5MB) into a directory of your choice.
The download might take several minutes. When the download is finished, the plugin starts automatically.

Now the plugin is ready to be used.

The database folder can be changed at any time. Please restart the plugin to activate the changes.
1.2. Troubleshooting

1.2.1. Allocating more memory
Some of the networks are very large, especially when using LHGDN or ALL as source databases. In order to visualize large networks, you need to allocate more memory to Cytoscape. Memory usage depends on the number of nodes/edges and number of attributes. For detailed information check the Cytoscape manual available at http://www.cytoscape.org/. For Cytoscape version 2.7.0, you can find the information here: http://www.cytoscape.org/manual/Cytoscape2_7Manual.html#How%20to%20increase%20memory%20for%20Cytoscape

1.2.2. Download and installation problems

- Make sure you have writing permission for the Cytoscape subfolders
- Download is interrupted with NullPointerException (in Linux or Mac OSX)

⇒ Instead of starting Cytoscape via the icon, try to start it via command line from the installation folder, e.g.:
  sh /Applications/Cytoscape-6.x/cytoscape.sh
2. DisGeNET database

The DisGeNET database integrates human gene-disease associations from various expert curated databases and text-mining derived associations including mendelian, complex and environmental diseases (Bauer-Mehren, et al., 2010). The integration is performed by means of gene and disease vocabulary mapping and by using a gene-disease association ontology as described below.

2.1. Original data sources

OMIM: Online Mendelian Inheritance in Man (OMIM) focuses on inherited or heritable diseases (Hamosh, et al., 2005). Gene-disease associations were obtained by parsing the mim2gene file for associations of type “phenotype” (data was downloaded from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene on June, 6th 2009). All associations were labelled “phenotype” as provided in the mim2gene file and classified as Marker in our gene-disease association ontology. In total, we obtained 2198 distinct genes and 2473 distinct disease terms resulting in 3432 gene-disease associations. After mapping of disease vocabularies, the OMIM network contained 2417 distinct diseases.

UNIPROT: UniProt/SwissProt is a database containing curated information about protein sequence, structure and function (Apweiler, et al., 2004). Moreover, it provides information on the functional effect of sequence variants and their association to disease. We extracted this information from UniProt/SwissProt release 57.0 (March 2009) as described in (Bauer-Mehren, et al., 2009). All protein identifiers were converted to Entrez Gene identifiers in order to allow integration with the other data sources. All gene-disease associations were classified as GeneticVariation. UniProt provided 1746 distinct gene-disease associations for 1240 distinct genes and 1475 distinct diseases.

PHARMGKB: The Pharmacogenomics Knowledge Base (PharmGKB) is specialized on the knowledge about pharmacogenes, genes that are involved in modulating drug response. Genes are classified as pharmacogenes because they are (i) involved in the pharmacokinetics of a drug (how the drug is absorbed, distributed, metabolized and eliminated) or (ii) the pharmacodynamics of a drug (how the drug acts on its target and its mechanisms of action) (Altman, 2007). Hence, it covers less broadly human gene-disease associations but was found to be complementary to the other sources, as it contains some gene-disease associations not present in the other repositories. We downloaded the genes.zip, diseases.zip and relationships.zip from http://www.pharmgkb.org/resources/downloads_and_web_services.jsp on June 6th 2009 and parsed the files to extract gene-disease associations. We furthermore made use of the perl webservice to obtain all available annotations and supporting information. We included 1772 associations for 79 distinct genes and 261 distinct diseases. PharmGKB associations were classified as Marker if the original label was “Related” and as RegulatoryModification if the original label was “Positively Related” or “Negatively Related”.

CTD: The Comparative Toxicogenomics Database (CTD) contains manually curated information about gene-disease relationships with focus on understanding the effects of environmental chemicals on human health (Mattingly, et al., 2006). We downloaded the CTD_gene_disease_relations.tsv file from http://ctd.mdibl.org/downloads/ on June 2nd 2009 and parsed it for gene-disease associations of type “marker” or “therapeutic” (see
http://ctd.mdibl.org/help/glossary.jsp for description of the original labels). CTD includes associations from OMIM but with some differences (i) for some associations extra information such as cross-links to PubMed are available and (ii) some associations are missing in either of the two databases. Hence, we kept all available gene-disease associations from both sources. All CTD gene-disease associations were classified as Marker if the original label was “marker” and as Therapeutic if the original label was “therapeutic”. All cross-links to PubMed were kept. In total CTD data provided 6469 associations for 2702 distinct diseases and 3345 distinct genes.

