SeqGrapheR Manual

Petr Novak (petr@umbr.cas.cz)

September 3, 2012

Contents

1 Introduction

The SeqGrapheR is a tool which is designed to complement RepeatExplorer, a computational pipeline for discovery and characterization of repetitive sequences in eukaryotic genomes(http://repeatexplorer.umbr.cas.cz) RepeatExplorer uses high-throughput genome sequencing data as an input and performs graph-based clustering analysis of sequence read similarities to identify repetitive elements within analyzed samples. Graph representation of repeats can be interactivelly visualized and analyzed using SeqGrapheR. The principle of this approach is decribed in Novak et al. (2010).

2 Installing SeqGrapheR

SeqGrapheR requires installation of the program GGobi (www.ggobi.org) and for full functionality it is necessary to install NCBI blast with programs blastall and megablast in the path. NCBI blast can be obtained from . The R packages gWidgets, Rgtk2, igraph, rggobi, gWidgetsRGtk2 and cairoDevice can be installed from CRAN with the command as

```
> install.packages('igraph', dep=TRUE)
```

Additionally, the required package Biostrings can be obtained from Bioconductor website (). To install the package start R and enter:

```
>source("http://bioconductor.org/biocLite.R")
>biocLite("Biostrings")
```

When all required libraries are installed in R, SeqGrapheR can be installed with command:

```
>install.packages('SeqGrapheR~0~.4.1.tar.gz',repos=NULL)
```

OS compatibility: Linux -recommended, Windows will require the installation of gtk libraries, functionality in MacOS was not tested.

3 Input files

As mentioned above, SeqGrapheR is intended to be used for interactive analysis of output from RepearExplorer pipeline. Several files generated by RepeatExplorer can be loaded to SeqGrapheR:

- `CLXY.GL` graph file in GL, this file must be loaded first
- `ACE_CLXY.ace` file with contig assembly
- `CL00XY_blastx.csv` results with simmilarity search against protein domain database
- `reads.fas` sequences in fasta format
- `interconnected.txt` is file multiple Ids which has similarity to other clusters

Location of these files is described in the RepeatExplorer manual.

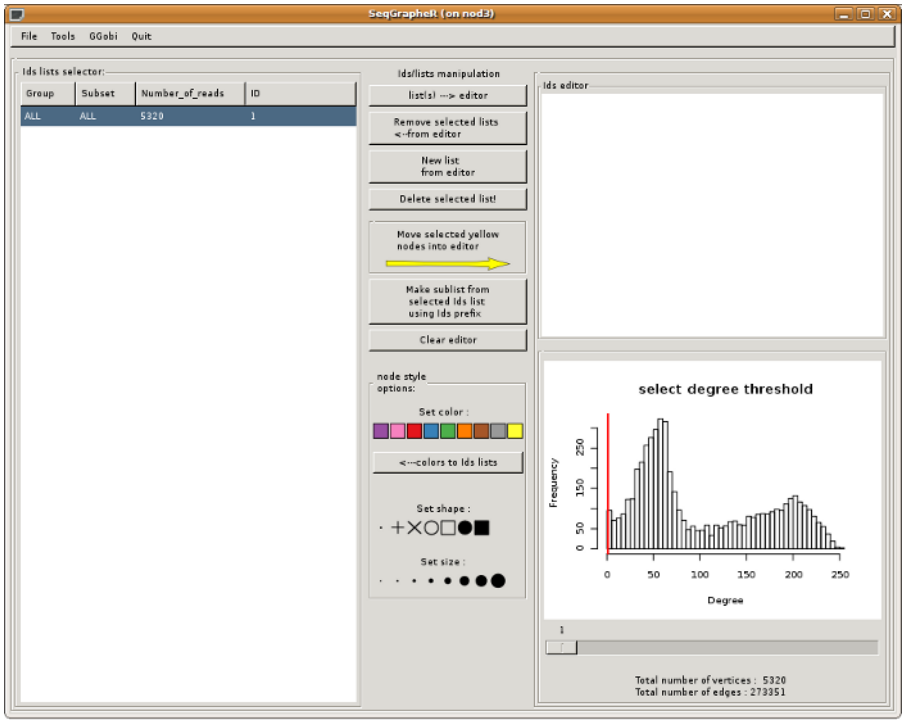

Figure 1: SeqGrapheR main window

4 Starting SeqGrapheR

To run SeqGrapheR GUI start R and enter:

```
>library(SeqGrapheR)
>SeqGrapheR()
```

