

## IJS internal report 12235

---

# Upgrading the Biomine heterogeneous network

Jožef Stefan Institute

Version 1.1 (final)

---

### Abstract:

We present the latest development of the Biomine system, its heterogeneous networks and the Biomine web application. In total, 3 heterogeneous Biomine networks providing data about 12 organisms have been constructed by merging the data from 10 databases that are available online. The largest, most general Biomine network consists of more than 1.3 million nodes and more than 26 million edges. The latest update was performed on the 1 July 2016 and the resulting databases were integrated into the Biomine web application.

---

Document administrative information	
Project acronym:	HinLife
Project number:	J7-7303
Deliverable number:	D1.1a
Deliverable full title:	BioMine update at M12
Document identifier:	IJS delovno poročilo 12235
Lead partner short name:	Jožef Stefan Institute
Report version:	1.1 (final)
Report preparation date:	23/12/2016
Lead author:	Vid Podpečan
Co-authors:	Dragana Miljković, Nada Lavrač
Status:	Final

## Introduction

The Task 1.1 of WP1 is focused on data preprocessing and transformation into a heterogeneous information network. More specifically, we are focused on updating the Biomine system, its heterogeneous networks, and developing an application which enables the exploration of these networks. The Biomine system consists of several parts the most important of which are:

1. the data preparation subsystem (data downloaders, data parsers, data importers),
2. the database and the cache server,
3. the graph search subsystem, and
4. user interfaces.

The results of this task upgrade the contents of the Biomine database and the cache server by providing the latest updates of its data sources, extend and improve user interfaces by developing a web application, and provide maintenance for various parts of the data preparation subsystem (e.g. adaptation to newer data formats, providing export of Biomine networks, etc.)

In the following we present the updated Biomine system and the constructed heterogeneous networks while giving examples how the updated heterogeneous networks can be explored by the Biomine search engine and the network visualization interface.

## Methodology

The Biomine project [1,2] develops methods for the analysis of public biological data sources (annotated sequences, proteins, domains, and orthology groups, genes and gene expressions, gene and protein interactions, scientific articles, and ontologies). All information is handled as graph and Biomine provides probabilistic graph search algorithms to automatically extract the most relevant subgraphs. Biomine was conceived at the University of Helsinki [2] and developed there until 2012 [3]. Since 2013 the development continues at the Jožef Stefan Institute.

Biomine system is implemented as a large collection of software components which perform various tasks. Several UNIX operating system tools and languages are also essential, e.g., *sed*, *Awk* and *Bash* are used when downloading, parsing and transforming data. A block diagram with the most important components is outlined in Figure 1. In the following we describe the underlying data model and the key components of Biomine.

## Data model

The data stored in the Biomine system can be described formally as a directed, labelled, weighted graph

$$G=(V, E, p) \quad (1)$$

where  $V$  is a set of nodes,  $E$  is a set of edges and  $p$  is a function  $p:E \rightarrow [0,1]$  that associates a probability  $p(e)$  to each edge.

Nodes represent biological entities of different types:

- Article
- Protein
- Gene
- Homology group
- Biological process
- Gene variant
- Gene location (Locus)
- Protein family
- Molecular function
- Drug
- Protein feature
- Phenotype
- Enzyme
- Cellular component
- Pathway
- Tissue

Edges represent relations between nodes (biological entities) and the most important types are:

- Protein is associated to Protein
- Article refers to Node
- Node has annotation GO
- Protein contains Feature
- Gene is homologous to Gene
- Gene codes for Protein
- Gene is located in Locus
- Protein belongs to Family
- Protein interacts with Protein
- OMIM refers to OMIM
- Gene participates in Pathway
- GO has parent GO
- InterPro has parent InterPro
- Compound participates in Pathway
- Gene affects Phenotype

- Phenotype is mapped to Locus

Because not all relations are equally important, edges are weighted by computing the value of the function  $p(e)$  for each edge. Function  $p(e)$  is defined as a product of three factors:

$$p(e) = \min(q(e) * i(e) * r(e), 1) \quad (2)$$

where  $q$  represents relevance,  $i$  represents informativeness and  $r$  represents reliability. According to Equation 2 the weight can be interpreted as the probability that the edge  $e$  represents existing, relevant and informative relationship [1]. A detailed procedure how to compute  $p(e)$  is described by Eronen et al. in [1] and [2].

