

STARNET USER'S MANUAL

This set of notes is intended as a user's manual, detailing how to use StarNet as an analytical tool. While the website is not overly complicated, it is best to go through each field on the submission page, explaining how the user should fill in the field, and what information that field contains which will be passed on to the visualization script.

We will also indicate, in the appropriate settings, the implications of the choices made by the user. That is to say, we will explain, for example, why the network drawn using the Levels methodology gives different biological information than that drawn using the Highest methodology.

Lastly, we will describe results pages, detailing what information appears on these pages and how to use it.

1. INTRODUCTION

We provide here a very brief overview of and background for the StarNet application.

The motivation for this project was the attempt to reconstruct genetic regulatory networks. Various techniques are available to tackle this problem. Unfortunately, however, these techniques are often not equipped to handle more than a small network, or a small group of genes. Additionally, it is often the case that the user must provide some a priori expert knowledge to begin the network discovery process. This comes in the form of specifying prior distributions for certain kinds of Bayesian networks, or limiting the choice of possible network topologies.

Our goal was to generate putative networks based on experimental data, on which further study could be based. We do not attempt to build the entire regulatory network in one pass. In fact, our method is very gene-centric; we focus on a single gene at a time, attempting to learn how that specific gene is related to others. No user input concerning the network topology is required, the user merely specifies which is the gene of interest.

StarNetSpecies allows exploration of ten species. They are listed in the Excel file species.xls, to be found on the help page. This list includes GEO accession numbers, Affymetrix chip names, numbers of arrays used, and numbers of genes on each array. It also details how many arrays appear in the "Development" cohort, where applicable.

We move now to user generation of networks. The StarNet user enters a gene ID and several other parameters which influence both the topology of the network, and how it is visualized. Our script builds networks, and displays them on a results page. Additional information concerning the networks, discussed below, is displayed. We remark as well, and this is discussed further later, that the user can draw more than

one network at a time.

It is our aim that a user can examine the network, and compare two networks generated (if two networks were generated). User knowledge, as well as a supplemental graph indicating known genetic relationships gleaned from the literature, enhance the information contained in the graphs.

2. STARNET SUBMISSION PAGE

2.1. Help and Tools. We provide several resources online to assist users.

Documentation

Clicking on this link will take the user to a help page, which contains an FAQ, this manual, and an email address to which one can send questions regarding StarNet. Lists of species studied, arrays used, array samples used, numbers of genes on the arrays, numbers of arrays within each cohort, etc., are also available.

Gene ID Lookup

While a user may have a specific gene which they would like to explore, they may not know the Entrez gene ID or official symbol for the gene. Alternately, a user may be interested in exploring all genes annotated with a particular GO term, or with GO terms containing a specific word. Clicking on this link will take the user to a page where a very general search can be performed. Upon entering a search term, four searches are carried out on our server:

- (1) We search the descriptions field for each gene database entry in Entrez Gene; all genes whose description contains the user search term are returned.
- (2) We search the GO database for all GO terms which contain the user search term. Then we determine which genes are annotated with the discovered GO terms, and return those genes.
- (3) We search the Entrez gene database for that gene whose official symbol is the user search term. That gene is returned.
- (4) We search the Entrez gene database, looking at all genes, and find those whose list of synonyms includes the user entered search term. Those genes are returned.

In all cases the Entrez gene ID, official symbol, unofficial synonyms and description fields from Entrez Gene are displayed on the results page. The results from the GO search contain three additional fields: GO category, GO term and GO ID, all obtained from the GO website.

Only genes for the species we are studying are returned.

2.2. Network Specification (Basic Parameters). The choices which the user makes in selecting the parameters below have direct influence on the topology of the network which is produced, which genes are included in the network, and which correlations are included. These choices determine which central (seed) gene is selected,

which type of network is drawn, and which distribution of correlation coefficients is used as the pool from which the edges in the network are drawn.

Before reading any further, it is advisable to read Sections 4 and 5 for explanations of the various types of networks which may be built, and the distributions from which correlations may be taken.

First Cohort

The user chooses which species, and which subset of that species, is to be examined.

This field cannot be left blank; if it is the program will terminate.

