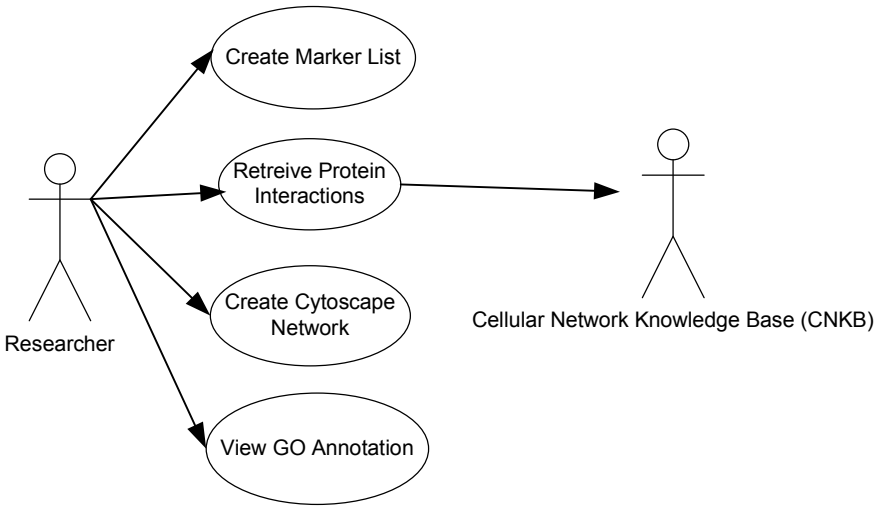
Revision History

Date	Version	Description	Author
09/24/2006	0.1	Initial draft	Eileen M. Daly
11/21/2006	0.2	Updated to reflect feedback	Eileen M. Daly
1/3/2007	1.0	Add display preference subflow	Eileen M. Daly

Table of Contents

1 Cellular Network Knowledge Use Case

1.1 Use Case Model



1.2 Brief Description

This use case describes how to access protein interaction data residing in the Cellular Network Knowledge Base (CNKB) and how to visualize this data as an interactions network within Cytoscape. The CNKB is a repository of interactions between pairs of proteins (these interactions can be computationally or experimentally derived). Both direct, physical interactions can be captured as well as indirect transcriptional relationships (where an interaction is between a transcription factor and its transcriptional gene target). For the purposes of the present document, an interaction is assumed to be a binary relationship between two genes. Each interaction has an associated confidence indicator (a value between 0 and 1) characterizing our level of confidence for the method aggregate (computational or experimental) used to derive the interaction.
Briefly, the use case workflow is as follows: users begin by selecting a group of markers and specifying for each marker the interaction type of interest (Protein-Protein and/or Protein-DNA). All interactions (of the designated types) involving the selected markers are retrieved from the CNKB and displayed (along with associated information such as markers GO annotation, interaction attributes, etc) in two geWorkbench visualization modules: (1) Cellular Network KB, and (2) Cytoscape.

1.3 Glossary

Term	*Definition*
I(A, B)	An interaction involving genes A and B. When A and B are unambiguously defined by the context (or are not germane to the discussion) we will drop the parentheses and just use I to refer to the interaction.
conf(I)	The confidence indicator for interaction I. conf(I) assumes a value between 0 and 1. The higher the value the more confident we are that the interaction is real (roughly speaking conf(I) is the probability that the interaction is true).
S(G), S_{PP}(G), S_{PD}(G)	The set of (respectively) all interactions, all protein-protein interactions, and all protein-dna interaction reported in CNKB that involve gene G.
S(GS), S_{PP}(GS), S_{PD}(GS)	Here GS = {G_1, G_2, …, G_N} is a set of genes and S(GS) is the algebraic union of the sets S(G_1), S(G_2), …, S(G_N). I.e., S(GS) contains all the interactions involving genes G_1, G_2,… G_N. S_{PP}(GS) and S_{PD}(GS) are defined accordingly, using only protein-protein or protein-dna interactions.
S(G, a), S_{PP}(G, a), S_{PD}(G, a)	The subset of S(G), S_{PP}(G), S_{PD}(G) comprising only those interactions I for which conf(I) >= a.
S(GS, a), S_{PP}(GS, a), S_{PD}(GS, a)	Again, GS is a set. As with the definition of S(GS), S(GS, a) comprise the union of S(G_1, a), S(G_2, a), …, S(G_N, a). And accordingly for S_{PP}(GS, a) and S_{PD}(GS, a).

1.4 GUI Overview

1.4.1 Cellular Network KB

This module comprises three sections: Activated Markers, Selected Markers and Throttle Graph.

- *Activated Markers*: Table that lists all markers that belong to activated marker groups. Contains three columns: Marker Name, Gene Name (the gene name corresponding to the marker, if known; empty otherwise), and Gene Type. The Gene Type column can assume one of three possible values: TF (if the gene in the Gene Name column is a Transcription Factor), K (if the gene is a Kinase) and P (if the gene is a Phosphatase). Gene types are derived from the GO annotation associated with a gene name (specifics are provided in the flows below). If none of the TF, K or P is applicable, the contents of the Gene Type column are left empty. See non-functional requirements for a precise description of how the contents of the Gene Type are derived from the GO annotation of a gene.

The table supports multiple selections. Users can select any number of markers from this list and move them to the Selected Markers table.