**LHGDN:** The literature-derived human gene-disease network (LHGDN) is a text mining derived database with focus on extracting and classifying gene-disease associations with respect to several biomolecular conditions. It uses a machine learning based algorithm to extract semantic gene-disease relations from a textual source of interest. The semantic gene-disease relations were extracted with F-measures of 78 (see Bundschus, et al., 2008) for further details). More specifically, the textual source utilized here originates from Entrez Gene’s GeneRIF (Gene Reference Into Function) database (Mitchell, et al., 2003). This database represents a rapidly growing knowledge repository and consists of high-quality phrases created or reviewed by MeSH indexers. Hereby, the phrases refer to a particular gene in the Entrez Gene database and describe its function in a concise phrase. Using this textual repository for text mining has recently gained increasing attention, due to the high quality of the provided textual data in the GeneRIF database (Bundschus, et al., 2008; Lu, et al., 2007; Rubinstein and Simon, 2005). LHGDN was created based on a GeneRIF version from March 31st, 2009, consisting of 414241 phrases. These phrases were further restricted to the organism Homo sapiens, which resulted in a total of 178004 phrases. We extracted all data from LHGDN and classified the original associations using our ontology. In total, LHGDN provided 59342 distinct gene-disease associations for 1850 diseases and 6154 distinct genes. The LHGDN is also available in the Linked Life Data Cloud (http://linkedlifedata.com/sources).

### 2.2. Generation of gene-disease networks

Gene-disease associations were collected from several sources. The source databases use two different disease vocabularies (MIM and MeSH). Entrez Gene identifiers are used for genes (except for UniProt/SwissProt which uses UniProt identifiers). Moreover, the kind of association differs among the databases and ranges from the generic term “related” to more specific terms such as “altered expression”. In order to merge all gene-disease associations and to present them in one comprehensive gene-disease network, we (i) mapped UniProt identifiers to EntrezGene identifiers if necessary, (ii) mapped MIM to MeSH vocabulary if possible (see Mapping of disease vocabularies) and (iii) integrated associations through our gene-disease association ontology (see Gene-disease association ontology). We furthermore constructed different gene-disease networks for each source (OMIM, UNIPROT, PHARMGKB, CTD, LHGDN), as well as two integrated networks CURATED (containing gene-disease associations of OMIM, UNIPROT, PHARMGKB or CTD) and ALL (containing all gene-disease associations). Our comprehensive database is also available as SQLite database (DisGeNET.db). All gene-disease networks are represented as bipartite graphs. A bipartite graph has two types of vertices and the edges run only between vertices of un-like types (Newman, 2003). The bipartite graphs are multigraphs in which two vertices can be connected by more than one edge. In our networks, the multiple edges represent the multiple data sources reporting the gene-disease association. We generated two projections, one for the
2.3. Mapping of disease vocabularies

We used the MeSH hierarchy for disease classification. The repositories of gene-disease associations use two different disease vocabularies, MIM terms for OMIM diseases (used by OMIM, UniProt, CTD) and MeSH terms (used by CTD, PharmGKB, LHGDN). We used the UMLS metathesaurus to map from MIM to MeSH vocabularies. This step was performed to merge disease terms representing the same disorder, thus reducing redundancy. We were able to map 497 MIM terms directly to MeSH using UMLS and we additionally mapped 23 MIM terms by using a string mapping approach. Briefly, we searched the UMLS metathesaurus for MeSH terms for which there is at least one synonym exactly matching one of the synonyms describing the MIM term of interest. The resulting 63 matched terms were manually checked and reduced to 23 terms. For disease classification, we considered all 23 upper level concepts of the MeSH tree branch C (Diseases), plus two concepts (“Psychological Phenomena and Processes” and “Mental Disorders”) of the F branch (Psychiatry and Psychology). Moreover, we added one disease class “Unclassified” for all disease terms for which a classification was not possible. We categorized all diseases into one or more of the 26 possible disease classes. For MeSH disease terms we directly used its position in the MeSH hierarchy, for MIM disease terms that were not mapped to MeSH, we used the disease classification of (Goh, et al., 2007). Then, we mapped their disease classification to the MeSH hierarchy and extended the mapping using a disease classification available at CTD (CTD_disease_hierarchy.tsv downloaded August, 8th 2009). In total, we were able to classify 3980 (98.39 %) diseases. The disease classification allows filtering and searching of the network restricted to disease class.

