TETRAMER: A Temporal Transcriptional regulation modeler to predict master regulators

Marco-Antonio Mendoza-Parra, Pierre-Etienne Cholley, Julien Moehlin, Alexia Rohmer, Vincent Zilliox and Hinrich Gronemeyer.

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Strasbourg; France.

1.0 INTRODUCTION

Describing living systems through the reconstitution of their genomic-regulatory functions stands for the biggest challenge of the current "big-data omics" era. Moreover, understanding the reorganization of such regulatory wires – as a consequence of external/internal cues – represents a new approach to interpret the acquisition of novel physiological or aberrant system states. In a cellular context, the detailed comprehension of these reorganizations, known as cell fate transitions, is a major component of the novel therapeutic developments in regenerative medicine.

Here we present TETRAMER, a cytoscape app providing a user-friendly framework for the reconstruction of cell fate transition-specific GRNs by integrating user-provided temporal transcriptomes with generic GRNs derived from (i) the analysis of multiple publicly available gene expression profiles associated to several mouse or human cell type/tissues (CellNet; http://cellnet.hms.harvard.edu/ *Cahan P. et al; Cell 2014*); (ii) the genome-wide mapping of human promoters and enhancers in multiple cell type/tissues from CAGE data generated by the FANTOM5 consortium (regulatory circuits; http://regulatorycircuits.org/ *Marbach D. et al; Nat. Methods 2016*); (iii) the systematic analysis of most publicly available ChIP-sequencing data corresponding to TF-binding in a variety of human or mouse cell type/tissues (http://ngs-qc.org/ *Mendoza-Parra MA et al; NAR 2013*).

Furthermore, TETRAMER provides an iterative approach for interrogating the capacity of each TF, retrieved on the GRN, to drive cell fate transformation. For it the temporal transcriptional regulation cascade derived from each TF is scrutinized as a way to verify its influence on the reconstitution of the differential gene expression patterns associated to the cell fate transformation.

TETRAMER has been initially designed for the study of retinoids-driven neuronal/endodermal cell fate decisions (*Mendoza-Parra MA et al; GenomeRes; 2016*), and has been further validated on studies concerning cell reprogramming; trans-differentiation as well as aberrant tumorigenic transformation. In all these cases, TETRAMER did not only identified several major TFs as master regulators of the cell fate transition under study, but in addition it provided a temporal inter-regulatory view among them, clearly demonstrating their temporal inter-transcriptional regulation interdependency.

Finally, TETRAMER has been also challenged for the generation of an atlas of candidate master regulators for the cellular reprogramming of more than 300 human cell lines. This effort allowed not only to confirm major TFs predicted by previous efforts (*D'alessio A. et al; Stem Cell Reports; 2015; Rackham O. et al; Nat. Genetics 20115*) but in addition it

provided new insights on their inter-regulatory relationship as explanatory for their predicted cell fate transformation capacity.

2.0 INSTALLATION

TETRAMER has been designed for Cytoscape version 3.4 and higher. The corresponding JAR file is available for download on this website.
There are two ways to install the tool:

- Open the "App Manager" window (Apps → App Manager) and use the "Install from File" function. The name of the plug-in will appear in the "Currently Installed" tab.

- Put the JAR file into the Cytoscape configuration folder:
  /home/"USER_FOLDER"/CytoscapeConfiguration/3/apps/installed/

In both cases, TETRAMER should be automatically available in Cytoscape; if it is not the case, feel free to restart Cytoscape.
When installed, a new tab "TETRAMER" will be visible in the Control Panel (left panel).

3.0 BUILD A GRN NETWORK WITH TETRAMER

While users can perform temporal transcriptional regulation propagation modeling with their own TF-TGs networks, TETRAMER can also build a network from temporal transcriptional information provided by the user. This is done by the use of generic TF-TGs relationships established by different means. Currently, TETRAMER allows to reconstruct temporal GRNs from:

(i) ***CellNET:*** TF-TG relationships assembled by Cahan *et al* from the analysis of more than 3 thousand publicly available gene expression profiles of diverse cell types and tissues (Cahan et al. 2014). (Mouse and Human collection available).

(ii) ***qcGenomics:*** TF-TG relationships established from publicly available ChIP-sequencing studies and qualified with our NGS-QC Generator approach (Mendoza-Parra et al) (Mouse and Human collection available).