## Key components

As outlined in the Introduction, there are four key subsystems forming the backbone of Biomine. In Figure 1 the connections between these subsystems and the rest of Biomine components are shown. In the following we provide a compact description of each subsystem.

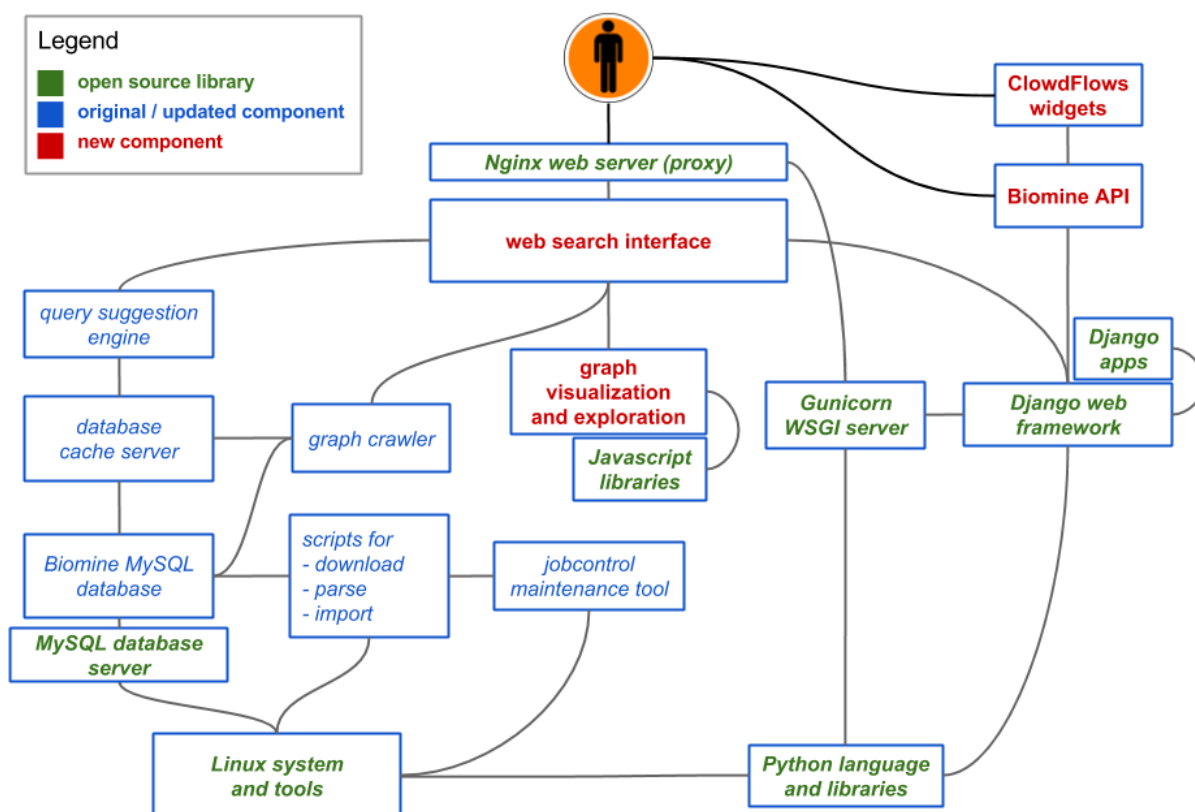


Figure 1: A block diagram of Biomine components and software libraries and programs.

### *Data preparation subsystem*

The data preparation subsystem consists of a large collection of scripts in various languages (mainly Bash, Awk and Python) which provide data collection, transformation and import for each of the supported data sources. The process of refreshing the main Biomine database is a three step process. First, data sources are checked in order to see if a new version is available. When all updates are downloaded, the data is parsed, extracted and transformed into an intermediate format. In the last stage, the prepared data is imported into the MySQL database.