It is possible to issue the SeqGrapheR() command multiple times and run independent instances of GUI. To start with, a graph must be imported first from the SeqGrapheR main window(Figure 1)

Currently there are four possibilities of how to import a graph. First, the graph can be imported from GL format. This file is generated during clustering processe by Repeat-Explorer pipeline. GL format is actually an external representation of an R object GL where GL*Gisanigraphobjectand*GLL is a matrix with layout coordinates calculated by the Fruchterman-Reingold algorithm. The second supported format is the ncol format. The .ncol edge file is a simple 2 (or 3) columns file where two vertices are on each line separated by tabulator:

```
vertex1name vertex2name [optionalWeight]
vertex1name vertex3name [optionalWeight]
...
```

When the .ncol file is loaded, the graph layout is calculated. This can take some time, from minutes for graphs with hundreds of vertices to couple of hours from graphs with 20,000 vertices. The progress of the layout calculation is show in the terminal. The third option is to load sequences in fasta format. In that case, megablast is run first for all-to-all comparison to find all edges of the graph, then the graph layout is calculated. The fourth option is to load a graph as the whole project. This option will also import additional project-associated information like lists of Ids, contig information, or results of a similarity search. When a graph is loaded, the histogram of degree should appear in the right bottom corner and the graph in the ggobi window will pop up. Also, the first Ids list is created from all sequence identifiers. The graph is shown in separate scatterplot window of GGobi program(Figure 2)