(iii) ***regulatoryCircuits:*** TF-TG relationships established by Marbach D. *et al* from the cap analysis of gene expression (CAGE) performed by the FANTOM consortium combined with sequencing binding motif analysis (Marbach et al. 2016) This resource is composed by GRNs defined for 394 human cell types/tissues, which has been collected in a single GRN.

In all cases, TETRAMER extracts transcription factors (TFs)-target gene (TGs) connections from one of the aforementioned generic GRNs by selecting relationships in which the target genes are differentially expressed. For it, users might display the dialog window by going to Apps → TETRAMER → Build a Gene Regulatory Network.

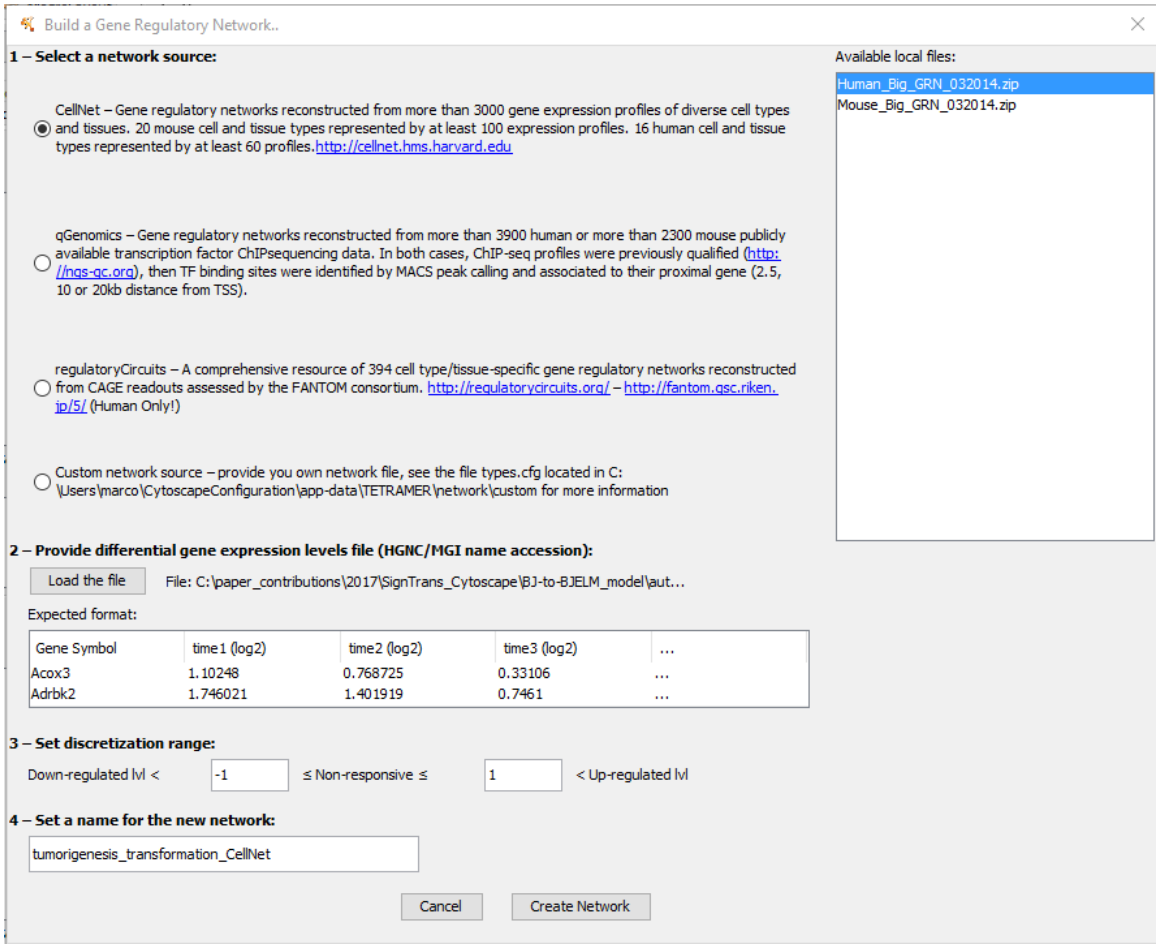

Figure 1. TETRAMER interface allowing to build GRNs from differential gene expression data.

4 Steps are required to construct a network:

1. Select a network source (CellNet, qcGenomics, regulatoryCircuits or Custom network source). The access to any of those proposed network sources might initially require to be downloaded (either automatically done by TETRAMER when an internet connection is available). On the right, the different available networks will appear and give the user the possibility to choose one of them.