- *Selected Markers*: Table that lists the subset of activated markers that were moved over from the Activated Markers table. The table supports multiple selections. Users can select any number of markers and remove them from the list (i.e., send them back to the Activated Markers table; details provided in the flows below). This table has the following columns:

    a. Marker name: as above.

    b. Gene name: as above.

    c. Gene type: as above.

    d. Entrez Id: The Entrez ID of the gene name (if known; empty otherwise). When present, the Entrez ID is hyperlinked to the relevant entry within Entrez Gene (we need to provide specifics here, exactly which Entrez page we link to for any given Entrez ID).

    e. GO Term: Associated GO Term. Users can browse the GO terms associated with a gene (using an AmiGO browser-like interface) and select one of those terms for display.

    f. Prot-Prot #: number of protein-protein interactions (reported in the CNKB) involving the gene of bullet (b) above.

    g. Prot-DNA #: number of protein-dna interactions (reported in the CNKB) involving the gene of bullet (b) above.

  Additionally, each row has associated check boxes that can be used to designate if all the interactions for a gene are to be (both check boxes are checked) or only protein-proteins or protein-dna interactions.

  Using a “preferences”-type tab user can control how many of the columns listed above they want visible within the Selected Markers table.

- *Throttle graph:* The graph allows users to throttle (for the genes in the Selected Markers table) which interactions to work with using as a criterion the interactions’ confidence indicator. The X axis of the graph ranges from 0 to 1, in increments of 0.01. For any value 0<=a<=1 on the X-axis, the throttle graph contains the point (a, y) where y is computed as follows:

    a. Consider the set GS of all genes in the Selected Markers table.

    b. For each G in GS, let Y(G) be defined as follows:

- If both checkboxes are selected for G, then Y (G) = S (G, a).
- If only the protein-protein checkbox is selected for G, then Y (G) = S_{PP}(G, a).
- If only the protein-DNA checkbox is selected for G, then Y(G) = S_{PD}(G, a).
- If neither check box is selected for G, then Y(G) = empty set.

c. The value y for the graph point (a, y) is then defined as the size of the set that results from the union of all the Y(G), for all G in GS. Note that this is an algebraic union, i.e., there is no double counting of interactions (for interactions I(A, B) where both A and B belong to GS).

d. Users can select a value for the X-axis by using either the graph slider (has a range of 0-1 and moves in intervals of 0.01) or the associated threshold box (only positive real values between 0-1 are allowed in that box; values are rounded to the closest second decimal point). When a value X=a is selected by the user for the X axis, the GUI is modified as follows:

• The vertical bar indicator on the graph moves on the X-axis, to the position X=a.

• For each gene G in the Selected Markers table, the contents of the columns Prot-Prot # and Prot-DNA # are updated as follows:

- Prot-Prot # = S_{PP}(G, a). Prot-DNA # = S_{PD}(G, a).
- If the Protein-Protein checkbox is **not** selected for G, then the contents if the “Prot-Prot #” column appears grayed-out (but the number therein is still visible). The graying out indicates that this group of interactions will not be taken into account when building the Cytoscape network (see below).
- If the Protein-DNA checkbox is **not** selected for G, then the contents if the “Prot-DNA #” column appear grayed-out (number therein is still visible). As above, in this case the Prot-DNA interactions will not be taken into account when building the Cytoscape network.

So, as the user interacts with the throttle graph slider the contents of the Selected Markers table (specifically, the contents of the “Prot-Prot

#" and "Prot-DNA" columns) are updated on the fly, using the numbers specified above.

1.4.2 Cytoscape Network

When the user clicks on the "Create Network" button the application builds an adjacency matrix comprising those interactions that were used to compute the contents of **non-grayed out** cells of the "Prot-Prot #" and "Prot-DNA" columns in the Selected Markers table. I.e., for the currently selected value X=a on the X-axis, and for any gene G on the table, the interaction set S (G, a), S_{PP}(G, a), S_{PD}(G, a) or null is used depending on if (respectively):

- Both Protein-Protein and Protein-DNA checkboxes are selected for G.
- Only the Protein-Protein checkbox is selected for G.
- Only the Protein-DNA checkbox is selected for G.
- None of the Protein-Protein and Protein-DNA checkboxes are selected for G.

The project panel is updated to reflect the addition of an adjacency matrix results node which can be viewed in the Cytoscape visualization component. A number of conventions/rules apply when drawing the adjacency matrix in Cytoscape (more details in the flows below):

- **Network nodes** (which correspond to genes) have a shape determined by the gene type (TF → circle, K → square, P → diamond). Users can right click on any node and get additional information about the gene, including gene description, GO annotation, T-test value and associated p-value for cases vs. controls, average expression in controls, average expression in cases (the last 3 bits of information are displayed only if cases and controls have been specified on the input microarray data set). Further, also when cases and controls have been specified on the input microarray data set, the nodes are color-colored using a blue-red gradient scheme (blue for negative t-test values; red for positive). Specific instructions about the color coding are provided in the non-functional requirements portion of this doc.
- **Network edges** (corresponding to interactions) are drawn as directed arrows (for protein-DNA interactions) or as undirected arrows (for Protein-Protein interactions). Edges I(A, B) are color-coded according to the Spearman correlation between genes A and B (positive correlation = red; negative correlation = blue) and have

thickness proportional to the value of the confidence indicator conf(I). Further, for any interaction edge I(A, B) users have the ability to query GeneWays in order to retrieve literature-reported relationships between genes A and B (the GeneWays feature can be implemented in a subsequent phase).