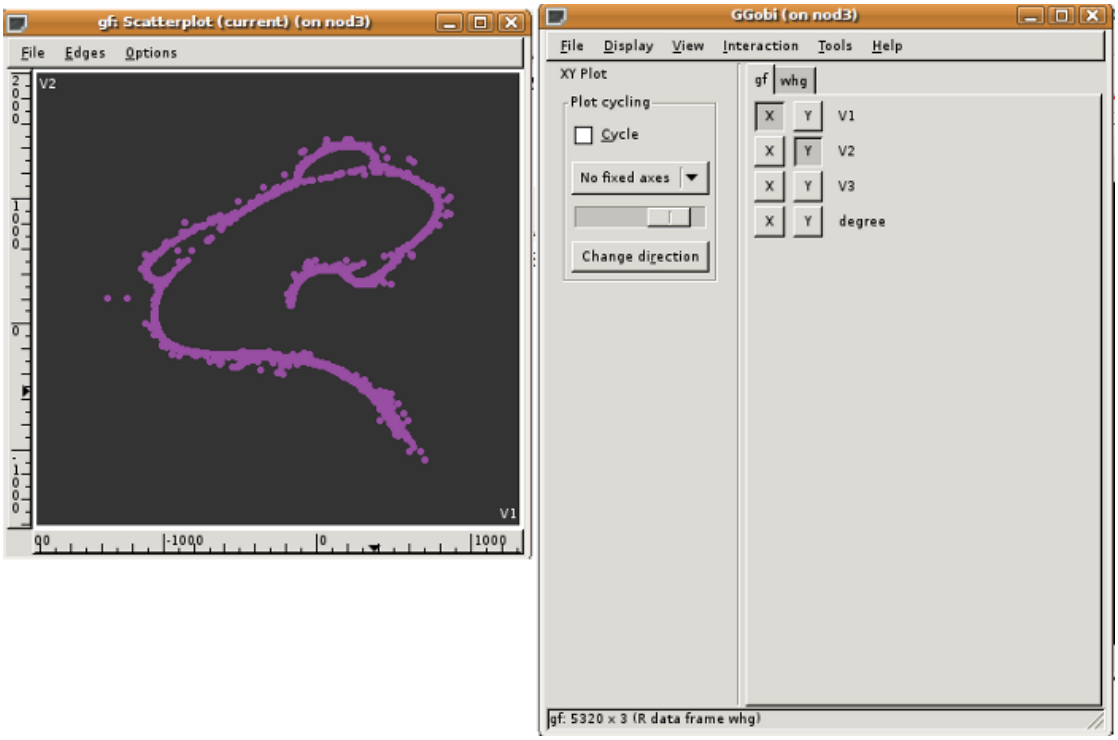

Figure 2: GGobi window

By default, only nodes without vertices are shown. Hiding the edges significantly speeds up the rendering and the manipulation of the graph (in extreme cases GGobi can crash during attempts to display all edges). To see or hide the edges, one can use options *Edges/Show lines only* or *Edges/hide edges* in the graph window. The initial graph scatterplot shows only the first two dimensions. It is possible to interactively change the point of view by switching into 3D from the main GGobi window using the *View/Rotation* option. There are different modes of interaction with the graph which can be changed from main GGobi window. When 3D view is on (rotation) the possible interactions are *Rotation* - enable to change the point of view, *Brush* is used for the labeling group of nodes .*Identify* is used for identification of individual points (reads). More information can be found in the GGobi manual. Yellow color is used for selection. While it is possible to change the color scheme, it is not recommended for correct interaction of SeqGrapher and GGobi.

5 Ids list selector

Ids list selector is the table on the left side. This table contains all created or imported lists of sequence names (Ids). A list of Ids can be imported either from a file or by manual entry into the Ids editor. To import the list of Ids from a file, use the option *File/Import/Single Ids list to editor* which can upload a file with ids separated by spaces, tabs, or commas. Only valid Ids are imported. Another option is to import multiple Ids list by *File/Import/Multiple Ids list*. Multiple lists must be in format:

```
>groupname1 Subsetname1 Number~of~Ids
Ids1 Ids2 Ids3 Ids4 Id5...
>groupname2 Subsetname2 Number~of~Ids
Ids12 Ids22 Ids32 Ids42 Id52...
```

It is possible to use an alternative format where each Id is associated with frequency; in this case every Id must be followed by the frequency (integer):

```
>groupname1 Subsetname1 Number~of~Ids1
Ids1 Freq1 Ids2 Freq2 Ids3 Freq3.....
>groupname2 Subsetname2 Number~of~Ids2
Ids21 Freq21 Ids22 Freq22 Ids23 Freq22.....
```

For a list with associated frequencies, an histogram of frequencies can be shown by selecting a single list in the Ids list selector and then using the option from the *Tools/Plot/Frequency of Ids* menu from selected list. To create an Ids list based on a contig assembly, an ACE file containing the sequence assembly information can be loaded. Multiple lists will be created, each list containing an Ids name for each contig. Every Ids lists can be marked by a different color/shape/size in the graph simply by selecting the row(s) with the Ids list and clicking on the *Node style options* to select the desired style (Figure 3).For fast exploration, double click on the Ids list in *Ids list selector* will highlight specified list by yellow color.

6 Manipulation with Ids lists

Buttons in the upper middle panel can be used for creating and editing Ids lists (Figure 4).Ids can be moved from the Ids list table into the editor, edited manually, and multiple lists can be merged together.A list can also be created from the yellow nodes which were selected by a brushing, by using the histogram, or by another method.

Additionally, multiple lists based on the color of the graph can be created by clicking the button below the strip color selector <—*colors to Ids list*. Sequence Ids can be also used for creating prefix-specific sublists. With selected Ids list, click on the *Make sublist from selected Ids list using Ids prefix*, then you will be asked for the prefix length. Multiple subsets will be created; each subset with unique N-letter long prefix Ids. This option can be used for *comparative analysis* when multiple sequence sources were pooled (e.g. from different species) and were labeled with a subset-specific prefix. In example below,5 letter long species-specific prefixes were used to create four sublists (labeled as ALL prefix) to distinguish the species-specific part of the graph (Figure 5).

Figure 3: Node color and style selector

7 Histogram area

When a graph is loaded, a histogram showing the distribution of all node degrees will appear on the bottom left. The degree of a node is the number of edges which ends at that node. In the case of a graph derived from sequence reads, nodes with higher degrees are those which are similar to more sequences while reads with lower degree have less similar counterparts.A region in the graph with a higher degree thus represents a sequence which is in the genome in more copies or is more conserved. To identify such regions in the graph, a slider below the histogram can be used to select the threshold for coloring. All nodes above the threshold will be labeled by a yellow color(Figure 6). The figure bellow shows the an example of the graph derived from a LTR-retrotransposon where nodes with degree greater than 120 (yellow) correspond to the sequences derived from LTRs which are present in the genome in more copies than internal sequences. When sequences are also loaded, a histogram of the read lengths can be shown in histogram ar ea as well. To switch between histogram views, use *Tools/Plot/Read degree* or *Tools/Plot/Read length*.