2. Load a gene expression level file in order to extract the list of differentially expressed genes that will be used to construct the network. The file contains, for each entry, the gene name, and one or several (time series) differential gene expression level (relative to the gene expression level at T0). Note that, as illustrated in the example displayed in the panel, a **header line is required, the Gene symbol must be displayed at the first column and the differential gene expression levels must be in log2.**

3. Define the upper and lower limits for gene expression levels. Up-regulated and Down-regulated genes/TFs are used to construct the network from TF/TG interactions databases.

4. Provide a name for the new network that you can easily identify . Once the new network is generated, a new network collection will appear under the "Network" tab in the Cytoscape control panel.

NOTES:

- You can manually download interaction files on source websites and put them in the corresponding directory (more informations in the “Custom network source” section).

- On regulatoryCircuits, only “Human” data is available

- On the uploaded gene expression level file, the field separator is automatically detected, but the algorithm is only searching for TAB, COMMA, or SEMI-COLON character.

- The format of the file is checked upon loading, an error message is displayed if the file is ill-formatted.

4.0 MODELING TEMPORAL TRANSCRIPTIONAL REGULATION PROPAGATION

To identify key transcription factors (master regulators) which are critical for cell fate commitment, TETRAMER models a temporal transcriptional regulation cascade from a set of defined “**Start nodes”** towards a pull of “F**inal nodes”** expected to define the cell transformation state (e.g. cell differentiation, reprogramming; etc). This signal propagation is expected to take place through the defined network structure and in coherence with the differential expression state of each node at each defined time-point. Furthermore, the temporal transcriptional response might be coherent with the type of transcriptional regulation (positive or negative; i.e. transcriptional activation or repression) described between the assessed TF-TG relations. In that context, TETRAMER requires a certain number of informations which can be filled by the Control Panel in Cytoscape (**Figure 2**).
First of all, users might select a network in the drop-down list located on the top the TETRAMER panel (**Figure 2A**), then add the required information as requested over the following tabs:

4.1 "Nodes Selection" tab

Users might provide a list of “**Start nodes”** for which, their capacity to drive the cell state transformation will be evaluated. In the same manner, users might provide a list of “**Final nodes”**.
There are multiple ways to select nodes:

- double-click on a gene name on the list of proposed nodes (as retrieved on the selected network; left panel).

- Select multiple nodes and use the "arrow" buttons (**Figure 2B**) to add or remove nodes from the lists.

- Submit a list of gene names (**Figure 2C**), and click on the "Add" button.

- Select nodes in the view panel, use "Apps → TETRAMER → Add selected nodes to Start/Final Node list" in the pop-up menu brought by right-clicking on a empty area of the network view panel.

NOTES:

- If there are no gene names displayed for selection, make sure that a network has been selected in the drop-down list (**Figure 2A**).

- Users can search for a specific gene name by using the search bar located on top of either the start or final nodes' lists.

- Users' provided "Start nodes" presenting down-regulated expression levels over the whole kinetics study will be excluded from the analysis. On the contrary, non-differentially or up-regulated genes will both be evaluated for their capacity to induce transcription signaling cascades till the defined final list. For this reason, *users might exclude non-differentially expressed genes from their list of "Start nodes" in cases they are only interested in up-regulated transcription factors*.

Figure 2. Setup of Start and Final nodes. Multiple options to define lists of start and final nodes are available in TETRAMER. **(A)** Users might first select a GRN on which the analysis will be performed; then the list of Start and Final nodes can be loaded by: **(B)** selecting a set of genes from the proposed list; **(C)** pasting a list of gene names; or **(D)** by selecting a set of nodes from the Network visualization panel.

4.2 "Time Points" tab

Transcriptional propagation modeling performed by TETRAMER requires to have differential expression information. While the analysis can be performed with a single readout (See, *section 2.4*), TETRAMER was primary designed for the analysis of multiple time points readouts.

For this reason, if the network in use has not been built by TETRAMER, users might upload differential gene expression levels as node attributes (File → Import → Table → File..., please refer to Cytoscape manual instructions for more information). This being done, users might retrieve this information in the "Time Points" tab as illustrated in **Figure 3**. In order to specify the gene expression time points which will be used during the transcriptional propagation modeling, users might select and transfer them to the "Selected attributes panel" as suggested for the selection of the "Start" or "Final nodes" (**Figure 3A**). Furthermore users have the possibility to define differential gene expression

thresholds (**Figure 3B**). At this point, they can access the number of upregulated/downregulated genes by clicking on the "View discretization table". In that way, they could decide to keep a defined threshold or modify it prior initiating the transcription modeling process.

