BiKMi User Manual

The **Bi**ological **K**nowledge **Mi**ner (BiKMi) is a web application developed by Fraunhofer SCAI for interfacing with the Knowledge Graphs housed in the Graph Store. BiKMi contains a number of useful tools for exploring the available models and for comparing experimental data against the manually curated biological relationships extracted from primary literature. This interface is a highly adaptable tool that can be modified to suite the needs and wishes of the user. This user manual is meant to serve as a guide for how to use BiKMi's tools and features to their fullest extent.

Overview

The **Overview** tab contains a collection of bar charts and other various quantitative data about the Knowledge Graph. Here, users can find detailed statistics on the types of nodes, edges, and sources of information that comprise our model.

Knowledge Graph

The Knowledge Graph itself is generated using OrientDB. You can access the Knowledge Graph directly by logging into OrientDB and running either your own SQL query or one of our bookmarked queries. The Graph Viewer in OrientDB will allow one to visualize the network directly.

Note that in order to access the knowledge graph you must login using the OrientDB applet.

Workspace

This section of BiKMi contains personalized user data repositories that can be used to compare against the manually curated data contained within the Knowledge Graph. Users can upload data to the different repositories described below which will then appear at the specified tools listed in their descriptions.

Note: In the current state of development, there is only one repository available, but more are in development. Thank you for your patience!

Differential Gene Expression Data

This repository is for uploading differential gene expression (DGE) data that will be compared to the manually curated BEL statements collected in the **Query: Path** tool. Once properly uploaded, pathway images generated in the **Query: Path** tool will also identify whether the uploaded data *contradicts* or *supports* the resulting BEL relations. For information on how to upload your own data, please see the Uploading Your Data section below.

Experimental vs. Published

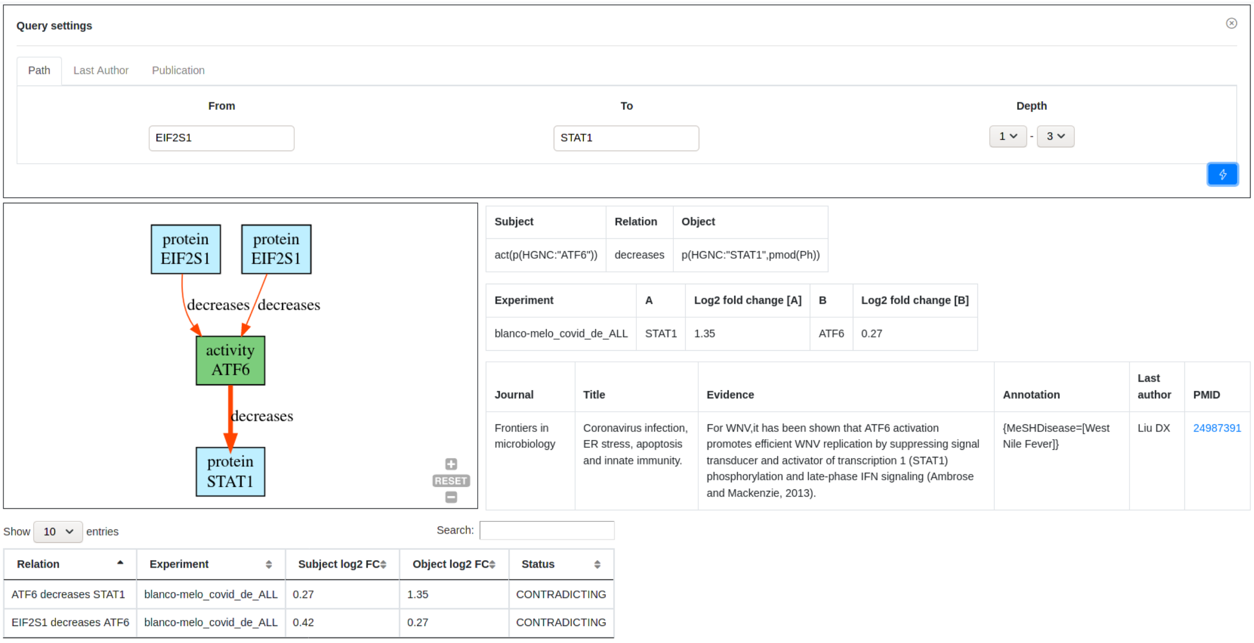
The DGE data uploaded in this repository can be used to check whether your experimental either *supports* or *contradicts* published biological relationships. It does this by comparing fold-change relationships in the experimental data to the relationships found in the Knowledge Graph. BEL relations are categorized as either direct (when A increases, so does B or vice versa) or inverse (when A goes up, B goes down or vice versa). A table describing how the BEL relation classes are categorized can be found below.