The user can mouse select sections of the network to create Marker Set. For additional information, refer to the Cytoscape Use Case.

1.5 Annotation

Source	Description
GeneWays	GeneWays maintains and tracks molecular pathway data from research literature.
Go	Gene Ontology www.geneontology.org provides a categorization of genes in terms of function their products perform, their cellular localization or involvement in high level biological process.

1.6 Actors

Primary Actor

- Researcher

Secondary Actors

- Cellular Network Knowledge Base (CNKB)

1.7 Preconditions

Expression data has been loaded into the workbench.

1.8 Basic Flow of Events

Actor Action	System Response
1. User activates a marker set of interest.	2. System updates the Activated Marker List in the Cellular Network KB visualization to include the markers in the activated market set.
3. Marker Selection: User selects one or more markers from the Activated Markers table, right-clicks anywhere on the Activated Markers table, and selects the option “Add to Selected Markers” from the resulting popup.	4. System updates the **Selected Markers** table to include the selected marker(s) and removes the selection(s) from the **Activated Markers** table**.** 5. For markers just added to the Selected Markers table, the contents of the corresponding row in the table are displayed italicized and using red font. This provides a visual cue to the user that the CNKB has not been yet queried in order to

	retrieve the interaction information for a new marker. Further, the Refresh button gets activated.
6. Marker Removal: To remove a group of markers from the Selected Markers table user selects the markers of interest; right-clicks anywhere on the Selected Markers table, and select the option “Remove Markers” from the resulting popup.	7. System removes the selection from the **Selected Markers** table and returns the markers to the **Activated Markers** table. Only markers that are still part of activated marker groups will appear back in the **Activated Markers** table.
8. If necessary, the user returns to step Marker Selection to add markers or Marker Removal, step to remove markers. 9. User clicks Refresh. If the Refresh button is not active, nothing happens. (<AF>we my want to change this later, to query the user if the would just like to re-retrieve de novo data for all markers in the **Selected Markers** table.	10. **Interaction Preview:** System updates the numbers in all the “Prot-Prot #” and “Prot-DNA #” columns within the Selected Markers table. The numbers are computed as described in the relevant portion of section “1.4.1 Cellular Network KB”. To complete this step the system will need to retrieve from CNKB the interaction sets for all genes G in the **Selected Markers** table that appear italicized and in red font (for all other genes, this information must have been acquired from the database during the last Refresh operation). After the Refresh is complete, all rows appear using regular font (no more italics and red font). Further, the Refresh button is deactivated.
11. Display: To modify the columns displayed in the **Selected Marker table**, the user clicks the Display Preference button; then go to subflow – Display Preference.	
12. User indicates the interaction type of interest for the Selected Marker; go to subflow - Interaction Selection.	
13. To view GO annotation for a selected marker, go to subflow - Add annotation.	
14. To modify the Throttle Graph display, go to subflow – Throttle Graph. 15. If necessary, the user repeats Marker selection steps.	
16. User clicks **Create Network.** 17. User clicks **Cancel,** the system cancels the operation.	18. System displays an indicator that the analysis is in progress. 19. System builds an adjacency matrix comprising those interactions that were used to compute the contents of **non-grayed out** cells of the “Prot-Prot #” and “Prot-DNA” columns in the Selected Markers table. 20. If successful, the system indicates that

	analysis is complete and a data node on the project folder for the Network created. System creates a network (to be viewed in Cytoscape **21.** If unsuccessful, the system displays a descriptive error message.
22. To view the network, go to subflow – network.	23. End

1.8.1 Subflow – Display Preference

Actor Action	System Response
	1. System displays the Display Preference pop-up window that lists the columns available for display in the Selected Marker List. (Display Pop-up)
2. To modify the visible columns: ▪ User selects the checkboxes of column headers to be included in the Selected Marker List. ▪ User unselects the checkbox for column to be removed from the Selected Marker List display.	
3. User clicks ok. 4. User close the pop-up – the operation is canceled and returns to the basic flow.	5. System updates the display of the Selected Marker List to reflect the columns selected.If needed, the table will include a horizontal slider to support viewing the entire table. 6. Return to basic flow.

1.8.2 Subflow – Add Annotation

Actor Action	System Response
1. User right clicks on a gene of interest and selects View GO Annotation.	2. System displays a pop-up window that lists the following GO annotation filters: Component, Function, and Process. (Pop-up #1)
3. User selects a category from the pop-up window.	4. System displays a linked pop-up window for the category selected. This pop-up window lists all associated GO Terms in that category for the gene at hand (Pop-up #2). For each term the popup displays both the GO term id and the associated text

	description.
5. User selects from the list of GO Terms or closes the display by left-clicking.	6. System displays a Tree view of the GO term selected. This view illustrates the parent –child relationship of the term. Each line in the tree contains one term and term description.(Pop-up #3). This view is very similar to the tree-view offered by the AmiGO browser (see AmiGO screenshot).
7. To add an annotation description to the **Selected Markers** table, user clicks on a GO Term in the tree view.	8. System updates the contents of the GO Term column of the **Selected Markers** with the selected GO Term. 9. The system includes a horizontal scroll bar to view all of the columns. The system supports resizing the column widths.
10. To modify column width: User drags the boundary on the right side of the column heading until the column is the width that desired.	11. System updates the display of the table to reflect the column width modification. 12. Return to basic flow.