Enter a gene symbol or Entrez Gene ID that corresponds to the first cohort

The user enters an Entrez gene ID, or official symbol/synonym, depending on the selection made in the next drop down list. This gene is the seed gene: the central gene for the network to be drawn. The gene of interest must be a gene from the species named in the first cohort.

Note: If the user enters an Entrez gene ID, the server searches for that gene on the array specified by the choice of the first cohort. If the gene is not on the array, or the gene has been discontinued or deprecated by Entrez, the program terminates and informs the user that the gene is not an appropriate choice.

If the user enters a symbol/synonym, the same check is carried out. Thus, if the gene is not on the array, or deprecated, the program terminates. If the parameter entered is a synonym, it is replaced in the execution of the rest of the program, and on the resulting graphs, by the official gene symbol. Certain synonyms represent more than one gene; if such a synonym is encountered, the program once again terminates with a warning.

Gene symbol/Entrez gene ID

The user chooses whether the ID entered above is a **Gene symbol** or an **Entrez gene ID**. Note that unofficial synonyms can also be entered if the user has chosen to enter gene symbols.

The default is for gene symbols to be entered.

Second cohort (optional)

The user may also choose a second species/subset to graph. The species need not be the same as that of the first cohort. Indeed, it may be of particular interest to compare the network for a gene within one species, and the corresponding network for that gene's homologue within another species.

The second cohort may be left blank.

Choose a sub-distribution of correlation coefficients

The user chooses one of the four possible types of distributions from which to pull the correlations which appear in the graph. The choices are: **100K highest magnitude from tail(s)**, the extreme tail with 100,000 coefficients; **Genecentric**, the

genecentric distribution; **Transcription Genecentric**, the transcription genecentric distribution; or **Transcription and Signal Genecentric**, the transcription and signalling genecentric distribution. The details of each distribution are given in Section 5.

The default setting is genecentric.

The number of connections each gene should make

Networks are built by making connections based on correlations, gene to gene. This parameter determines how many connections are to be made from each gene.

The default is five connections.

The number of levels (steps from the central gene) that should be drawn

The user does have several choices for which type of network to draw, with an additional choice as to which distribution correlations should come from. However, once these choices are made, the method of graphing is relatively consistent. The seed gene is the starting point of the graph, and is considered to be level 0. The seed gene is connected to various neighbours (as many connections as specified by the previous parameter), depending on the distribution chosen and the graph building method. These neighbours are the level 1 nodes. These nodes are connected to further nodes, the level 2 nodes, etc. The user chooses here which level is to be the highest level.

The default for number of levels is two.

2.3. Network Specification (Advanced Parameters). The following parameters need not be set, as all have default values. They are hidden in a drop down list, until the user accesses them.

Network type

The user chooses to draw the network using one of the five following functionalities: **Levels**, **Levels with Internal Edges**, **Weight**, **Highest**, or **Highest with Internal Edges**. The details of each methodology are explained in Section 4.

The default setting is highest with internal edges.

Correlations to view

We recall that for each distribution we built a negative, positive, and combined positive and negative collection of genes (see Section 5). Here the user selects one of **Both positive and negative**, **Positive only** or **Negative only**, indicating which “tail” of the distribution they will use. The choice is between examining only genes which have an up regulatory effect on each other, only genes that have a down regulatory effect on each other, or both.

The default is both positive and negative correlations.

Note: If examining only negative coefficients, some care must be taken with interpretation. That genes A and B have a negative correlation, and B and C have a

negative correlation, indicates that A and C are positively correlated. This assertion can of course be checked by changing the parameters of the graph and trying to find the specific relationship between A and C.

Highlighted GO terms (comma separated)

The user enters here a term of interest. For each gene in the graph, the GO annotations for that gene are examined. If a GO term annotating a gene contains the user entered search term, the gene is flagged, and its name appears on the graph in red.

The default search term is “transcription”. The user may enter more than one term to be searched for; these search terms must be comma delimited.

2.4. Visualization. The following choices do not influence *which* network is drawn, but rather *how* it is drawn. Here the user can choose the visual style and look of the network. The following parameters need not be set, as all have default values. They are hidden in a drop down list, until the user accesses them.