2.4. Gene-disease association ontology

For a correct integration of gene-disease association data, we developed a gene-disease association ontology. We classified all association types as found in the original source databases into Association if there is a relationship between the gene/protein and the disease, and into NoAssociation if there is no association between a gene/protein and a certain disease (in other words, if there is evidence for the independence between a gene/protein and a disease). The different association types from the original databases were mapped to the ontology for a seamless integration. In this study, we only considered gene-disease associations of type Association. The ontology is available at http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download.
Figure 1: Gene-disease association ontology
3. Tutorial

DisGeNET is a plugin for Cytoscape (Shannon, et al., 2003) to query and analyze human gene-disease networks. For this purpose, we have developed a new gene-disease association database integrating information from several expert curated databases and a resource containing text-mining derived associations (Bauer-Mehren, et al., 2010).

3.1. Basic functions

By selecting different data sources, association types and/or disease classes from their respective drop-down menus, you can generate different gene-disease association networks. In addition, gene-disease association networks can be generated around a specific disease or gene of interest using the search box provided with the plugin. Most of these functionalities are also available to generate disease and gene monopartite networks.

3.1.1. Generate gene-disease association network

In order to obtain a gene-disease association network without any restrictions on association type and disease class follow the next steps:

- Select the source of interest, e.g. CURATED containing information from all expert curated databases in our database (OMIM, PHARMGKB, UNIPROT and CTD).
- Set Association Type and Disease Class Any
- Press Create Network
- Apply a Cytoscape layout algorithm to generate the view of choice, e.g. select the layout Organic

Once the network is obtained, specific information on the nodes and their relationships
can be explored as detailed below:

- Select nodes and edges and check their attributes.
- For example, use the Cytoscape search function to query for Alzheimer Disease. For this purpose, modify the search options and select the attribute diseaseName.
- Search for a particular disease, e.g. Alzheimer Disease
- Zoom into the network and select the Alzheimer Disease node
- More information about this node is found in the Node Attribute Browser
- All available node and edge attributes are listed in Tables 1 and 2.
- For this purpose you might want to select attributes to be displayed in the Node Attribute Browser or Edge Attribute Browser of the Cytoscape Data Panel.
- Select an edge to display information about a particular gene-disease association such as associationType, data source providing this association, supporting evidence (PubMed identifiers), etc.
3.1.2. Generate gene or disease projection network

In addition to bipartite graphs representing gene-disease associations, DisGeNET allows generating monopartite networks representing the gene or the disease projection of the gene-disease association network. In order to obtain the disease projection of the network generated from CURATED source (described in 2.1.1) follow the instructions detailed below:

- Select the *Disease Projection* tab in the DisGeNET main panel.
- Select the source, e.g. CURATED
- Press *Create Network*

3.1.3. Restrict the network to a certain association type

Note: This option is only available for Gene Disease Networks.

- Select the Source, e.g. CURATED
- Select the Association Type, e.g. *Genetic variation*
- Press *Create Network*

3.1.4. Restrict the network to a certain disease class

Note: This option is available for all types of networks. The classification is based on the disease branch of the MeSH hierarchy.

- Select the Source, e.g. CURATED
- Select the Disease Class, e.g. Digestive System Diseases
- Press *Create Network*
3.1.5. Search for a particular gene/disease or set of genes/diseases

The search option included in the DisGeNET tab can be used to generate networks around a disease or gene of interest. In addition, it can be used to search for a given disease or gene of interest in a network already generated.

- If only current net is **not ticked**, a network only containing associations related to the query will be created (using Create Network).

- If only current net is **ticked**, the according node will be selected (highlighted yellow) in the current network (with active view) when pressing [Enter].

- The search is restricted to Source, Association Type and Disease Class as selected.

- In this example, we are searching for any kind of Alzheimer Disease (there are four different types) in the CURATED dataset without any restriction of association type or disease class.

- Note: The DisGeNET search allows the use of the wildcard symbol (*). For performance reasons only the first 50 matching terms are listed in the drop-down box but all are included in the generated network.