In order to simplify this process and to provide an abstraction layer for easy handling of several data source a tool called *jobcontrol* has been developed. It is a top level tool which aggregates several commands and also manages versioning of the processed data.

### *Database and cache server*

The data extracted from external data sources by the data preparation subsystem is stored in as a relational database in MySQL RDBMS. Biomine's heterogeneous networks are stored in several large interlinked tables which hold information about nodes, edges, weights, labels and other relevant data.

Because a relational database is not well suited for queries which arise in graph mining algorithms, the key data describing structures of the networks is duplicated in a custom, highly optimized cache server. This enables high performance of the Biomine query engine and enables the real time auto-suggest feature in the Biomine web application.

### *Graph search subsystem*

The graph search subsystem is the core of Biomine. It implements a crawler which takes into account the user query, performs a search in the selected network and returns a subnetwork which was estimated as the best answer to the query.

The crawler traverses the selected network while taking into account the computed link weights. Three types of search are supported:

1. neighbourhood search,
2. connection search (source-target), and
3. connection search on a single set.

The crawler can take into account the desired size of the resulting network. However, this is a *soft constraint* because there is no guarantee that the result can have exactly the specified size.

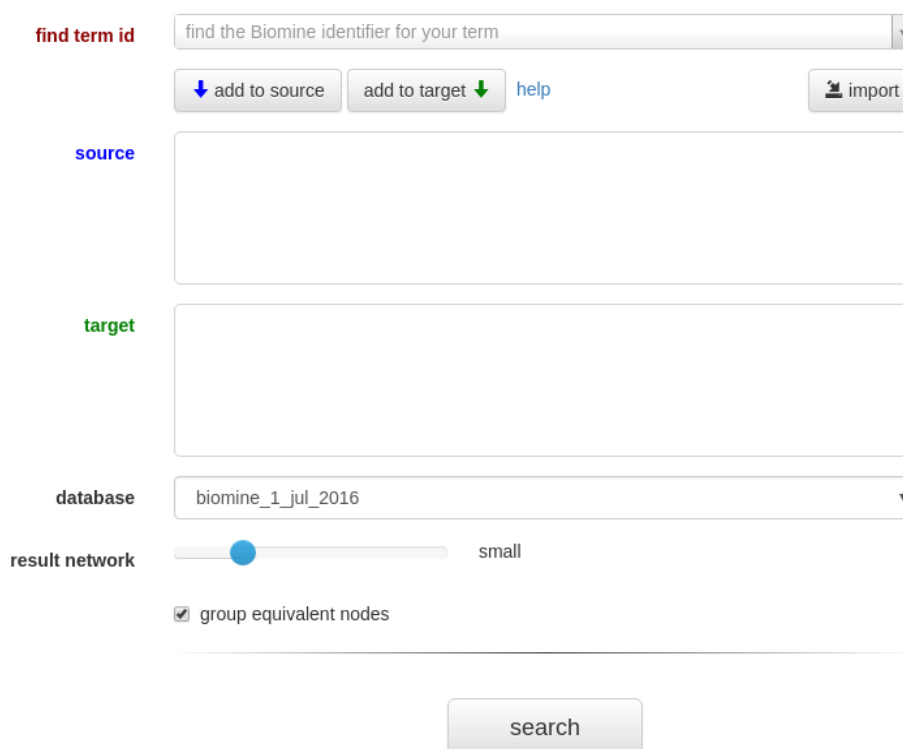
### *User interfaces*

Because Biomine is a complex system it is essential that it should provide an easy-to-use interface which abstracts the underlying complexity. We have implemented two user interfaces:

1. a web application with interactive network visualization and
2. an API which enables programmatic access.

The web application offers an easy-to-use search form and supports interactive network visualization. Figure 2 shows the search form of the Biomine web application. All three types of search that are supported by Biomine search subsystem can be invoked using the search form by entering data into the source query set, target query set or both. The database can be selected (only the latest databases are available) as well as the desired size of the result (as mentioned above this is a soft constraint). In addition, an option to group equivalent nodes is also available. This can reduce the size of the result network by grouping nodes that are of the same type and have the same connection pattern. For example, if there are 20 nodes representing articles which all refer to the same gene they can be grouped into one node thus reducing the complexity of the network without changing its semantics.