1.8.3 Subflow – Throttle Graph

Actor Action	System Response
	1. System displays the Throttle Graph.
2. User indicates interaction selection; go to Selection flow.	
3. User modifies the value of the x axis by 1) using the slider or 2) entering a numeric value in the Threshold text box.	4. If the slider is used, the slider moves in intervals of 0.01 **5.** If the threshold box is used, the system verifies that entry is valid. Only positive real values between 0-1 are allowed in that box; values are rounded to the closest second decimal point. 6. If valid, the system updates the graph to include a subset of markers whose confidence level is above the threshold value.
7. User can capture the graph as an image by right clicking > capture image.	8. System captures the graph as image data node in the project. Refer to Image use case for additional details. 9. Return to basic flow.

1.8.4 Subflow –Interaction Selection

Actor Action	System Response
1. Selection: User indicates the interaction type of interest for markers in the **Selected Markers** table by performing one of the following steps: 2. Protein- Protein ▪ Select the Protein-Protein checkbox individually for one or more markers. ▪ Select the Protein-Protein checkbox for multiple markers by clicking on the first check box and dragging the mouse vertically over consecutive check boxes. ▪ Select the "All Protein – Protein" check box which results in enablement of the "Protein-Protein" check boxes for all markers. 3. Protein- DNA: Similar as described above for Protein-Protein.	4. Systems updates the interaction dialog box for each marker to reflect selected.
5. De-selection: User follows corresponding steps to deselect the Protein-Protein and Protein-Dna checkboxes.	6. Systems update the checkbox for each marker to reflect cleared selection. The system updates the **Selected Markers** table contents to reflect the following: 7. Prot-Prot # column contents appear grayed-out for any marker whose Protein-Protein checkbox is unselected. 8. Prot-DNA # column contents appear grayed-out for any marker whose Protein-DNA checkbox is unselected. 9. Return to basic flow.

1.8.5 Subflow – Network

Actor Action	System Response
1. User selects the adjacency matrix data node from the project folder.	2. System displays the network created in Cytoscape visualization component. Shapes and colors of network nodes and edges are drawn as described in section 1.4.2 Cytoscape Network. Each node is labeled with the corresponding gene name.
3. To retrieve annotation, user mouse selects interaction line(s) and right click > View Annotation.	4. System retrieves the GeneWays annotation corresponding to the selected interaction lines in a pop-up window.

5. To close, the user clicks the close window icon.	6. System updates the display to reflect closing the annotation window ***<Aris>This functionality can be postponed for later</Aris>***.
7. The user right-clicks on a node in the network to retrieve information about the corresponding gene.	8. System retrieves the gene information corresponding to the selected network node and displays it in a pop-up window. This information includes: 9. The gene name.(Node Annotation Pop-up 1) 10. A menu option (titled "Description") to see the gene description. If the users select this option, systems display gene description from the Affy annotation file, if available. (Node Annotation Pop-up 2) 11. 3 menu options (titled "Component", "Function", and "Process"). Selecting any one of them brings up a subordinate popup that lists the GO terms associated with the gene in the selected part of GO. (Node Annotation Pop-up 3 & 4) 12. T-test value (and within parentheses, associated p-value) of gene in cases versus controls. This value is displayed only if in the Markers Panel there are activated panels which have been classified as cases and controls. (Node Annotation Pop-up 1) ○ Average expression of gene in cases. This value is displayed only if in the Markers Panel there are activated panels which have been classified as cases and controls. (Node Annotation Pop-up 1) ○ Average expression of gene in controls. This value is displayed only if in the Markers Panel there are activated panels which have been classified as cases and controls. (Node Annotation Pop-up 1)
13. To close, the user clicks the window.	14. System updates the display to reflect closing the annotation window.
15. To amend the display, user activates marker sets.	16. The network reflects the inclusion of the Activated Marker Set. ▪ red indicates positive t-test values ▪ blue indicates negative t-test values

1.9 Post Condition

- List of markers and the associated number of interactions (protein to protein or protein to DNA).