8 Similarity search

To use this option, NCBI blast must be installed (blastall 2.2.18 was used for SeqGrapheR testing). To run a similarity search, sequences corresponding to the graph nodes must be imported first. To search for similarity to proteins or DNA use *Tools/Similarity search/Protein blastx* or *Tools/Similarity search/DNA blastn*. The SeqGrapheR window will be non-reactive until the search is finished. Results are first shown in the histogram window which show a histogram of

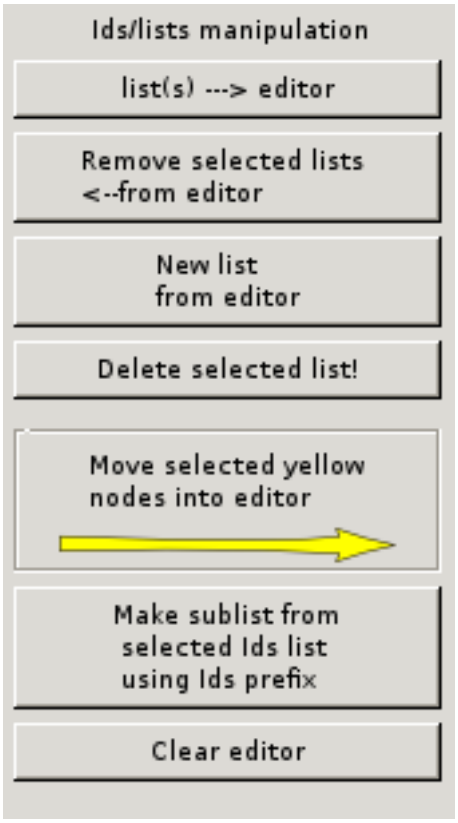

Figure 4: Manipulation with Ids lists

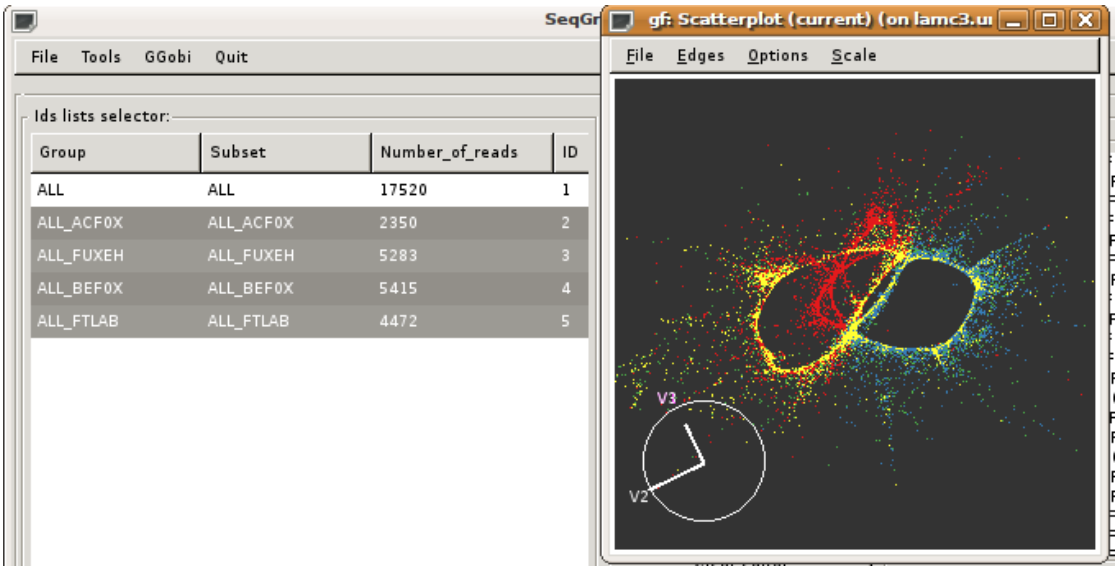

Figure 5: Example of separation of ALL Ids list into sublists based on the 5 letter long name prefix