3.1.6. DisGeNET LinkOut

In order to get more information about a gene or a disease node, you can linkout to the according website (Entrez Gene, OMIM or MeSH) using the DisGeNET LinkOut function. It is available in the node context menu, which can be accessed by right-clicking a selected node.

- For gene nodes, a linkout to Entrez Gene is given.

- For disease nodes, linkouts to MeSH or OMIM (depending on the type of disease node) are given.
3.1.7. DisGeNET Expand

In order to find all diseases/genes that are associated to a gene/disease node in an existing network you can use the DisGeNET Expand function. It can either be used to create new DisGeNET networks using the selected nodes for the query or to expand the existing nodes with edges found in DisGeNET.

Note: the function works with one or more selected nodes. To call the function, select one or more nodes, then click the right mouse button. This will open the node context menu containing the DisGeNET LinkOut and DisGeNET Expand functions. You can then choose between DisGeNET Expand -> Expand current net and DisGeNET Expand -> Build new net.

3.1.7.1. Expand DisGeNET networks

- This is a network generated with DisGeNET using as source OMIM, as AssociationType and DiseaseClass Any and as search term PSEN2.

- In OMIM, there is only one disease (Alzheimer disease-4) annotated to the gene PSEN2.

- The DisGeNET Expand function can be used to query for more associated diseases (click the right mouse button on the PSEN2 node to open the context menu). This function uses as data source the whole DisGeNET database. You can either add more gene-disease associations to the current net or build a new net.

- The result is this expanded network in which all found gene-disease associations for PSEN2 were added. You can see that there are 5 more diseases annotated to PSEN2.
• The expansion can be repeated various times. For instance, in a next step, we can expand this network by querying for more genes associated to *Alzheimer disease* - 4 and *Alzheimer Disease*.

• This results in a large network with 373 nodes and 893 edges. It is visible that there are many more genes associated to *Alzheimer Disease*.

### 3.1.7.2. Expand foreign networks

The same functionality to expand gene or disease nodes with more associations found in DisGeNET can be used to expand foreign network that were not created with DisGeNET but contain gene or disease nodes. In order to use the *DisGeNET Expand* function on nodes that were not built within DisGeNET, the node label needs to contain a valid Entrez Gene identifier or valid disease identifiers that are allowed by DisGeNET.

Note: DisGeNET only contains human gene-disease associations and hence can only be queried with human gene identifiers.

Examples for valid identifiers:

- **5080** for PAX 6 gene
- **mesh:D000544** for Alzheimer Disease
- **omim:217700** for Corneal endothelial dystrophy 2

In the following example, we show how a network not generated with DisGeNET can be expanded with DisGeNET gene-disease associations.
• First, we generate a network using the File->Import->Network from webservices function within Cytoscape.

• We query the Pathway Commons database for pathways containing the human gene PSEN2. For this, first set the Data Source to Pathway Commons Web Service Client, enter PSEN2 in the Search field and select the organism Human. Press search.

• We select a pathway we are interested in, for instance the NOTCH signalling pathway from the Cancer Cell Map database. Double-click the pathway to retrieve it.
• This results in a network with 113 node and 272 edges.
• The network contains the PSEN2 gene (PSN2_HUMAN). Moreover, there are various node attributes available among them the Entrez Gene identifier (biopax.xref.ENTREZ_GENE)

• In order to use DisGeNET expand, we need to ensure that the node labels contains the Entrez Gene identifier since DisGeNET uses node labels to query the database.

• To do so, we first create a new visual style in the VizMapper, for example called “ExpandDisGeNETStyile”

• Then, we set the node label to the attribute containing the Entrez Gene identifiers, here to biopax.xref.ENTREZ_GENE and use the Passthrough Mapping.
• Now, we can use the DisGeNET Expand function to search for gene-disease associations containing the selected node. Using the function for the PSEN2 node, we can search for all associated diseases in DisGeNET. We can either add the found associations to this net or create a new net.

• In the resulting network all diseases associated to PSEN2 are added

• For Cytoscape 2.7.0 users: You can make use of the Nested networks functionality to add the gene-disease association networks as nested networks to the nodes.

• To add the gene-disease association network as nested network to the PSEN2 gene node, right click on the node and select Nested Network -> Set Nested Network

• Now select the gene-disease association network for PSEN2 as created before using DisGeNET Expand
• The *PSEN2* node now contains the gene-disease association network as nested network, which can directly be opened by using the *Nested Network*->*Go to Nested Network* function.