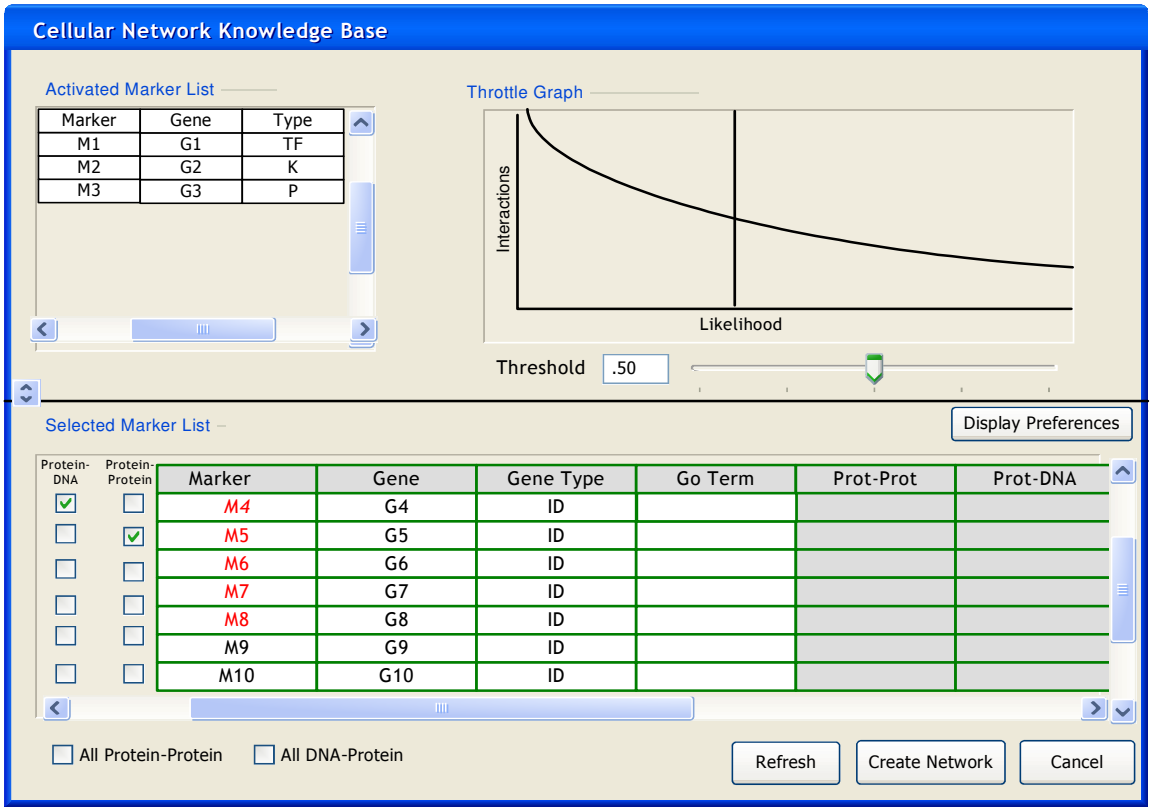
1.10 User Interface

The following is a list of some key components of the GUI.

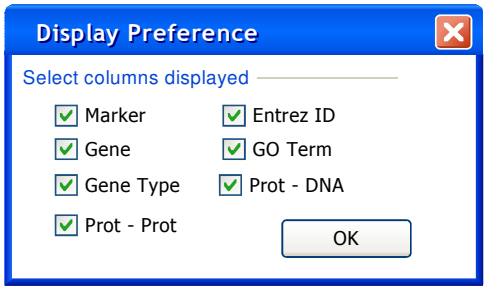
1.10.1 Interaction

Field Name	Data Type	GUI	Description
Activated Marker List		Display	Displays markers from the activated marker set.
Selected Marker List		Display	Displays markers selected from the Activated Marker List
Gene Name		Display	Gene name corresponding to the marker, in known; empty otherwise.
Entrez Id	Integer	Display/ Link	The Entrez ID of the gene name (if known; empty otherwise). When present, the Entrez ID is hyperlinked to the relevant entry within Entrez Gene.
Protein - Protein		Checkbox	Indicates the number of protein to protein interaction reported in the CNKB.
Protein - DNA		Checkbox	Indicates the number of protein to DNA interaction as reported in CNKB.
Create Network		Button	Loads to network to Cytoscape.
Refresh		Button	Updates the Selected Marker Table values.
Threshold		Slider/Text box	Default .5 The slider moves in 1% increments
Display Preferences		Button	Displays the preferences Pop-Up 1 or Pop-up 2
Cancel		Button	Cancels the operation.