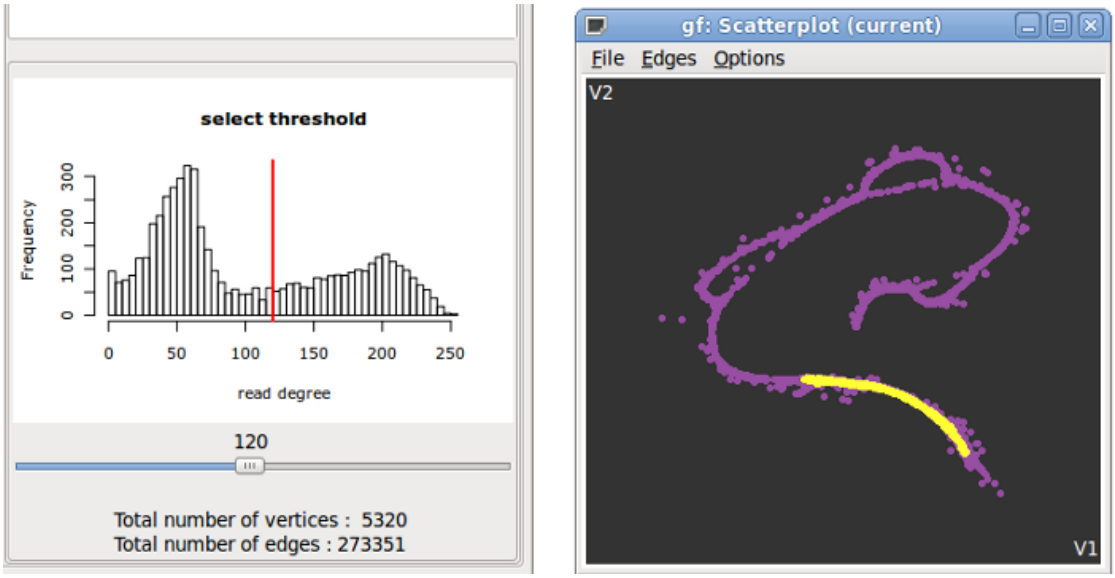

Figure 6: Example of selection of nodes with high degree using Histogram

the blast similarity bit-scores. Histogram with the last blast results can be also re-opened by Tools/Plot/Blast results. Example of Blast results in histogram window. All nodes with blast score greater or equal to the selected threshold (set here to 45) are highlighted by yellow color (Figure 7)

To see the table with blast hit information(Figure 8), use the slider to select a bit score threshold and press *show hits for selected nodes in the table* button below the histogram. Blast results in the table can be sorted by clicking on the corresponding column header. Selected rows can be used for an Ids list creation by moving them into Ids editor (*move selected to Ids editor* button).

A table with the blast results can be exported or imported using *File/Export/Blast results* or *File/Import/Blast results (tabular format)*. Similarity search against database or transposable elements protein domain is part of the output of RepeatExplorer pipeline (see RepeatExplorer manual). In table imported from RepeatExplorer pipeline, subject names in the blast table follow the format where type of sequence is specified by string before double underscore. In this case the button *Create Ids sublists from selected nodes* can be used to create protein domain specific Ids lists (Figure 9)

Currently only the last similarity search is stored in the memory. When a new search is run or new blast table imported, the previous results are overwritten.

9 GGobi window

When GGobi window is accidentally closed, it can be re-opened from the main SeqGrapheR window using *GGobi/Reopen*. When any GGobi window is still open this action will close all existing GGobi windows first and a new instance of GGobi will be created.