• If the node label does not contain valid identifiers for DisGeNET, an error message is shown.

No valid identifier
There is no data available for selected nodes. Please check if node label contains disease or gene identifiers that are compatible with DisGeNET. e.g. `mesh:D000544` (MeSH ID for Alzheimer Disease) `omim:217700` (OMIM ID for Corneal endothelial dystrophy 2) `5080` (GeneID for PAX 6 gene)
For more information please check our Online Tutorial.
3.2. **Specific use cases**

In this section some examples that illustrate the kind of questions that can be answered using DisGeNET are presented.

3.2.1. **Which are the genes annotated to breast cancer in expert curated databases?**

This is an example of a more general question that can be phrased as “Give me all the genes known to be associated to disease x from a given data source”.

In order to answer this question, query the Gene Disease network selecting CURATED as source, no restriction on the association type or disease class, but specifying *Breast Neoplasms* in the search field to restrict the search to the genes annotated to this disease term. This will generate a network with 277 nodes (one disease and 276 gene nodes) and 417 edges. The edges are coloured according to the association type.

![Gene Disease Network](image)

Many genes associated to *Breast Neoplasms* are also annotated to other diseases. We can inspect these diseases by exploring the node attributes `associatedDiseases` in the Node Attribute browser and also by colouring the nodes according to MeSH disease classification. For this purpose, use the function *Colour nodes with disease class*. 
Breast Neoplasms is classified as Neoplasms and Skin and Connective Tissue Disease. MAP3K1 is a gene annotated only to Breast Neoplasms in the CURATED data set, while COL7A1 is annotated to 8 different diseases belonging to 4 different disease classes.

In order to know if there are other genes described in the literature but not recorded in the set of curated databases considered, we perform the query on the LHGDN set. This query will retrieve annotations derived from text-mining. The result is a network composed of 1099 genes annotated to Breast Neoplasms (1100 nodes and 3321 edges).

If we inspect the association between gene CDH1 and Breast Neoplasms, we see that there are 12 edges connecting the two nodes. The associations belong to different classes (Marker, GeneticVariation, etc.), hence they are coloured differently.

Furthermore, we can explore the supporting evidence for each gene-disease association by inspecting the edge attribute browser. We can examine the associations by either linking out to the according publication (using the Cytoscape function Search on the web).
Or we can view the sentence that was found by text-mining that supports the association between the gene and the disease (using the node attribute *sentence*).

This example illustrates the value of incorporating information from literature, since the curated databases currently don't cover all knowledge about gene-disease association available in the literature.

### 3.2.2. Do comorbidities observed in patients reflect a common genetic origin of the diseases?

This is a specific example of a more general question such as: Are diseases x and y related by genetic origin?

Some diseases are known to co-occur in a patient, a process known as disease comorbidities (Park, et al., 2009). Disease comorbidities can be studied considering the common genetic origin of both diseases. *Alzheimer Disease* and *Myocardial Infarction* are one example of comorbidity. By querying the Gene Disease network we can answer the question if these two diseases share a common genetic origin.

First, we query the Disease projection (CURATED) for *Alzheimer Disease* with no restriction to Disease class and create the network. Then, we search this network for *Myocardial Infarction* using the DisGeNET search function with the option *only current net* ticked.

We immediately see that both diseases are connected.
Once we know that there is at least one gene shared between both diseases, we can go back to the CURATED Gene Disease Network (or create it) and then create a subnetwork containing the two diseases and their associated genes. For this purpose, we first select the four nodes representing subtypes of Alzheimer Disease (Alzheimer Disease, Alzheimer disease-2, Alzheimer disease-4, Alzheimer disease, type 3) and the two nodes for Myocardial Infarction (Myocardial Infarction and Myocardial infarction, susceptibility to) [see section 3.1.5] and their associated genes using the Cytoscape function Select -> Nodes -> First neighbours of selected nodes.
Then, we create a subnetwork containing all selected nodes and all edges using the Cytoscape function **File -> New -> Network -> From selected nodes, all edges.**

As a result we obtain a network containing 62 nodes and 126 edges. We can see that *Alzheimer disease* and *Myocardial Infarction* are both annotated to the genes NOS3, ACE and APOE, supporting the hypothesis that alterations in the function of these genes can result in the development of both diseases in the same patient.