1.10.2 Markers Added to the Selected Marker List



1.10.3 Display Preference Pop-Up

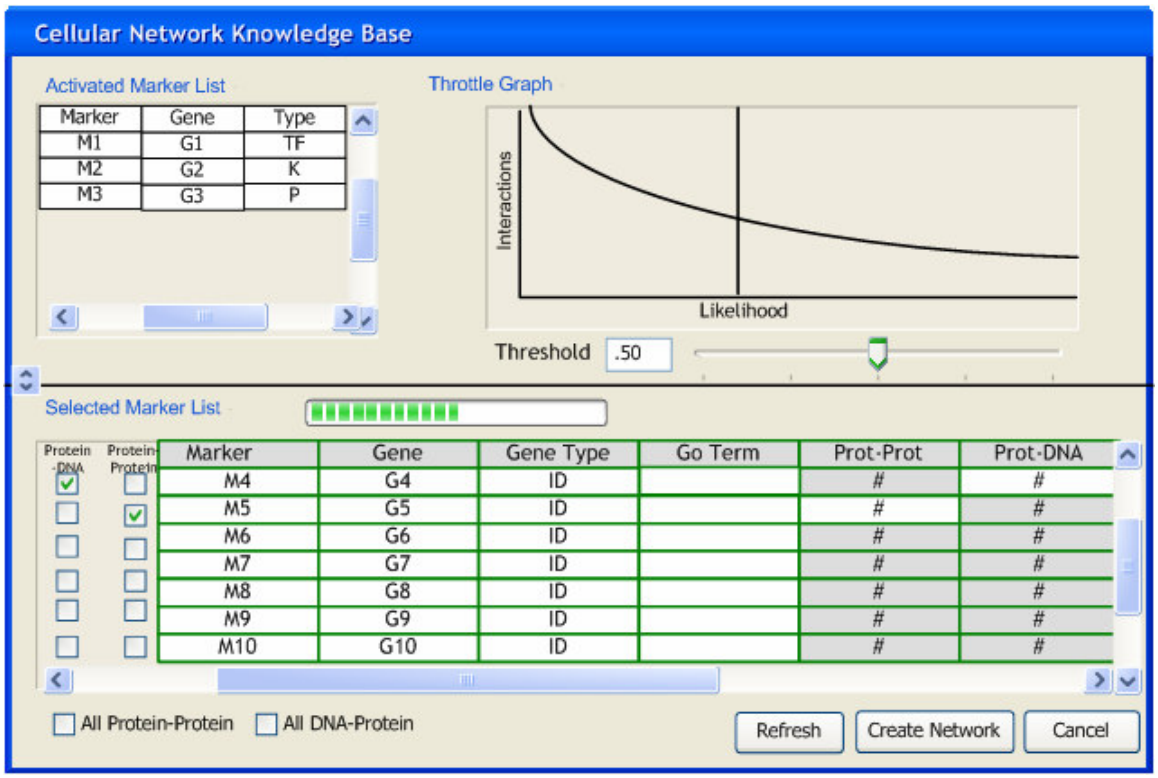


Field Name	Data Type	GUI	Description
Column Header	Display	Checkbox	Supports selection the column headers displayed in the Selected Marker List. The default selection includes: Marker, Gene, Gene Type, Prot –Prot and Prot-DNA. Additional columns include: Entrez Id,

			GoTerm.
OK		Button	Updates the columns displayed in the Selected Marker List

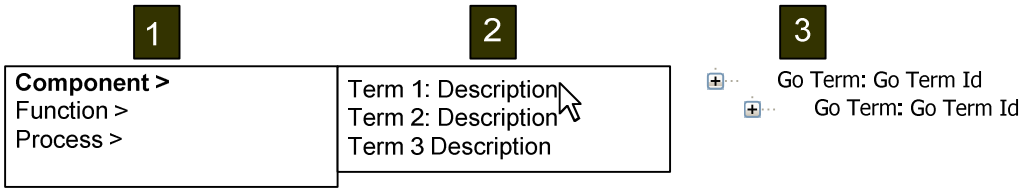
1.10.4 Interaction Graph

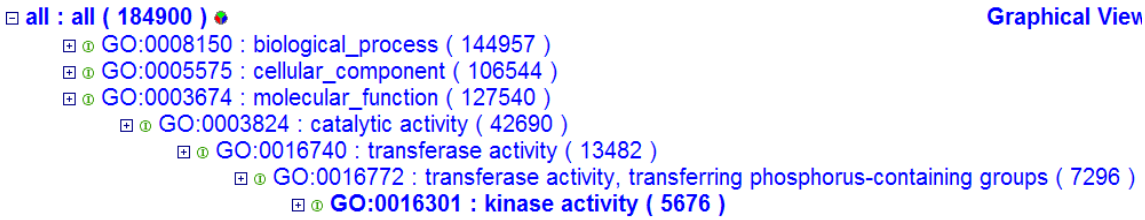
Field Name	Data Type	GUI	Description
Protein - Protein		Checkbox	Default – Unselected
All Protein - Protein		Radio Button	Default - Unselected
Protein - DNA		Checkbox	Default – Unselected
All Protein - DNA		Radio Button	Default - Unselected
Cancel		Button	Cancels the operation.
Create Network		Button	Loads to network to Cytoscape
Refresh		Button	Updates the display of the interaction selected for the **Selected Marker List**.
Threshold	number between 0-1	Slider/Text box	Default .5 The slider moves in 1% increments



1.10.5 Annotation

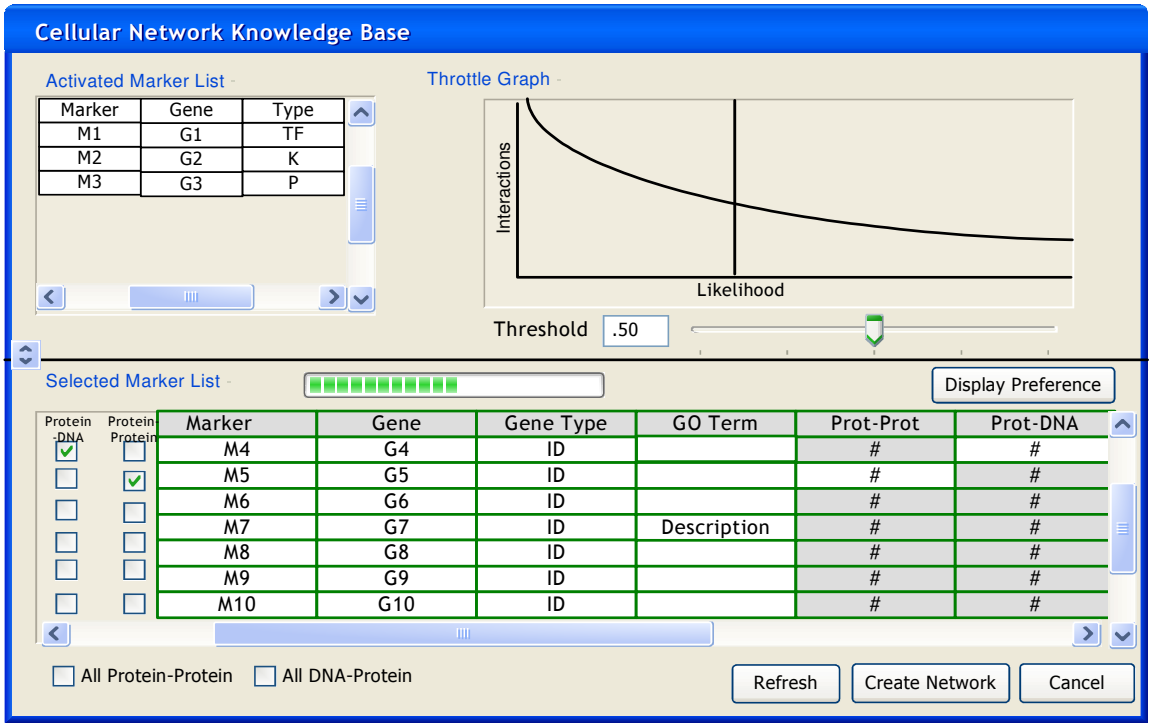
Right clicking on a marker in the selected marker list will display:



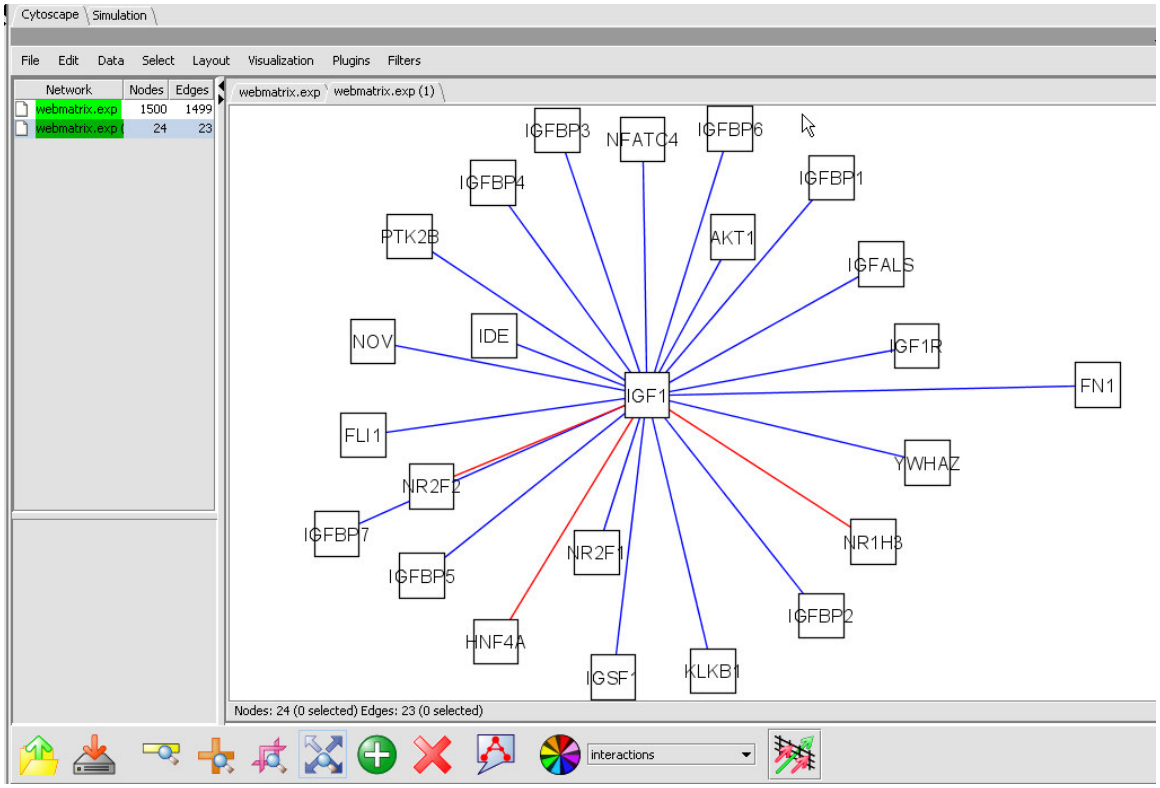


Field Name	Data Type	GUI	Description
Category Filter			GO category filter Component, Function and Process
Go ID: Go Term	Max Char:		Default – Unselected Each term is identified by

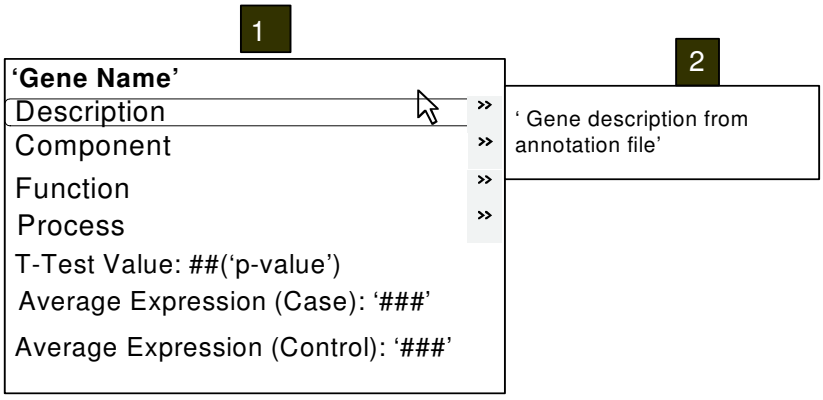
			its unique ontology ID followed by the name of the term.