Draw network with

The user chooses whether to draw the network with **Gene Symbol** or **Entrez gene ID**. Each gene in a network appears as a node in a graph. Each node is labeled with an ID for the gene which it represents. The choice here is whether that label will be the Entrez gene ID or the official gene symbol for the gene in question. The default is for the gene symbol to be used as the label.

Note: it does not matter whether the user entered the seed gene as an Entrez ID or symbol; regardless of that choice, genes can be labeled with either an Entrez ID or symbol. Additionally, if a gene which appears in the network has been deprecated by NCBI, or its record has been discontinued, the gene is represented by its Entrez gene ID with an asterisk appended. This is done regardless of whether the user asked that the graph be drawn with Entrez gene ID or symbol.

Node type

The user chooses whether each node in the graph will be drawn as a **Box** or **Ellipse**. This choice is merely one of preference and aesthetics; which shape produces a graph that is easier for the user to examine, or that the user finds more aesthetically pleasing. The default is for ellipses to be used.

Node style

The choice here is whether the nodes will be drawn as **Filled shapes** or **Outlined shapes**. That is to say, will the nodes be drawn as boxes/ellipses of a solid color, or will they be drawn as white boxes/ellipses with only the edges coloured. This choice is again one of aesthetics, and ease of use. The default is for the nodes to be filled.

Use brackets to indicate genes that appear in both cohorts

Certain genes (specifically, the seed gene) may appear in both networks, if the user chose to draw two networks. While these genes are flagged in the tables provided

below the graphs, it can be also be useful when visually examining the networks to have a cue for identifying those genes which are common to both networks. That cue is to surround the label of each common gene with brackets (e.g., “[OOG3]”). Here the user decides whether this cue will be provided (**Yes**) or not (**No**). The default is for the brackets to appear.

Further explanation of determination of common genes in cross species networks in given in Section 3.

Label edges with correlation coefficients

Each correlation coefficient in the network is represented by an edge in the graph, and each edge in the graph represents a correlation coefficient. While there is a scale provided below the graphs giving a rough indication of the value of the correlation represented by an edge, and there is also a table provided which lists each edge and corresponding correlation coefficient, there is no direct way to read correlation coefficients off the graph. We offer users the ability to draw the network with the edges labeled with their correlation coefficients. The user chooses either **Yes** or **No**. The default is for the edges not to be labeled; this choice was made in order to avoid cluttering in the graphs, especially larger ones.

Draw edges using

In drawing graphs, the Graphviz program follows a prespecified algorithm. We have chosen to draw our networks as concentric circles, each circle being one level of the network. While Graphviz is extremely efficient and produces the graphs which we desire, its basic functionality has one drawback. Edges are drawn as straight lines, with no care as to whether these lines are drawn over other nodes. In other words, in drawing the edge between node A and node B, it is entirely possible that the edge passes through node C, just as a side effect of the relative placement of the nodes, and not as any indication of relationship between node A and node C, or node B and node C. Not only can this clutter the graph and make it hard to read, it can cause confusion, if the user mistakenly sees a relationship which does not actually exist.

There is a way around this difficulty: drawing edges with splines instead of straight lines. Without going into too many mathematical details, a spline is simply a smooth curve passing through selected points. The splines drawn by Grahviz are chosen so as to avoid intersecting intermediate nodes (such as node C in the example above); each edge touches only the two nodes which it is supposed to connect. The net result is a graph which is cleaner, easier to read, and with no possibility of the confusion mentioned above. Drawing splines, however, does involve calculation of the curves, and thus more processing time.

The user chooses either to draw the network with **Splines (nicer)** or **Straight lines (faster)**. The default is for the network to be drawn with splines.

Note: If any node has fifteen or more edges coming out of it, and the user has to chosen to draw the network with splines, the script will produce an error and stop. The user will be asked to draw the network with straight lines. This is done

as computing the splines will place too heavy a load on our server.

Color scheme for node level with respect to the central node

For ease of reading graphs, we color code the levels (level 0 = seed gene, level 1 = direct neighbors of the seed gene, etc.). The user has three choices for color scheme of the scale for levels: **Pastels**, **Brown/Blue** and **Purple/Green**. The pastels scheme is a selection of pastel colors. The purple/green scheme ranges from dark to light purple and then light to dark green. Similarly, the brown/blue scheme ranges from dark to light brown and then light to dark blue. The default is that the scale is pastels.