The same result can be obtained using the Cytoscape plugin “Advanced Network Merge”. For this purpose, we can create two separate gene-disease networks for *Alzheimer Disease* and *Myocardial Infarction* and then use the “Advanced Network Merge” to merge these networks using for instance the geneld as matching attribute.
3.2.3. Which are the diseases that are associated to post-translational modifications such as phosphorylation?

This is an example of a more general search query that can be expressed as: Give me all the diseases for which there are alterations in post-translational modifications such as x.

This type of use case might be of interest for drug discovery projects in which the identification of disease genes able to be targeted by drugs interfering with phosphorylation is needed.

First, we create a network querying the complete Gene Disease network (ALL) and restricting the Association Type to Methylation/phosphorylation (no restriction to any Disease class). The query results in a network composed of 621 nodes (157 disease nodes) and 1117 edges.

By exploring the diseases (use Colour nodes with disease class), it can be observed that most of them belong to the class Neoplasms, but there are other diseases such as those belonging to Nervous systems Diseases, Hemic and Lymphatic diseases, Immune Systems Diseases. The supporting evidence for each gene-disease association can be explored using the edge attribute browser as explained in section 2.1.

3.3. Analyzing DisGeNET data using external tools

Some of the networks can get very large, especially when using LHGDN or ALL and for this reason the plugin will not create gene projections with the LHGDN or ALL setting. In order to analyze DisGeNET data with external network analysis tools such as the igraph library for complex network research (Gabor and Tamas, 2006), we provide all networks and attributes in a sqlite database available at http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download.
3.3.1. Extract data from DisGeNET database

Sqlite can be downloaded from http://www.sqlite.org/download.html. Please also check the sqlite documentation for more information.

Connect to the DisGeNET database using the following command (call from the folder containing the DisGeNET database, e.g. ../cytoscape-v2.6.3/plugins/DB/):

```
sqlite3 DisGeNET.db
```

Use the following commands to extract the whole (ALL) gene-disease network and to write them into a tab delimited text file named DisGeNET_ALL.txt:

```
sqlite>
.mode tab
sqlite> .output ./DisGeNET_ALL.txt
sqlite> select * from geneDiseaseNetwork;
```

3.3.2. Build networks using igraph library

Once you have saved the network, you can access and visualize it with any external tools for network analysis. Many tools can read tab delimited text files such as the igraph library for R. The igraph library can be downloaded from http://igraph.sourceforge.net.

Start R and use the igraph library.

```
R
R> library(igraph)
```

Read in the network and build a graph object.

```
R> edges <- read.csv(file="./DisGeNET_ALL.txt", sep="\t", header=F)
R> graph <- graph.data.frame(edges, directed=F)
```

Now you can make use of a variety of graph manipulation functionalities available in igraph. For further information check the igraph documentation.
4. Contact

4.1. Biomedical Informatics group

We are interested in the understanding of the mechanisms underlying biomedical related problems at the molecular scale. This involves the study of the network of interactions between molecules that underlay, for instance, the etiology of a complex disease. In addition to the study of diseases of complex origin, we are also interested in the mechanisms underlying the appearance of side effects after drug treatments.

One part of our research is focused on strategies to, once a network of molecular interactions is obtained, characterize the network and model its behavior in order to gain insight into the etiology of the disease phenotype. In particular, we are interested in the application of qualitative modeling approaches, such as Petri Nets and Boolean networks.

Another line of research involves strategies for obtaining the networks that are relevant for the biomedical related problems already mentioned. For this, we are developing software for the retrieval and analysis of data from public network repositories (databases of signaling pathways, gene regulatory networks and metabolic reactions). Although the publicly available network databases contain valuable information, we are aware that their coverage is not complete: a lot of information regarding interaction between biomedical entities (genes, proteins, phenotypes, chemicals, drugs, etc) still lies in the biomedical literature as free text.

This is where our third line of research comes in, which involves the use of text mining approaches for the extraction of relationships between biomedical entities from the biomedical literature. In the past years we have developed NER systems for the identification of mentions of gene sequence variants from MEDLINE abstracts, and linkage of the mentions found in text to the corresponding database identifiers (in this case dbSNP). In addition, we have developed a corpus with annotations for variation mentions for the evaluation of this kind of NER systems. Currently, we are working on the application of NLP approaches for the identification and extraction of different types of relationships between biomedical entities.