NOTES:

- At least one time point must be selected.

- Make sure to correctly order the time column : the first time point should appear on the top and the last on the bottom of the list. You can reorder selected items by using the "UP" and "Down" buttons (**Figure3C**).

Figure 3. Defining temporal differential gene expression attributes.

4.3 "Edge Correlation" tab

As previously mentioned, temporal transcriptional regulation propagation is expected to take place through the defined network structure in coherence with the differential expression state of each node as well as with the type of transcriptional regulation (positive or negative; i.e. transcriptional activation or repression) described between the assessed TF-TG relations. In the particular case of the gene regulatory networks established by Cahan et al. 2014; each edge contains an attribute "correlation (or corr)" which makes reference to the type of transcriptional regulation. This attribute is positive in the case of transcriptional activation, and negative in the case of transcriptional repression. Thus, during modeling, TETRAMER verifies that the correlation attribute matches the nature of the differential expression behavior of the interconnected nodes, and if it is not the case, the corresponding transcriptional regulation propagation cascade is terminated due to its incoherent behavior (**Figure 4A**).
To do so, , users might select the "corr" attribute on the “Edge correlation” tab (**Figure 4B**). Furthermore, users have the possibility to filter edges by modifying the threshold criterion for the "corr" attribute (See discretization range; **Figure 4B**).
In contrast to "CellNet", gene regulatory networks reconstructed either from ChIP-sequencing readouts (qcGenomics) or from CAGE information ("Regulatory circuits") do not provide a "corr" attribute. In fact, in the case of ChIP-seq-derived networks, the type of transcriptional regulation is not known since the approach for associating a gene with a given TF is simply based on the presence of a binding event in proximity to the corresponding gene without information on its transcriptional response. On the other hand, CAGE-derive networks are per definition associated with transcriptional activation since the presence of a TF has been linked to the presence of a CAGE readout on the proximal promoter.
This being said, in the case of ChIP-seq derived networks, users can filter edges on the basis of either the TF-enrichment confidence p-value (predicted with the peakcaller MACS2) or the corresponding read count intensity. In the case of CAGE-derive networks, a "p-value" edge attribute is available. In both cases, users could select these attributes (when available), then filter edges using the "discretization" panel and selecting "ignore correlation sign" (**Figure 4B**).

NOTES:

- TETRAMER allows to assign different attributes to different time-points (custom networks). Thus, users might select as many attributes as time-points and order them in the same manner.

- The button "View discretization table" allows to visualize the number of edges available for the analysis after threshold application. This way, users might evaluate the number of available edges prior performing propagation runs.

4.4 "Advanced options" tab

4.4.1. Start nodes iteration method

By default, TETRAMER tests the capability of each gene provided in the "Start nodes" list to drive the temporal transcriptional regulation cascade towards the set of defined "Final nodes". This mode is defined in the "Advanced Options" as the iteration method "one by one" and is very useful when the user does not have a clear idea of the potential master regulators of the system under study. Assuming that the user possesses a list of potential candidates (one or more), TETRAMER provides the possibility to perform the analysis from a group of defined genes (provided on the "Start nodes" list). In this mode - defined as "iteration by group" - all temporal transcriptional regulation cascades derived from the genes retrieved on the "Start nodes" are evaluated at once. Note that in this mode, users have the possibility to include a name to the group of genes.

If users are interested in predicting master regulators from a list of "Start nodes", TETRAMER provides the possibility to predict combination of master regulators able to provide higher performance for the signal propagation towards the “Final nodes”. Thus, users might use the "one by one" iteration mode, then check the option "Search for start nodes combination". Then, parameters concerning the minimum yield and specificity associated with the nodes which have to be combined or those which will be obtained with the combination can be defined. Furthermore, the maximum number of nodes which have to be included in the combination can also be set (see figure 5).

4.4.2. Knock-out nodes as part of the transcriptional regulation model

TETRAMER has been designed to perform temporal transcriptional regulation cascade models in the context of "wild type" situations. Nevertheless, the possibility to incorporate *in-silico* "knock-out" events has also been considered during its implementation. Thus, users have the possibility to include a list of genes to consider as absent via the "Knockout" tab. Then in the "Advanced Options" panel, users will have the possibility to choose between two “KO nodes” iteration methods: "one by one" or "by group". Like in the case of the "Start nodes", the mode "by group" uses the list of genes as an ensemble of KO events (i.e. their transcriptional response state will be defined as zero). In the case of "one by one" mode, one gene at a time is considered as absent during the transcriptional regulation propagation assay; thus TETRAMER will evaluate the capacity of signal propagation for each gene retrieved on the KO list.