3. STARNET RESULTS PAGE

We now examine the result pages produced by StarNet, describing in detail the information which they contain. StarNet produces a main results page, from which the user can hyper-link to a page specific to each of the cohorts drawn. We first describe in detail the main results page, as the other two pages are very similar.

We remark that we are describing here the results page, not those warning messages produced if an error has occurred.

Central node

The top of the results page contains general information detailing the choices which the user made in the construction and visualization of the networks. In particular, the Entrez gene ID, official symbol, unofficial synonyms, and description from Entrez are listed. This is done regardless of whether the user entered an Entrez ID or symbol, and how they want the graph drawn.

Parameters

The user has entered many parameters on the submission page. We echo these parameters, with some supplemental information, here. Most fields are self explanatory, but a few deserve comment.

First, the “ID provided” field will always contain either an official symbol or an Entrez gene ID. The ID will be followed by either “[symbol]” or “[entrez]”. If an unofficial synonym has been entered, it will be replaced here by the symbol.

The distribution used is also reported. If the user has chosen any of the genecentric distributions, their names are simply reproduced: “genecentric”, “transcription”, or “signal”. If they have chosen the extreme tails distribution, this fact is indicated: “Top100”.

Notifications

This is a list of warnings to the user; nonfatal errors of which the user should be made aware. Two types of warnings may appear.

First, if an unofficial synonym has been entered by the user as a gene ID, and a suitable symbol has been associated with that synonym, the user is informed that

the synonym was not an official symbol, and has been replaced in all further analysis with the gene symbol. The symbol and Entrez Gene ID are supplied to the user.

Second, if a gene which appears in the network has been deprecated by Entrez, or its record has been discontinued by Entrez, the user is thus notified, and is supplied with the Entrez ID. The user is also told which of the networks the gene appears in. (Such genes will appear in the graph with an asterisk appended to their Entrez IDs.)

Graphs

The graphs are, in some sense, the heart of the analysis. One graph is produced for each cohort which the user has asked to draw. (For specific details on graph construction see Section 4.) Each node in a graph represents a gene, each edge a correlation between the nodes (genes) which it connects.

The levels in a graph are colour coded for ease of identification, and the levels are organized as concentric circles. It is thus often easy, looking at a graph, to see the direct neighbours of the seed gene, lying in a circle around the seed gene. The neighbours of these genes lie on a larger circle containing the first.

Nodes are labeled either with Entrez gene IDs or official gene symbols, depending on the user's specifications. If a gene in the network has been deprecated, or its record discontinued, by Entrez, its node is labeled with the Entrez ID followed by an asterisk.

The edges (correlations) are themselves colour coded. Darker correlations are stronger, lighter correlations weaker. Positive correlations are drawn as blue, negative as red. Each edge is drawn from one node to another, as an artifact of the network building process. That is to say, we start with the seed gene. From it we draw edges to its direct neighbours. The fact that these edges are drawn from the seed gene to the neighbours, and not the other way around, is not an indication of causality; it is merely due to the algorithm being used. It is possibly useful to the user to know which direction edges were drawn; the edge drawn from node A to node B has a small ball attached at the B end. (If there is a ball at both ends, then it is the case that the edge was drawn twice, once in each direction.) If the user so specified, the edges are labeled with the correlation coefficients to which they correspond.

If so specified, genes common to both graphs are represented with square brackets around their labels.

If any gene in a graph is annotated with a GO term containing a user specified search term, that gene's label is written in red.

The other parameters which the user set and affect how the graph is visualized will, of course, have an effect on the appearance of the graph. Most of those parameters which we have not mentioned have self explanatory influence on the graph.

Scales

Directly below the graphs are the accompanying scales. We provide one scale for positive and one scale for negative correlations, for each graph. Note that the scale is not a fixed absolute scale; in particular it does not run from 0 to 1 (or -1 to 0 for the negative correlations). The scale is determined on a per graph basis; given a

graph the positive scale does not run from 0 to 1, but rather from the value of the smallest correlation in the graph to the value of the largest correlation in the graph.

There are two exceptions to this rule. First, if there is only one correlation, we use an absolute scale, running from 0 to 1 (-1 to 0 in the negative case). The second exception is if all the positive (negative) correlations within a graph are the same.