The Biomine API enables programmatic access to the Biomine search subsystem. It is based on JSON protocol and offers two functions: listing of databases and search.



The screenshot shows the search form of the Biomine web application. It includes a 'find term id' input field with a placeholder 'find the Biomine identifier for your term'. Below this are three buttons: 'add to source' (with a blue arrow), 'add to target' (with a green arrow), and 'help'. There is also an 'import' button with a document icon. The 'source' and 'target' fields are large empty text boxes. The 'database' field is a dropdown menu currently showing 'biomine\_1\_jul\_2016'. The 'result network' field has a slider set to 'small'. At the bottom, there is a checked checkbox for 'group equivalent nodes' and a 'search' button.

Figure 2: The search form of the Biomine web application.

The API should always be used so that the following two steps are carried out:

1. call the **list\_databases** function to get a list of available Biomine databases,
2. call the **api** function with the selected database and source and optionally target query terms (few other optional parameters can also be adjusted).

Upon the successful invocation the API returns a network encoded in the JSON format.

## Results

### Biomine's heterogeneous networks

We have extended and updated Biomine to support in total 12 organisms (of which 7 are plants):

1. human (*Homo sapiens*)
2. mouse (*Mus musculus*)
3. rat (*Rattus norvegicus*)
4. fruit fly (*Drosophila melanogaster*)
5. nematode (*Caenorhabditis elegans*)
6. arabidopsis (*Arabidopsis thaliana*)
7. potato (*Solanum tuberosum*)
8. rice (*Oryza sativa*)
9. tomato (*Solanum lycopersicum*)
10. tobacco (*Nicotiana tabacum*)
11. beet (*Beta vulgaris*)
12. cacao tree (*Theobroma cacao*)

These organism are grouped into three *organism sets*:

- **biomine** (1-5),
- **human** (5), and
- **plants** (6-11).

*Organism set* is a Biomine termin denoting organisms which data is merged into a single database (network). Note that the homo sapiens (human) is included in the *biomine* organism set but is also available as a separate organism set. The reason for this redundancy is that a separate network containing only human-related data is of great value to biology experts when searching human-specific information.

Each organism set has a corresponding database (network) which contains nodes, links, and other related data. Biomine networks are constructed using the following public data sources: UniProt, GO, EntrezGen, MIM, InterPro, SwissProt, Trembl, HomoloGene, STRING, PubMed, and GoMapMan. As of the latest update, three databases (networks) exist. Table 1 shows their names and sizes.

network name	number of nodes	number of links
biomine_1_jul_2016	1.371.592	26.291.591
human_1_jul_2016	770.414	6.931.319
plants_1_jul_2016	897.651	2.368.585

Table 1: The latest Biomine networks and their sizes.

## Experiments

Suppose we are interested in the famous tumour protein **p53** which has been described as the guardian of the genome because of its role in conserving stability by preventing genome mutation [4,5,6]. It is crucial in multicellular organisms, where it prevents cancer formation and thus functions as a tumour suppressor. In humans, this protein is encoded by the *TP53* gene. We will demonstrate how Biomine can help us collect and presents knowledge about the *TP53* gene in the form of a compact biological network.

We will begin the experiment by opening the URL <http://biomine.ijs.si/> in the browser. Default settings will be used:

- database: *biomine\_1\_jul\_2016* (the largest, most general database)
- result network size: small
- grouping: ON
- type of search: neighbourhood

When entering the keyword "TP53" the search interface offers many related entities among which we select "TP53: tumour protein TP53" that has the unique internal identifier EntrezGene:7157 and can be found in the category "Gene (human)".

By pressing the "add to source" button the search term is created and added to the source query set. Since we want to perform a neighbourhood search the target query set will remain empty. The search is started by pressing the "search" button. The preview of the results will appear on the right hand side. To launch the advanced network exploration interface we can click the "advanced view" button which opens the Biomine network explorer in a new tab. The resulting network is shown in Figure 3.