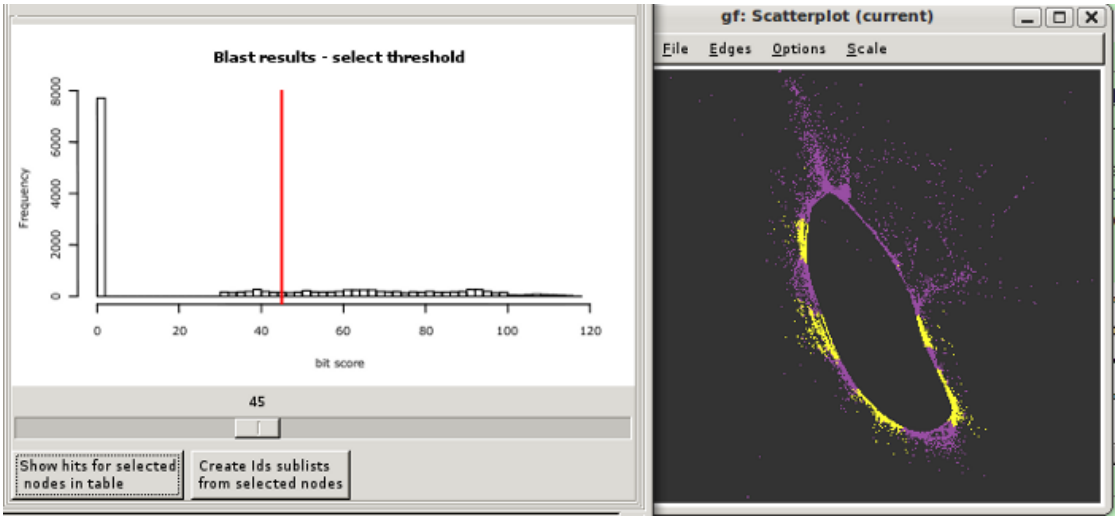

Figure 7: Histogram showing distribution of score blastx search. Reads with score above threshold (red line) are shown by yellow color on the left panel

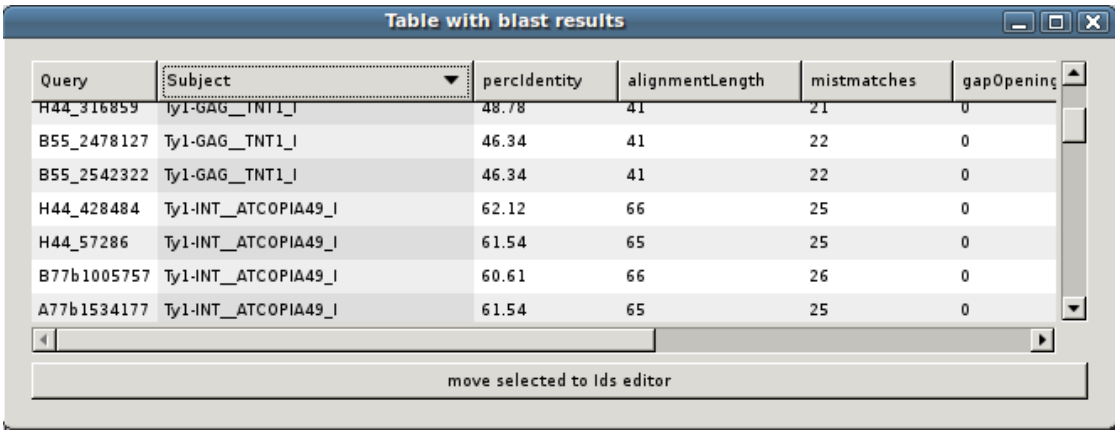

Figure 8: Output of blastx similarity search in tabular format

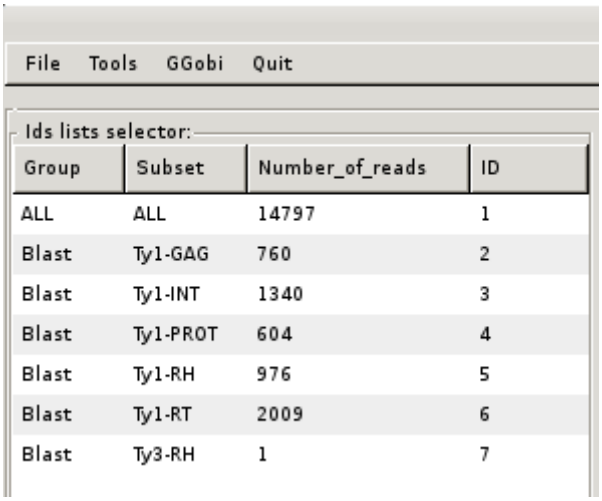

Figure 9: Ids lists created from blastx results based on the prefix specifying type of protein hits

10 Project

The current graph with all Ids lists, the last blast results, and sequences can be saved in one file as a project to work with later using *File/Project/Save and File/Project/Open*. Note that the current graph coloring will not be preserved. Current graph image can be saved into png file (*File/Export/Graph image*). The current projection of the graph as seen in the ggobi graph window can be saved as a new graph in a GL file. This file is the same as the original GL graph except the coordinates of the layout are correspondingly rotated.