We remark as well that there are several situations where no scale at all is drawn. If the user asks that a graph be drawn using only positive coefficients, then no scale for negative correlations is drawn. If they do ask for both positive and negative correlations, but there are no negative correlations in the graph, no scale for negative correlations is drawn.

Next is the level scale, common to both graphs. While it is often not difficult to visually differentiate between the levels in a graph, the scale is a further visual aid.

Known interactions: graphs

Certain of the genes which appear in our graphs will be well known and well researched. It is thus quite possible that certain interactions between these genes and others which may or may not also be in the graph will be known. We provide the user with this information, derived from Entrez's Reference into Function (RIF) file, which can be found on their FTP site.

The information is presented first in graph form, one graph for each cohort. Again, each node represents a gene, but in these graphs each edge represents an experimentally verified interaction. Nodes and edges are labeled to visually provide information to the user. Nodes whose names appear blue represent genes which appear in our correlation derived graphs.

We document several types of interactions, detailed below, with corresponding labeling explained.

- (1) Protein-Protein: edges black, no arrowheads.
- (2) Protein-DNA: edges red, arrowhead pointing at the DNA node.
- (3) Protein-RNA: edges blue, arrowhead pointing at the RNA node.
- (4) RNA-RNA: edges red, no arrowheads.
- (5) RNA-DNA: edges red, the DNA head of the edge is a diamond.
- (6) DNA-DNA: edges red, both ends of the edge are dots.

We remark that we have determined that a gene represents a RNA, DNA, or protein, respectively, if the annotation in the gene RIF is (NM or XM) and (NT or NC), or (NP or XP), respectively. Any other annotations are labeled as proteins. We have chosen to identify the molecules in this fashion based on RefSeq's notation for accession numbers. (See <http://www.ncbi.nlm.nih.gov/RefSeq/key.html>.)

Note: For each of the lists described below we provide one list for each cohort.

Gene list

We provide a list of all the genes which appear in the graph. The Entrez gene ID and official symbol are hyper-linked to the Entrez Gene page for the given gene.

Unofficial synonyms and descriptions from Entrez Gene are also provided. Last, a column with either “Yes” or “No”: yes the gene appears in both graphs, or no it does not. (This column is left blank if only one graph was drawn.)

If a gene has been deprecated, or its record discontinued, by Entrez, the symbol does not appear, but is replaced instead with the Entrez ID followed by an asterisk.

Edge list

This is a list of all of the edges (correlations) in the graph. For each edge the two genes connected by the edge are indicated. The symbol for each is given, and the Entrez ID for each is placed in square brackets and hyper-linked to the Entrez Gene page for the given gene.

For each edge the correlation coefficient, and 95 and 99% confidence intervals are given.

Note that if an edge is drawn from node A to node B, and an edge is also drawn from B to A, the correlation will appear in the list twice, once as “A [A] - - - B [B]” and once as “B [B] - - - A [A]” .

Again, if a gene has been deprecated, or its record discontinued, by Entrez, the symbol does not appear, but is replaced with the Entrez ID followed by an asterisk.

Known interactions

We provide a list of each of the literature documented relationships. The first column contains the gene symbol, and hyper-linked Entrez Gene ID of the first gene in the relationship. The second column contains the Entrez Gene description of the gene. The third and fourth columns contain similar information for the second gene. Next we list the interaction type, followed by the gene RIF (the description of the interaction), the PubMed reference (hyper-linked to PubMed), and finally the source database (either BioGrid or Bind).

Matches for GO search term(s)

We provide a list of those genes annotated with GO terms containing the user provided search term. The first column contains the official gene symbol (or the Entrez ID with an asterisk if the gene has been deprecated or discontinued). The second column contains the Entrez ID hyper-linked to the Entrez page corresponding to the gene. Finally, the third column contains a list of the GO IDs for those terms which contain the user search term, and annotate the given gene. The GO IDs are hyper-linked to the appropriate GO entry on the GO website.