In both cases, an additional results table (see *section 2.5*) displaying the effect of the “KO nodes” on the signal propagation cascade is available.
Finally, in the case of the "one by one" mode, users have the possibility to predict combination of “KO nodes”. Thus,, users might check the option "Search for KO nodes combination", then parameters concerning the minimum yield and specificity associated to the “KO nodes” which have to be combined or those which will be lost with the combination can be defined. Furthermore, the maximum number of nodes to include in the combination can also be set.

NOTES:

- When the “Start nodes” iteration method is defined as "one by one", the “KO nodes” iteration method can only be used "by group". On the contrary, when the mode "by groupe" is selected for the “start nodes”, users have the possibility to use either“KO nodes” iteration methods.
- If a gene is defined as both "Start node" and "KO node", a message will be displayed and the knockout will not occur.
- A “Final node” can be selected as “KO node”, consequently this “Final node” will never be activated during the regulation cascade.

4.4.3. Vertical versus horizontal temporal transcriptional regulation models

TETRAMER has been designed to perform temporal transcriptional regulation propagation from a defined "Start node" towards a group of defined "Final nodes". While this approach implies that multiple time-point readouts should be available for the transcriptomes, TETRAMER has also shown capacities to predict master regulators in the context of a single time-point; for instance when the only differential expression information available provides from the comparison between a terminal and the initial state. In this context, TETRAMER might require to evaluate the capacity to propagate the transcriptional information over a single temporal state, which corresponds to propagate over the available horizontal edges (in contrast to a vertical propagation in the case of multiple temporal readouts). For it users can activate the use of horizontal edges in the tab "advanced options" (see figure 5).

4.5 Tab "Run"

Once all the parameters are setup, users are ready to run the signal propagation model. This is possible by the tab "Run" in which the following options are available:

- Save the parameters by the "Save configuration" button or reload others previously saved by using the " Load configuration" button.

- Perform multiple network randomization runs as a way to evaluate the confidence of the predicted transcriptional propagation yields/specificity. This is performed by randomizing the nodes interconnection but by keeping constant the number of nodes and edges per evaluated network.

Finally, users can start the transcriptional regulation propagation modeling by the "Start the simulation" button.

4.6. Modeling a temporal transcriptional regulation propagation

Once the simulation is initiated, users could get the following warnings:

- Information concerning the number of nodes for which differential expression readout is not available. This warning does not necessarily imply that there is problem with the gene regulatory network in use, but in cases the number of nodes without such information appears relatively high, users might verify the reason why this information is not available in the GRN in use.

- Information concerning the number of Start nodes that were discarded for further analysis due to the fact that they are down-regulated over all time-points.

After these two potential warnings, two subsequent panels provide information concerning the number of final nodes that are differentially expressed at each time-point. The first panel provides the time-point at which the final nodes are observed as differentially expressed for the first time, and it gives the possibility to define whether to use all final nodes at all time-points for the analysis (by default) or to exclude some of them due to their temporal behavior. In addition it indicates the number of final nodes for which their expression levels are not bypassing the defined thresholds (non included in the analysis by default). Once these selection is performed, a second panel displays the frequency of expression of all final nodes at each selected time-points in the previous panel. This second panel aims at providing to the user an enhanced control on the selection of the final nodes to be used during the modeling.
Once the group of final nodes to be used is defined, TETRAMER performs the temporal transcription propagation modeling and provides the following items at the end of this process:

- An "edge statistics" panel, in which the number of total directed edges, those in a self-looping organization, as well as those that are duplicated are displayed. Finally this panel includes the number of edges that were used for the network randomization process.
- A new table panel containing the results of the signaling propagation is generated under the name "TETRAMER results".

4.6.1. Yield and specificity predicted by TETRAMER

TETRAMER aims at evaluating the capacity of each gene, retrieved on the "Start nodes" list, to drive the transcriptional regulation propagation towards the defined group of final nodes. For it, two major characteristics are computed:

- *the yield of final nodes activation*: Defined as the fraction of final nodes (in percent) that are retrieved in the temporal transcriptional propagation cascade relative to the total number of final nodes defined by the user.

