XMLCreator

Version 1.0

June 1st, 2009

Hong Sun, Karen Lemmens, Tim Van den Bulcke, Kristof Engelen,
Bart De Moor and Kathleen Marchal
KU Leuven, Belgium

XMLCreator

The XMLCreator will use the input and output data of DISTILLER [1], together with the additional data to create the XML file and expression data file that are required for visualization by ViTraM.

Although the XMLCreator currently only includes the possibility to derive the XML file from the output and input of the module detection tool DISTILLER [1], more algorithms will be included in the future.

1. Introduction to DISTILLER

DISTILLER [1] is a module detection tool that identifies sets of genes that are co-expressed in a set of conditions (or modules) and the regulatory program of the genes in these modules. The regulatory program consists of regulators and/ or their corresponding regulatory motifs.

Although in the original publication, only expression and regulatory motif data were used, the data integration framework is very flexible for adding more data sources, for instance ChIP-chip data. The input data for DISTILLER [1] thus consists of expression data, usually in combination with another data source like motif screening data or regulator binding data. DISTILLER also requires that all data sources consist of the same number of genes and that genes are ordered in the same way in all data sets.

2. XMLCreator

2.1 Requirements

Developed in JAVA, the XMLCreator is platform-independent and is expected to work under other operating systems (Windows, Linux, Mac) that support the JRE (1.5 or higher) and with sufficient memory depending on the size of the input data.

Please direct comments and questions related to the software to kathleen.marchal@biw.kuleuven.be.

2.2 Installation of the software

The software can be downloaded from the download section on:

ftp:/ / ftp.esat.kuleuven.be/ sista/ klemmens/ ViTraM/ Index.html

After downloading the package, please follow these steps:

1. Unzip the downloaded file
2. Open the unzipped folder
3. Double click on the file XMLCreator .jar to run the software or open a command line window, and execute the command "java -jar –Xms64m –Xmx256M XMLCreator.jar" in the folder in which the files of the XML-Creator are stored.

If everything is OK, the XMLCreator should start right now and the following window appears (Figure 1):

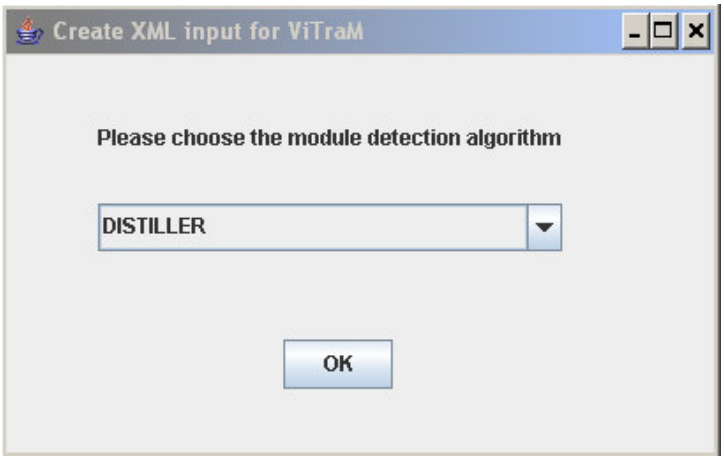

Figure 1: The interface of XMLCreator.

By choosing DISTILLER in the previous step, the following window will appear (Figure 2):