Enriched GO terms

We provide a list of all the GO terms which are enriched in our graph. The first column contains the GO term, preceded with an asterisk if more than one gene in the graph is annotated with that term. The second column contains the GO ID, hyper-linked to the GO entry corresponding to that term. The p-value and Bonferroni adjusted p-value calculated for the given GO term are provided; only terms with adjusted p-values less than .05 are included. Finally, the last column contains

a list of all the genes in the graph annotated with the given term. This list has symbols (replaced with an asterisk in the case of deprecation or discontinuation), and, in brackets and hyper-linked to Entrez, Entrez IDs.

Genes common to both networks

In drawing graphs for cohorts from different species, common genes are identified as follows: A gene is considered to be common if it appears in the one cohort, and its homologue in the species of the other cohort appears in the other cohort. If both cohorts are from the same species, the calculation of common genes is done in the obvious fashion. Each common gene gets one row, containing its Entrez IDs and symbols in both species. These are again hyperlinked to Entrez Gene, and deprecated genes are again flagged with an asterisk.

The common gene table appears only if two cohorts were drawn.

HeatSeeker

To make further analysis of gene coexpression visually easy, we include a link to the HeatSeeker utility. This option is only available if two cohorts were drawn.

On clicking the HeatSeeker button, the HeatSeeker utility carries out the following analysis. First, it finds all genes in each of the two networks which have a homologue in the other. Next, from these lists of homologues, a list is constructed containing all homologues which appear in either network: the union of the two lists just built. We only examine those genes which appear on the appropriate arrays (those corresponding to the networks), as these are the only genes for which we have expression values. We disregard all other genes.

For each network we examine each gene in the list, and pull out its expression values in the corresponding cohort. With this information we build a false color map representing one minus correlation of expression values between the genes in the list within that cohort. (We compute one minus correlation, in order that this measure is zero when comparing a gene to itself.) We also build a false color map for the differences between correlations within the two cohorts. These three heat maps are displayed on a new output page.

Further, for each cohort, using one minus correlation between expression values, and the complete linkage algorithm, we carry out hierarchical clustering of gene expressions, and output the associated dendrograms.

3.1. Cohort Specific Results Pages. The amount of information available on the main results page is, perhaps, overwhelming. A user may be interested in examining one (or both) of the two cohorts in isolation. This option is available. By clicking on the graph for a given cohort, the user is taken to a page specific to that cohort. The graph and lists for that cohort are displayed, those for the other cohort are not. Additionally, the graph is drawn much larger, and each node on the graph is hyper-linked to the corresponding entry in Entrez Gene.

The layout of the cohort specific results page is the same as that of the main results page. Information regarding the central gene and the user specified parameters is

provided. The graph is drawn, followed by the scales for the correlations and the levels. The known interactions graph appears. Finally the tables (genes, correlations, known interactions, GO search term matches, GO enrichment) appear.

The warning messages (Notifications) are not, however, reproduced.

4. NETWORK TYPES

Throughout this manual we use the words “network” and “graph” almost interchangeably. In our setting a graph is a collection of nodes (for our graphs these are genes) connected with edges (in our case, correlation coefficients). In other words, drawn on a piece of paper or a computer screen, a graph looks like a collection of dots, some of which are connected to each other with lines.

The star networks which we draw are a simple class of graphs, with many options for the user. In these networks we are examining only one gene (the seed gene), and we are looking to find all the genes to which it is connected.

There are three methods to build a star network.

Levels

Simply include all correlations from the seed gene. This is the first level. On the second level, connect these direct neighbours of the seed gene to their direct neighbours. Repeat the procedure. There is also a variant: **Levels with Internal Edges**, explained below.

Weights

Essentially the same as above, except that we only allow a new gene to be part of our network if the product of the correlation coefficients connecting the seed gene to the new gene is greater than a user specified cutoff.

Highest

Again, essentially the same as ‘Levels’. In this instance, though, we only allow the n highest correlation coefficients, where n is user specified. There is also a variant: **Highest with Internal Edges**, explained below.

There are two important comments to be made. First, concerning the Weight functionality. In building the network in this case, as we are building a given level, we prepare for the construction of the next level by recording for each gene added to this level the weight to that gene. That is, for each gene which we connect to on the current level, we record the product of the correlation coefficients connecting the seed gene to our new gene. As we process the current level, we may add a gene to that level more than once. This amounts to finding more than one path from the seed gene to a gene on the current level. The weights of the different paths may be different. We record the highest weight. Here is an example.