BEl Relation	Relationship Category
increases	direct

BEI Relation	Relationship Category
directly_increases	direct
decreases	indirect
directly_decreases	indirect

When generating pathway images in the **Query: Path** tool, users can click on the edges between any two nodes to bring up information on that relation. BiKMi will also parse all of the uploaded DGE data and identify matching pairs within the datasets (if they exist). For example, in the image below, a path was found when EIF2S1 was selected as the start point and STAT1 as the end point. All of the gene symbols that appear in the resulting image are compared against the symbols in the uploaded DGE data, and a table containing all possible pairs is created beneath the pathway image. This table includes information about how the manually curated BEL relations were mapped to the DGE data as well as if the DGE data *supports* or *contradicts* the BEL relation as indicated in the `Status` column. If the pair of gene symbols in the DGE were found to have the same type of relationship (direct/indirect) as the manually curated BEL relation, then it would receive a "SUPPORTING" tag in the `Status` column, otherwise if they differ, then it would instead receive a "CONTRADICTING" tag.

In some cases, the generated pathway image may be quite large and the collective table underneath may contain several comparisons. In this instance, users can select individual edges in the image to obtain comparative information between the involved nodes which will appear underneath the BEL relation on the right. In the below example, an edge was selected that joins nodes involving ATF6 and STAT1. These symbols are then identified in the uploaded DGE data and the corresponding information is displayed on the right side.



Query settings

From: EIF2S1

To: STAT1

Depth: 1 - 3

Subject	Relation	Object
act(p(HGNC:"ATF6"))	decreases	p(HGNC:"STAT1",pmod(Ph))

Experiment	A	Log2 fold change [A]	B	Log2 fold change [B]
blanco-melo_covid_de_ALL	STAT1	1.35	ATF6	0.27

Journal	Title	Evidence	Annotation	Last author	PMID
Frontiers in microbiology	Coronavirus infection, ER stress, apoptosis and innate immunity.	For WNV,it has been shown that ATF6 activation promotes efficient WNV replication by suppressing signal transducer and activator of transcription 1 (STAT1) phosphorylation and late-phase IFN signaling (Ambrose and Mackenzie, 2013).	{MeSHDisease=[West Nile Fever]}	Liu DX	24987391

Show 10 entries

Search:

Relation	Experiment	Subject log2 FC	Object log2 FC	Status
ATF6 decreases STAT1	blanco-melo_covid_de_ALL	0.27	1.35	CONTRADICTING
EIF2S1 decreases ATF6	blanco-melo_covid_de_ALL	0.42	0.27	CONTRADICTING

Uploading Your Data

DGE data can be uploaded in this workspace by clicking on the symbol in the top-right of the page. When you decide to upload your data for comparison, it is important to properly format it prior to import. Files should be imported as comma-separated (`.csv`) or tab-separated (`.tsv`) format. Additionally, your data **must** contain the following columns (named exactly as shown here):

- "HGNC": Column with the approved HGNC symbol for the measured expression value.
- "log2FoldChange": The log2 fold change of the experimental value compared to its matched control.
- "pvalue": The calculated p-value for the calculated fold change. If not applicable, then type 0 in this column.