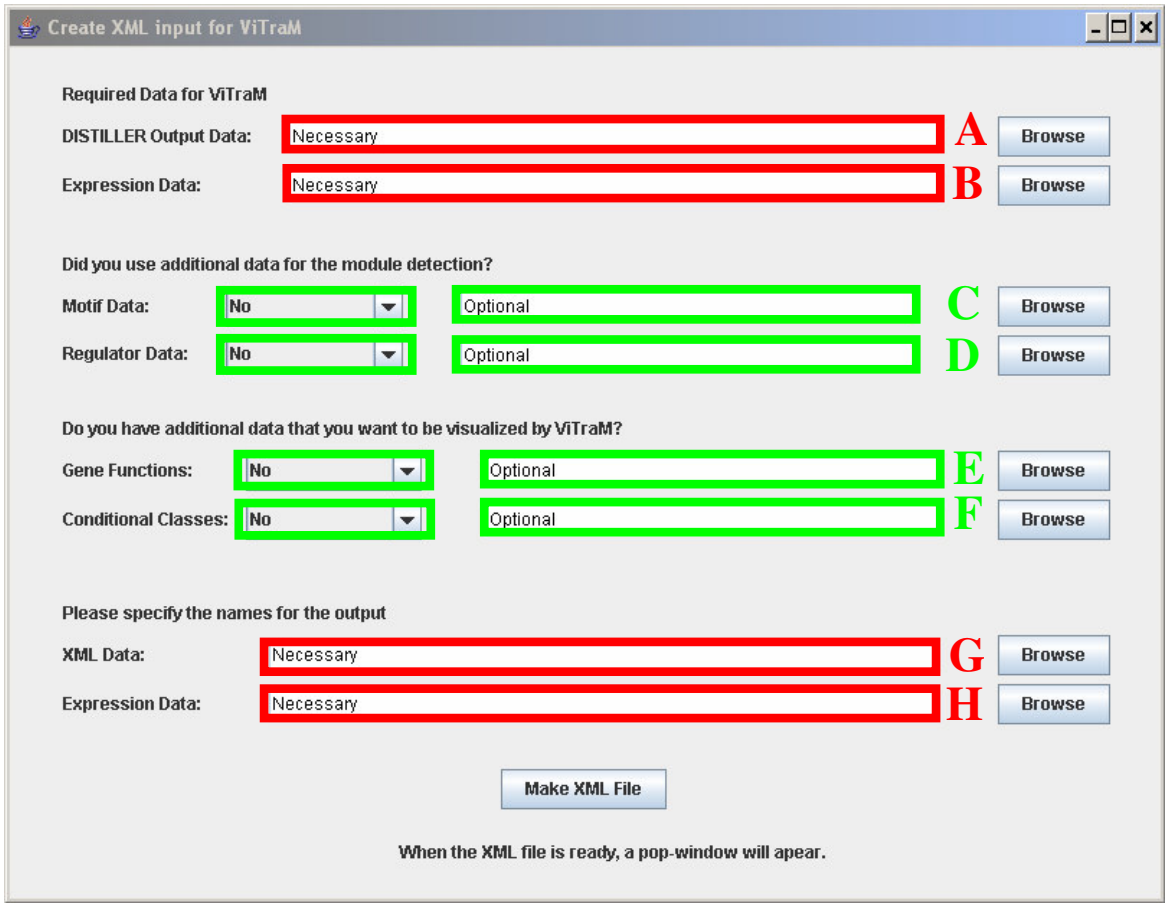

Figure 2: The XMLCreator software. The output (A) from DISTILLER and the expression data (B) used as input for DISTILLER are required for creating the XML file (G) and corresponding expression file (H) for ViTraM. There are also optional files (C, D, E, F) that can be specified for creating the output files.

When the data is created, a pop-window will be shown (Figure 3):

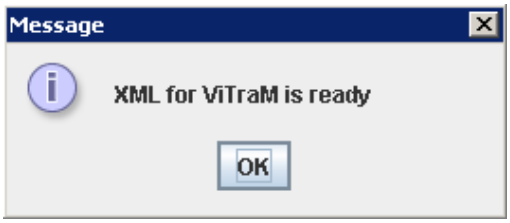

Figure 3: The pop-window will indicate that the files for ViTraM were created.

2.3 Data

The following data files are required for visualization by ViTraM and thus for generating the XML file that is required by ViTraM:

- DISTILLER output data: an .m file that contains information on the modules, i.e. which genes are co-expressed in which experiments. If not only expression data is available but also regulator and/ or motif data, DISTILLER will not only derive modules, but also information on the regulation of the module. Information on which regulators and/ or motifs are in control of which modules will then also be present in the output file (Figure 2.A).
- Expression data: The expression data contains the array names in the first row, while the first column consists of the gene names (Figure 2.B). This is the expression data that was used as input for DISTILLER.

If regulator and/ or motif data were used as input for DISTILLER, these data should also be provided to the XMLCreator. The order of the gene names must be the same as the ones in the expression data (as required by DISTILLER).

- Motif data: The motif data consists of motif names in the first row and gene names in the first column.
- Regulator data: The first row of the regulator data are the regulator names, whereas the first column contains the gene names.

In addition, the user should indicate which data file was used as input by DISTILLER (see Figure 4). It is possible to include only one additional data source, such that the user should indicate "Data source 1". If both motif and regulator data were included in the analysis by DISTILLER, the user has to indicate which data source is the first data source and which one is the second in the input of DISTILLER. Otherwise the output of DISTILLER can not be interpreted well. The gene, motif, regulator and experiment names that are used in these data sets will be displayed in the visualization by ViTraM and gene names should be used consistently in all files.