Our seed gene is A, and it is correlated to B and C with correlation coefficients .9 and .9, respectively. B and C are correlated to D with correlation coefficients .9 and

.8, respectively. The two possible paths from A to D are:

A to B to D, with weight .72,

A to C to D with weight .81. D will be recorded as having a weight of .81.

The second comment concerns connections from higher to lower levels, as well as those within levels. There is a contrast between the standard functionality and the **Internal Edges** variants of Levels and Highest.

In the basic network building (Levels), there will be no connections from higher to lower levels. If you were to need to connect to a lower level, you would already have been connected to from the lower level. (If a gene on a higher level has to connect to a gene on a lower level, then that connection has already been made, from the lower level.) There is, however, the possibility that a gene may be connected to another gene on the same level. In the basic network functionality this is not allowed; these edges are simply not drawn. In the Levels with Internal Edges functionality these edges are drawn. As an example, consider the following simple network, with seed gene A:

A B .9

A C .9

B C .9

B D .99

In the standard functionality the edges are: A-B, A-C, and B-D. In the Internal Edges functionality the edge from B to C (and the edge from C to B) is also drawn.

Note that with the Highest functionality, we can potentially feed back to our ancestors. Here is an example. Our genes are A through F, correlated as follows, with seed gene A:

A B .9

A C .8

A D .7

B C .95

B D .91

D E .6

D F .5

We consider A as our seed gene, and use the two highest coefficients. A is connected to B and C. B is an ancestor of D, so that A is as well. D should connect to A, because that connection to A is one of its two highest coefficients. This is not allowed by the script, as A is D's ancestor. The two connections from D are to E and F.

In a similar vein, two genes on the same level may not be connected to each other. In our example there is only one connection from B - to D. Thus the edges drawn are: A-B, A-C, B-D, D-E, and D-F.

In the Highest with Internal Edges the two connections from D are to A and B. Similarly, there are two connections from B: to C and D. The edges here are: A-B,

A-C, B-C, B-D, C-B, C-A, D-B, and D-A.

Effectively, the upshot of the above is that in the basic functionalities a gene can be “connected to” multiple times, only in the following way: It has several (direct) parents. In other words, all connections to it are made from only one level, the level immediately below its own.

We remark that the user can tell by looking at the graph where the edges come from. Edges are recorded in the network file with the parent node sitting in the first column, and the child sitting in the second. The script sees this, and draws the line from parent to child with a ball at the child end.

5. DISTRIBUTIONS

In the above discussion of network building we did not mention where the correlation coefficients came from, we simply said that they were chosen. Implicit, then, is the fact that in building a network we are choosing the correlation coefficients which become edges from a specific subset of the millions of correlation coefficients which have been computed between the genes on our array. We choose these subsets for two reasons. First, the number of correlation coefficients between all the genes is far too large to be manageable, and far too large to be searched in any efficient way. Second, working with different subsets of the correlation coefficients from which to select those coefficients which appear in the graph imparts different biological meanings to the resulting graphs. In other words, the graphs generated for a given seed gene, but using different distributions, will give the user very different pictures of genetic relationships.

This section details the various subsets (distributions) which we have chosen. We describe in each section choosing coefficients from a generic distribution. In reality, of course, we choose our coefficients from the correlations computed from various cohorts of arrays.

The “Extreme Tails” Distributions

These are the simplest distributions. Let us assume that we have the list of all the pairwise correlation coefficients between all the genes on the array. We select from these coefficients the 100,000 largest positive coefficients. Similarly, we choose the 100,000 largest (in absolute value) negative coefficients. Thus far we have two distributions. We note that in each of these distributions we have only positive, or only negative, correlations. We finally combine the positive and negative tails, creating one more distribution.

The Genecentric Distributions

The extreme tails distributions certainly captures the most extreme behaviours, and the most extreme correlations. Many genes, however, will not have such strong

relationships; their correlations will be too small to appear in any of the extreme tails built above.

To remedy this, we build gene-centric distributions. For each gene we find three groups of correlations:

- (1) the ten largest positive correlations which it has to any other genes,
- (2) the ten largest negative correlations which it has to any other genes, and
- (3) the union of the two above sets of correlations.