  (ADD FORMULA HERE)

- *the specificity of the final nodes activation:* Defined as the fraction of final nodes (in percent) that are retrieved in the temporal transcriptional propagation cascade relative to the total number of nodes.
  (ADD FORMULA HERE)

In the example illustrated in Figure 7A, The user have selected the 5 final nodes in the last two time points (green color). The transcriptional propagation assay gave rise to 6 nodes (displayed with a blue border), among them only three of the defined final nodes are retrieved. Thus, the yield in this assay is of 3/5 *100 = 60% and the specificity value is 3/6 * 100 =50%. In this manner, the best master regulators might be those presenting the highest yield and specificity.
The result table generated by TETRAMER, present all Start nodes ranked by the yield and specificity. For each of them their corresponding yield and specificity (in percentage) are displayed. In addition, the number of directly activated ("+ direct regulated") or directly repressed ("- direct regulated") genes are also displayed, both in absolute numbers as well as in percentage.

4.6.2. Other Results tables

4.6.2.1. Node (random)
The number of sub-tab depends on the parameters of the modeling, for instance a sub-tab "Node (random)" will only be visible, if you have selected the option "Test the network" found in the panel run. In this case, genes are still ranked on the basis of the yield and specificity computed during the temporal transcriptional regulation propagation, but in addition a "mean yield" and a "mean specificity" - accompanied by their associated standard deviations (sd)- derived from the analysis over the multiple randomized GRNs are displayed. Finally, a p-value for the yield and the specificity are derived from a Normal distribution model.

4.6.2.2. Node (TP selected)

In the case that the final nodes are retrieved on multiple time-points a sub-tab called "Node (TP selected)" is displayed. This table contains the same structure than the sub-tab "Node", but illustrated per time-point. For it, the displayed yield and specificity are calculated by taken in consideration only the nodes present at each time-point.

4.6.2.3. Node (KO)

The information displayed in this table depends on the start nodes / KO nodes iteration method selected in the tab "Advanced Options".
With start nodes iteration "one by one" or "by group", KO nodes iteration "by group":
Displays the scores of each start node or group of start nodes where all KO nodes are inactivated at the same time. In addition to the structure of the previous described tables, a "delta yield" and a "delta specificity" column is displayed, corresponding to the

difference in yield or specificity as consequence of the inactivation of the aforementioned KO nodes.
Furthermore, With start nodes iteration “by group”, KO nodes iteration “one by one”, this table displays the scores of the group of start nodes with each KO nodes inactivated. The column “Start Node” is renamed as “KO Node”. In this case, the "delta yield" and a "delta specificity" makes reference to the effect of the absence of each of the listed genes in the Knockout nodes list relative to the yield and specificity computed for the start nodes iteration "by group".

4.6.2.4. Node Combinations

This table displays the scores of the best combinations of start nodes selected based on the parameters given in the advance options panel. In constrast to the previous result tables, in this case, the yield and specificity are computed for node combinations, for which their gene names are displayed. Furthermore a "delta yield" and a "delta specificity" column are available displaying the gain on these parameters relative to the best yield obtained with one of the single nodes being part of the combination.

4.6.2.5. Node Combinations (KO)

Like in the previous case, this table displays the score of the best combinations of KO nodes selected based on the parameters given in the advanced options panel. As previously indicated, a "delta yield" and a "delta specificity" column are displayed in addition, corresponding to the lost in such parameters as consequence of the KO combination.

4.7. Visual options as part of the results generated by TETRAMER

Besides of the results table, TETRAMER provides three further features:

4.7.1. “View signal propagation network”
compute and display the cascade of regulation network for the current selected result (row), if multiple rows are selected, It generates multiples networks unless you have checks the option “Merge into a single network”. The networks created can be visible under the tab “Network” of the Control Panel.

NOTES:

- In case of KO nodes, Node and edge inactivated by the knockout are dotted and black colored.

- This button is disabled for the results in the tabs “Node (TP selected)”, “Node (Random)”, “Node Combinations” and “Node Combinations (KO)”

4.7.2. “View relationship network”
displays the sub-network between the start nodes/group node selected. Only the direct connections are shown, this may result in having single nodes. The option “Display only

edges in accordance with time activation” limits the display to connections having a source node activated before the target node.

NOTE:

- This button is disabled for the results in the tab “Node (TP selected)” and“Node (KO)”.

4.7.3. "Save results"

Users can save their results (results tables and settings) by clicking on the button “Save results” and selecting a directory.