4.2. Citation

If you are using DisGeNET for your own research, please cite:


4.3. **Acknowledgements**

This work was generated in the framework of the EU-ADR project co-financed by the European Commission through the contract no. ICT-215847 and the eTOX project from the European Community’s Seventh Framework Program (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement no. 115002. The Research Unit on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and member of the COMBIOMED network. We thank the Departament d’Innovació, Universitat i Empresa (Generalitat de Catalunya) for a grant to author ABM.

4.4. **Contact**

If you have question, comments or suggestions, please contact us.

Laura Furlong (lfurlong@imim.es)
Tel:  (+34) 93 316 0521
Fax:  (+34) 93 316 0550
http://ibi.imim.es/

Integrative Biomedical Informatics Laboratory
Research Group on Biomedical Informatics (GRIB) - IMIM/UPF
Parc de Recerca Biomèdica de Barcelona
Dr. Aiguader, 88
E-8003 Barcelona
5. Attribute tables

Table 1: Edge attributes in the gene-disease network

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>associationType</td>
<td>Association type of the gene-disease association according to the gene-disease association ontology (see Gene-disease association ontology).</td>
</tr>
<tr>
<td>interaction</td>
<td>Unique identifier for this association.</td>
</tr>
<tr>
<td>label</td>
<td>Association type as originally assigned by the source database.</td>
</tr>
<tr>
<td>pmids</td>
<td>List of PubMed identifiers of publications supporting the reported gene-disease association, if available.</td>
</tr>
<tr>
<td>sentence</td>
<td>The actual sentence in which the gene-disease association was detected (only available for LHGDN).</td>
</tr>
<tr>
<td>source</td>
<td>Database in which this gene-disease association was reported (OMIM, UNIPROT, PHARMGKB, CTD, CURATED, LHGDN, ALL)</td>
</tr>
</tbody>
</table>

Table 2: Node attributes in the gene-disease network

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>associatedDiseases</td>
<td>List of disease identifier associated to a gene node.</td>
</tr>
<tr>
<td>associatedDiseaseNames</td>
<td>List of disease names which are associated to a gene node.</td>
</tr>
<tr>
<td>associatedPathwayNames</td>
<td>List of KEGG and Reactome pathways the gene is annotated to (only for gene nodes).</td>
</tr>
<tr>
<td>associatedPathways</td>
<td>List of KEGG and Reactome pathway identifiers the gene is annotated to (only for gene nodes).</td>
</tr>
<tr>
<td>nrAssociatedDiseases/</td>
<td>Number of associated diseases or genes (number of first neighbours of the node).</td>
</tr>
<tr>
<td>nrAssociatedGenes</td>
<td></td>
</tr>
<tr>
<td>diseaseClass</td>
<td>List of disease class identifiers (disease classes according to MeSH hierarchy).</td>
</tr>
<tr>
<td>diseaseId</td>
<td>MIM or MeSH identifier for the disease node.</td>
</tr>
<tr>
<td>diseaseName</td>
<td>Name of the disease according to MeSH or OMIM morbidmap.</td>
</tr>
<tr>
<td>genelD</td>
<td>Entrez Gene identifier of the gene.</td>
</tr>
<tr>
<td>geneName</td>
<td>Name of the gene.</td>
</tr>
<tr>
<td>nodeType</td>
<td>The type of node (gene or disease).</td>
</tr>
<tr>
<td>styleName</td>
<td>Name of gene or disease, needed for the DisGeNET visual style.</td>
</tr>
<tr>
<td>styleSize</td>
<td>Number of first neighbours of the node, needed for the DisGeNET visual style.</td>
</tr>
</tbody>
</table>
6. References


7. GNU Free Documentation License

GNU Free Documentation License
Version 1.3, 3 November 2008


Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant.

The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "publisher" means any person or entity that distributes copies of the Document to the public.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses
following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History.") To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document numbering more than 100, and the Document’s license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document’s license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to
Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means; this is the first time you have received notice of violation of this License (for any work) from

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means; this is the first time you have received notice of violation of this License (for any work) from
that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (C) YEAR YOUR NAME.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.3
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU
Free Documentation License".
If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.