This gives us three distributions: positive connections from each gene, negative connections from each gene, both positive and negative connections from each gene. Now we are guaranteed that whatever gene a user may be interested in, that gene will appear in at least one of our distributions.

The Transcription Gene-centric Distributions

We build another gene-centric distribution. This time, however, we only consider correlations between genes, both of which are GO annotated with a term containing the word “transcription”.

The motivation is to focus attention on those genes implicated as transcription factors.

The Transcription and Signalling Gene-centric Distributions

We build yet another gene-centric distribution. This time we only consider correlations between genes which are GO annotated with a term containing the words “transcription” or “signal”.

The motivation is to focus attention on those genes implicated as transcription factors or with connections to signaling.

6. IMPLICATIONS OF NETWORK CHOICES

Here we explore the biological and logical significance of choosing the different distributions, as well as the different network types.

It is our feeling that the most useful choice for network type and distributions is Highest with Internal Edges in conjunction with the Gene-centric distribution. These choices will give rise to the network closest to the actual biological network found *in vivo*. This is the case for the following reasons. Choosing one of the “extreme tails” distributions limits what will appear in the graph. Indeed, certain genes do not appear in these extreme tails. This does not mean that the relationships which they have with other genes are not important; they simply are not the most extreme of relationships.

While the Highest methodology for graphing produces a larger neighbourhood, the examples provided in Section 4 indicate that this can come at the cost of not actually including the highest correlations from a gene. The Highest with Internal Edges methodology captures all of the largest edges out of each given gene in the network.

Why have the other options been included, if our feeling is that the above choice is, in some sense, the best? Each of the other choices has its advantages and disadvantages, each has its own utility. The two variants of the gene-centric distribution are included for obvious reasons. We were originally interested in the gene regulatory network of murine cardiac development. Key players in genetic regulation are transcription factors. This option allows us to examine the interplay between the transcription factors, in isolation. The fact that the distribution is gene-centric guarantees that each transcription factor on the array is considered. The user may be interested in broadening their search slightly, allowing interactions between transcription factors/signalling proteins and other transcription factors/signalling proteins (or the genes which code for them).

As to the extreme tails distributions, these are useful in aiding the user in identifying those connections to a given gene which are, in relative terms, strong. That is to say, while the gene-centric distribution gives the largest connections to a gene, it gives no notion of how those highest connections compare with the rest of the correlations on the array. If a given gene has connections in the extreme tails, those relationships are guaranteed to be rather strong even as compared against the strongest relationships on the array.

We note that 100,000 is a very small percentage of the entire set of correlations. Relatively few correlations, and correspondingly relatively few genes, appear in these tails. In fact, given the scale free topology that many biological networks have, and the fact that in such networks the characteristic path lengths are relatively short, while clustering coefficients are rather large, this is not at all surprising. In a sense, then, the extreme tails may be useful in identifying the “hub” genes, those few highly connected genes central to the network.

While the other network building methodologies build neighbourhoods in levels, where distance is measured in the number of steps between one gene and the next, the Weights methodology measures distance in terms of correlations. In other words, a very different type of neighbourhood is built. We get a feeling for those genes which are most closely connected to the seed, rather than those which can be reached by few intermediate steps. (Note that while a gene may be close in weights, it may be very far away in terms of connections. As an example, a gene cascade may have many levels, all of which are very tightly connected: the farthest upstream and downstream genes are far apart in terms of levels, but close in terms of the product of the correlation coefficients along the path connecting the two genes.)

The Levels and Highest methodologies with Internal Edges are perhaps more biologically accurate than those methodologies without internal edges. However, the graphs which are drawn with internal edges may be more difficult to read. This is not surprising, as these graphs may well contain feedback loops. Drawing the graphs with internal edges creates a dense network. In some sense more edges are drawn; we allow feedback and edges within levels. With the highest methodology, as noted

in the examples presented in Section 4, we may actually end up with fewer edges if we allow internal edges. This is so as the highest edges from a gene on a higher level may feed back to a lower level, duplicating edges which already exist. This means that one fewer new edge is drawn. Having said this, however, each edge drawn from a gene is actually amongst the highest correlations involving that gene. This is in contrast to the Highest method without internal edges. Here the edges drawn may not actually be among the highest involving the gene. See, again, Section 4.