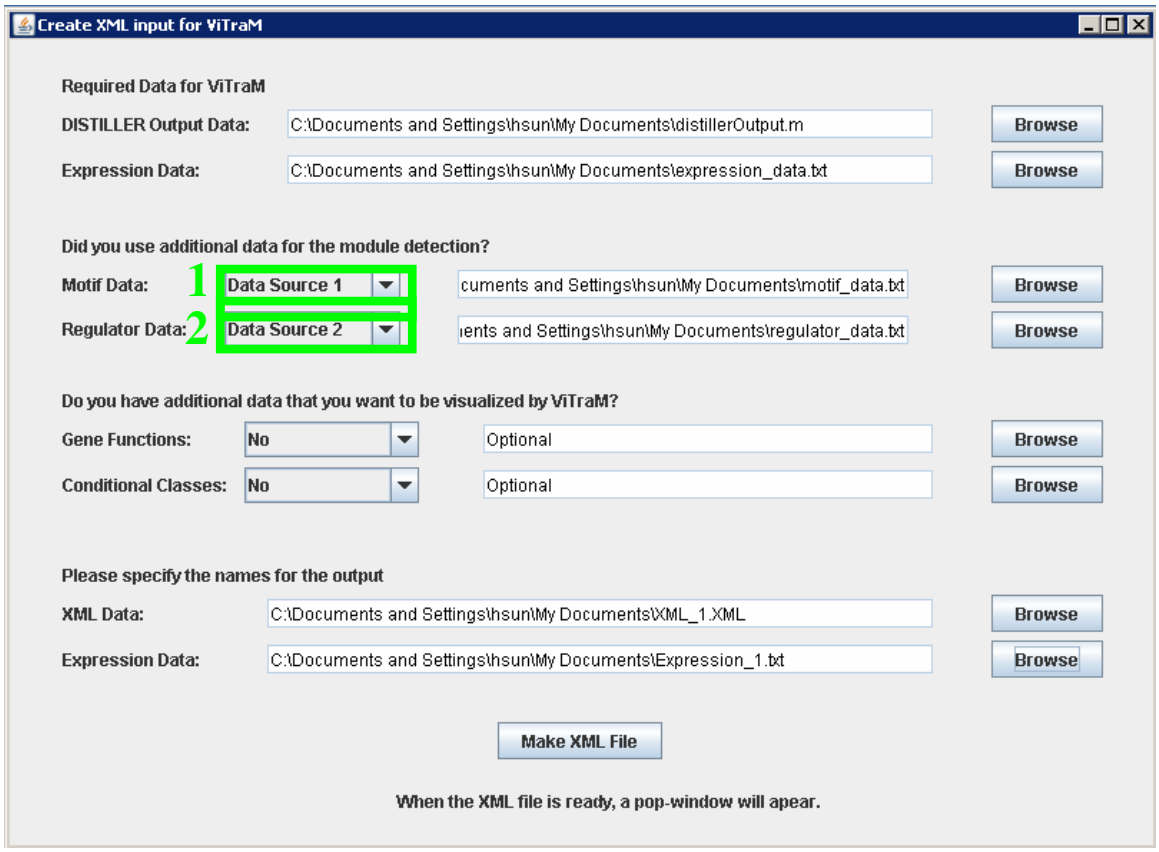

Figure 4: When you have two data sources, e.g. regulator and motif, it should be indicated which data source was used as first input for DISTILLER and which data source was used as second input for DISTILLER. As such the XMLCreator can interpret the output file of DISTILLER correctly.

Finally the user can also include additional data sets like information on the gene function or on the experiments. The latter files are however not required if not available.

- Gene function data: The gene function file consists of binary data indicating whether a gene is for instance member of a particular gene ontology category. The first row contains the gene names, whereas the first column contains the functional categories (see Figure 5).
- Conditional classes: A conditional class gives information on the major cue that was measured during the experiment. A similar file as for the gene function data can be derived for the conditional classes.

The structure of the gene function data

	DNA	TCA_cycle	aerobic	amino_acids	carbon_compounds
gene_1	0	0	0	1	0
gene_2	0	0	0	0	1
gene_3	0	0	1	0	0
gene_4	1	0	0	0	0
gene_5	0	1	0	0	0
gene_6	0	0	0	0	1
…					

Figure 5: A small example of the structure of the gene function data. The first column consists of the gene names; the first row consists of the gene functional classes. A "one" indicates that a gene belongs to a particular functional class.

REFERENCES:

[1] Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J, Engelen K, Marchal K., "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*". *Genome Biology* 2009, 10:R27.