Your data can contain extra columns, but it must have these three, properly labelled for it to successfully upload.

IMPORTANT NOTE: We highly recommend that you filter your dataset for significant values prior to upload. BiKMi does not filter uploaded data! We suggest filtering for values with a log2 fold change of more than 2, and a p-value < 0.05.

Description

Once the data is properly formatted and filtered, users can then upload their data. Users will be taken to a page where they can give a brief description on the dataset being uploaded. This will be a personalized workspace, therefore, this description is solely for the user to be able to quickly identify their datasets.

Query

In order to properly explore the data available inside the Knowledge Graph, we have developed several tools that allow users to both extract specific information on proteins/genes of interest as well as explore the molecular relationships that were manually curated from primary literature sources. Here, we discuss each of the tools and how users can best utilize them.

Jump to:
Protein Profiler
Protein List
BELish
Path

Protein Profiler

This interface is used for gathering information on proteins found inside the Knowledge Graph. To use it, simply enter an approved HGNC symbol for a given protein in the query box. Our Knowledge Graph compiles information on proteins from multiple sources including: UniProt, Reactome, HGNC, Ensembl, and MGI, as well as related drugs from Drugbank. This tool will present all of the data associated with the queried protein that is available in our database.

In its current state, you will only find information on a protein if has been added to the Knowledge Graph via manual curation.

Protein List

The **Protein List** tool assembles summarized information on multiple proteins at once. This is useful for comparing proteins of interest. The **Protein List** tool collects most of the information gathered by the **Protein Profiler** tool for each protein entered in the query box and, when possible, links to the source of the information directly.

To search for information on multiple proteins, simply enter HGNC approved symbols separated by commas (",") or tabs.

BELish

Molecular relationships stored inside the Knowledge Graph are manually curated as **BEL** statements prior to import.
A basic BEL statement has the following structure:
FUNCTION(NAMESPACE:
VALUE) RELATION
FUNCTION(NAMESPACE:
VALUE)

For more information on the parts of a BEL statement, and their options, see the
Defining A BEL Statement section below.

This tool allows users to build custom BEL statements and search the Knowledge Graph for any relationships that match. BEL Statements are composed of a *subject*, *relation*, and an *object* with the *subject* and *object* being represented as nodes in the Knowledge Graph and the *relation* as the edges.

Relations

When choosing a *relation*, users can be as general or as specific as they choose. The BEL relation types encoded in our database were designed using a hierarchical schema, therefore, each relation class inherits from a parent class. Thus, users can search for a class of relations instead of a specific instance (e.g. *causal* relations instead of *increases*) which will increases the number of matching molecular relationships in the results. Relations in the drop down box are ordered by their inheritance schema as indicated by the indentation level, with the more-indented child classes inheriting from their above, less-indented parent classes. Below is an example:

- causal (parent class)
    - causes_no_change (child class)
    - decreases (child class)
    …
- compiler (parent class)
  …

Subjects and Objects

The *Subject* and *Object* boxes represent the physical nodes in the knowledge graph.
Because users may want to a more generalized search, or they may not know the exact parameters (i.e. function, namespace, or value), our tools enables one to use wildcards in their search. This means that users can substitute any term for a wildcard
(represented as a question mark: "**?**") and the BELish query tool will search for relationships that best fit.

Examples

Protein-Protein Interactions (PPIs)

Subject	Relation	Object
p(?)	causal	p(?)

Human Only PPIs

Subject	Relation	Object
p(HGNC:?)	causal	p(HGNC:?)

Associated Diseases

Subject	Relation	Object
path(?)	association	path(?)

Defining a BEL Statement

Functions

The **FUNCTIONS** parameter defines the type of node. BEL is capable of representing both
concrete entities (genes, RNA, proteins, etc.) as well as abstract concepts (biological processes, patholgoies, etc.)
Below is a table of commonly used **FUNCTION**s that users can reference
when building *Subject* and *Object* nodes. When building the BEL term, be sure to use the shortcode in the `BEL Shortcode` column.