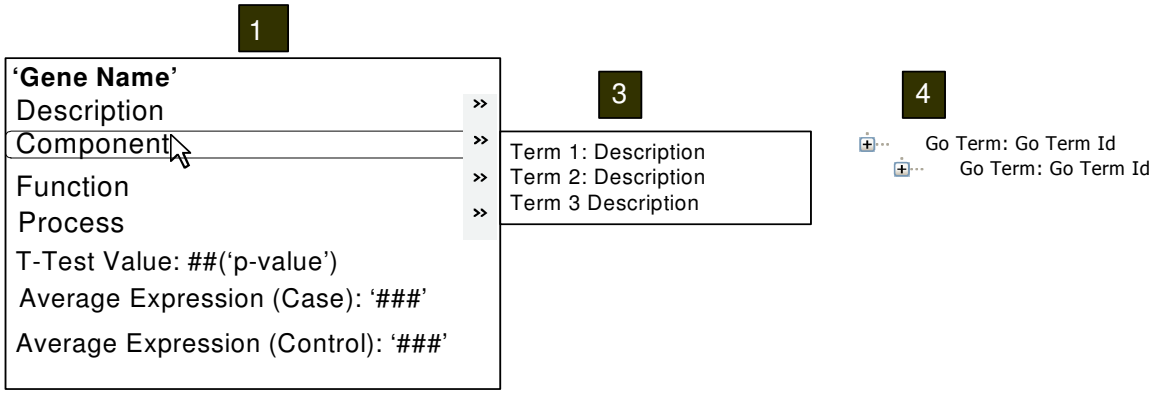


1.10.6 Network



1.10.7 Node Annotation





Field Name	Data Type	GUI	Description
Gene name	Character Max Char:30	right click node	Name of the gene which corresponds to the network node selected.
Gene description	Character Max Char: 30	right click node	Description of the gene which corresponds to the network node selected (from the Affy annotations file, if available).
Component	Character Max Char: 30	right click node	Go annotation which corresponds to the network node selected.
Function	Character Max Char: 30	right click node	Go annotation which corresponds to the network node selected.
Process	Character Max Char: 30	right click node	Go annotation which corresponds to the network node selected.
T-Test Value: (p- value)	Integer	right click node	Displayed only when cases/controls have been specified.
Average Expression in cases	Integer Min: Max:	right click node	Displayed only when cases/controls have been specified.Displayed for T-Test Results Only
Average Expression in control	Display Min: Max:	right click node	Displayed only when cases/controls have been specified.Displayed for T_Test Results Only
Network edges	Display		Size of line is proportional to

			probability of interaction in increments of 10%. Relationships that have a positive spearman correlation are color coded (blue = positive negative spearman, red = negative positive spearman) Lines corresponding to DNA-Protein-DNA interactions: TF to a non-TF are directed (arrows) from TF to the non-TF. Lines corresponding to DNA-DNA interactions are undirected:. (TF to TF and non-TF to non-TF) are directed. (arrows)

1.11 Non-functional requirements

1.11.1 Determination of Gene Type

Both the "Activated Markers" and the "Selected Markers" tables contain a column titled Gene Type whose contents can assume any of the values TF (for "Transcription Factor"), K (for "Kinase") and P (for "Phosphatase"). The algorithm to use for deriving these characters for any given Gene G is as follows:

- First collect all GO terms T(G) = {t1, …, tn} associated with G under the GO category "Molecular Function".
- For every ti in T(G) let Ti be defined as:
  Ti = {ti} U {t: t is a GO term that is a parent of ti in the "Molecular Function GO tree}.
- Let T_total(G) = union of all Ti.
- The value of the Gene Type column for Gene G is:
  - TF, if T_total(G) contains the term "transcription factor activity" (GO ID = GO:0003700).
  - K, if T_total(G) contains the term "kinase activity" (GO ID = GO:001630).

- P, if T_total(G) contains the term “phosphoprotein phosphatase activity” (GO ID = GO:0004721).

1.11.2 Derivation of color for Cytoscape nodes

When ***activated*** marker groups in the Markers Panel have been classified in cases and controls, then the nodes of the Cytoscape network are color-coded to indicate differential expression. To compute the colors we need to have a mapping between the gene G corresponding to a Cytoscape node and the marker P acting as the probe for G on the underlying chip (it is possible that for some G there will be no known P; or that the associated P is not present in the input data due to different chip type or due to prior data filtering). For all G-P pairs for which P is a marker on the input data, the color of the G-node in Cytoscape is computed as follows:

- First compute the t-test and associated p-value for the expression of P across the arrays that have been classified as cases and controls. The determination of the p-value should be done using the appropriate degrees of freedom (#cases + #controls - 2).

- If the t-test is negative, the node is color coded using a grade of blue. Specifically, each node is colored with the following RGB value: (X(p), X(p), 255), where 0 <= X(p) <= 255 depends on the p-value of the t-test as follows:

  - If 1 < p <=$ 0.01: X(p) = 255 – 1.27/p.
  - If 0.01 < p <= 0.0001: X(p) = 127 – 0.0127/p
  - If p > 0.0001: X(p) = 0.

  The rational behind this function is as follows: a p-value of 0.01 gets the mid-point of the blue gradient range, RGB value = (127, 127, 255). A p-value of 0.0001 (and everything above it) gets the extreme becomes extreme blue, RGB = (0, 0, 255). Other p-values are drawn accordingly somewhere in between this range.

- If the t-test is positive, the node is color coded using a grade of red. Specifically, each node is colored with the following RGB value: (255, X(p), X(p)). The formula for calculating X(p) is exactly as specified above for the negative t-test case.