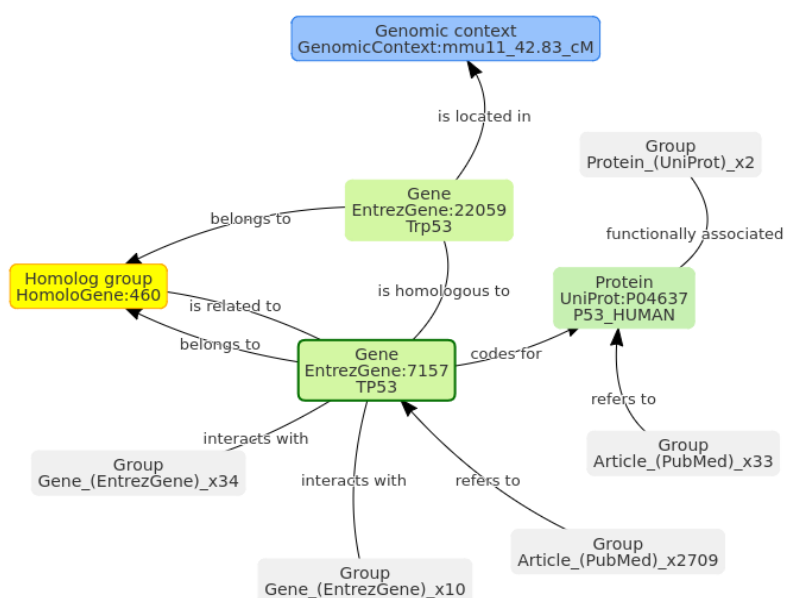


Figure 3: A result of Biomine neighbourhood search for the *TP53* gene using default search settings.



The network in Figure 3 reveals that gene TP53 is indeed well researched because a group node containing 2709 scientific articles refers to TP53. In addition, 33 articles refer to the P53 protein which is encoded by TP53. There are also a group of 34 genes which interact with TP53 and a mouse gene TRP53 which is its homologue.

Increasing the desired result network size yields a larger network which is shown in Figure 4. It contains significantly more information but the interpretation of which requires a biology expert.

To demonstrate the usefulness of the node grouping feature we also include a picture of the network from Figure 4 with the grouping feature turned off. The result is shown in Figure 5. Note that the absence of grouping expands the article group node to 2709 nodes and hopelessly clutters the network.