Node Type	BEL Shortcode	Description
gene	g	Represents a gene symbol.
RNA	r	Represents mRNA.
microRNA	m	Represents a microRNA.
protein	p	Represents a protein.
complex	complex	A physically bound complex consisting of two or more proteins, RNA, genes, or abundances.
composite	composite	List of multiple abundances synergize to produce an effect.
abundance	a	Can represent small molecules, chemicals, cell populations, tissues, organs, and other biological components.
biological process	bp	A defined biological process or molecular pathway.
pathology	path	Disease or pathology.

More information on BEL functions can be found at the BEL language website.

Namespaces

Namespaces are a controlled, standardized, collection of values that belong to a particular category, ontology, or data source. BEL uses namespaces in order to control and regulate how nodes are defined. Users can use the below list of namespaces that are most often used during BEL curation. The list is divided into categories since specific BEL functions tend to use a particular collection of namespaces:

Genes, RNA, Proteins	Biological Processes	Pathologies
HGNC	MESH	MESH
UNIPROT	MESHBP	MESHBP
MGI	GO	GO
RGD	GOBP	GOBP
		DO
		HP

Values

Values are the individual terms, phrases, symbols, or biological entities that the user may search for. Values differ from namespace to namespace, with values from HGNC being comprised of gene symbols while values from the Disease Ontology (DO) are clinically defined diseases and disease subtypes.

For a list of values which compile each namespace, please visit our BEL Namespace repository.

Path

Path is a tool designed for visualizing subgraphs within the Knowledge Graph. Users can use various methods to generate images of all of the *causal* relations between BEL nodes including: start and end points, last author, and information from publications. The resulting image is interactive, with users being able to click on the individual edges in the generated image which will bring up additional information on that particular relation including its provenance and, if applicable, a comparison to any uploaded differential gene expression data.

The different visualization options can be accessed by their individually labelled tabs:

- Path
- Last Author
- Publication

Path Tab

The **Path Tab** allows one to identify all *causal* pathways between a start and end point defined in the `From` and `To` boxes, respectively. The terms entered into these boxes must be values from an included namespace as described in the Values section above. The `From` and `To` values can be from any namespace, but this tool will only compile an image of causal relations between the two.

The `Depth` parameter controls how much separation should exist between the queried values. Users can set a minimum and maximum depth for thier pathway search, with "1-1" indicating that only pathways with a single edge between the start and end points should be considered, while "1-3" expands the search by allowing all pathways with either 1, 2, or 3 edges between the queried terms. The maximum depth is limited and depends on the size of the Knowledge Graph itself, with large Knowledge Graphs having a lower maximum. This is to ensure that **Path** queries do not crash the system.

Last Author Tab

This tab allows one to generate an image of *causal* relations which were curated from publications by a given last author. Users must enter the last author's name as it appears in the Pubmed publication (e.g. last name and first initial such as "Einstein A") into the `Last Author` box. Additionally, the resulting image may be filtered so that it only contains particular BEL relations as selected from the `Relation` drop-down box. The image can be further filtered for particular node types by selecting a biological entity from the `Bio. Entity` box.

Publication Tab

There are instances in which a user may want to visualize all of the BEL relations contained in the Knowledge Graph which were derived from a particular publication or all publications from a specific journal or year. This tab gives users the power to generate *causal* relation models using specific publication information:

- `Journal` : The name of a scientific journal. Image will contain all *causal* BEL relations from specified journal.
- `Title` : The title of a scientific publication.
- `Year` : The publication year of the scientific publication(s). Can be used to filter the image.
- `Last Author` : Similar to the Last Author tab.
- `PMID` : The PubMed ID of a scientific publication.
- `BEL Relation Type` : Filter used to control which types of BEL relations are included in the pathway image.