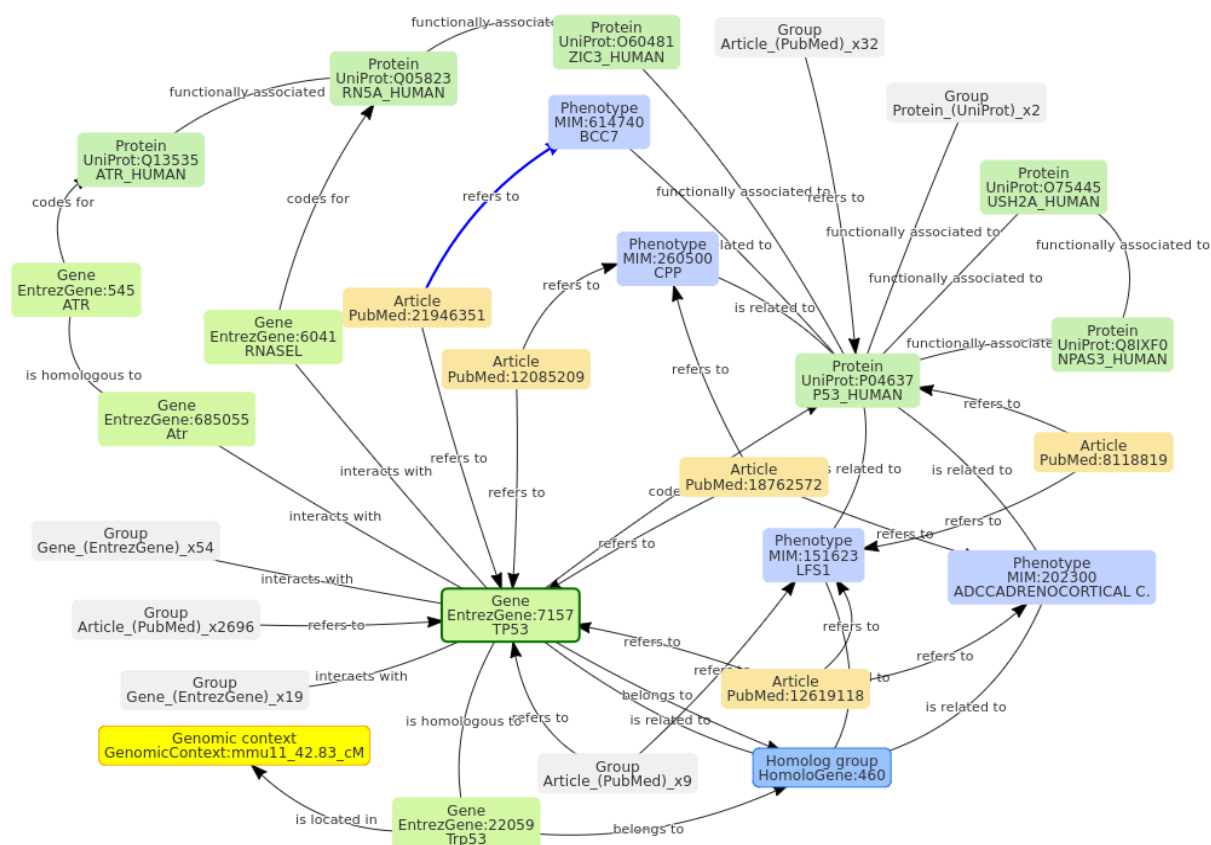
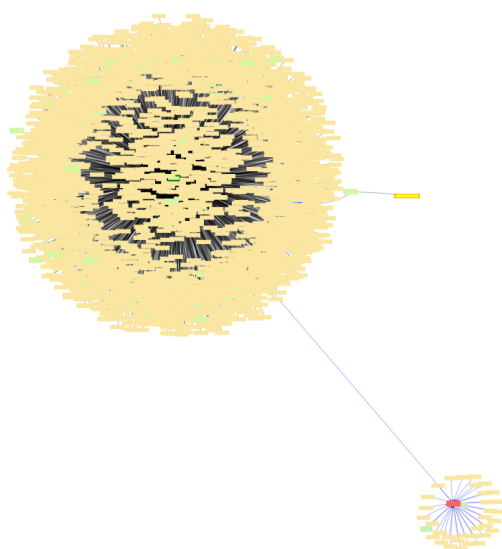


Figure 4: A result of Biomine neighbourhood search for the TP53 gene using default search settings while the desired network size is increased by one step.



*Figure 5: A result of Biomine neighbourhood search for the TP53 gene using default search settings but with the desired network size increased by one step and the grouping feature turned off. The yellow nodes represent scientific articles.*

## Conclusion

We have described the updated Biomine system and its heterogeneous networks. An overview of Biomine's system architecture was given and we have presented its four major components. A short summary of Biomine's graph-based data model was given and we have listed the supported node and link types.

We have summarized the updated heterogeneous networks, identified their sources and exact sizes. To demonstrate the updated system an experiment was performed where we searched the most general Biomine network for the famous TP53 gene. We have demonstrated the effect of the selected search parameters and presented a basic interpretation of the results.

## References

1. Eronen, L. and Toivonen, H. (2012) Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13 (1), 1–21.
2. Sevon P, Eronen L, Hintsanen P, Kulovesi K, Toivonen H: Link Discovery in Graphs Derived from Biological Databases. In *DILS*, Volume 4075 of *Lecture Notes in Computer Science*. Edited by: Leser U, Naumann F, Eckman BA.

3. Podpečan, V., Lavrač, N., Mozetič, I., Kralj Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K. SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics* 12, 416 (2011)
4. Wikipedia: TP53. <https://en.wikipedia.org/wiki/TP53>
5. Surget S, Khoury MP, Bourdon JC (19 December 2013). "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective". *OncoTargets and Therapy*. 7: 57–68.
6. Read, A. P.; Strachan, T.. *Human molecular genetics 2*. New York: Wiley; 1999. Chapter 18: Cancer Genetics.