

Guide for Bioinformatics Project Module 2

Genetic and Physical Interactions and Expression Data

In the first module we learned the basic information about your gene, where in the genome it is located, what size of protein it is predicted to encode and if there are any conserved domains present in that protein. Before we investigate more properties of your gene and/or protein itself we want to try to get an overall picture of the function of your protein. We started this hunt by doing BLAST searches in the last module, looking for proteins of similar sequence in other organisms and then inferring a function for your gene based on what is known about the function of these orthologs. *Another way to try and predict function is to look at who your protein works with in order to do its job [physical interaction] (what proteins actually binds each other in the cell); to look at when your gene is mutated who exacerbates a phenotype [genetic interaction] (what other gene mutations make it worse); and to look at what genes are expressed or repressed in similar patterns to your gene [expression and response patterns] (what genes show up and leave the scene at the same time).* Once these other proteins/genes are identified we can determine what is known about their functions, thereby creating a list of possible functions that your gene might be involved in. Then, as we then look into more properties of your protein we can narrow down this list by determining if they support the functional role we are predicting, i.e. if you find out your protein interacts with other proteins that bind to DNA it is reasonable to expect that your protein will localize to the nucleus.

In this module GeneMania will allow us to assess your gene in the context of other genes, starting by searching for genetic interactions, then examining physical protein interactions, co-localization and some co-expression data. SPELL data will allow us to study gene expression data generated from microarrays (explained below) in response to various stimuli.

GeneMania

GeneMANIA helps you predict the function of your favorite (assigned) genes and gene sets. GeneMANIA finds other genes that are related to a set of input genes (or in this case your gene of interest), using a very large set of functional association data. Association data include protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. We will specifically be using GeneMANIA to search for any genes that have a connection to yours, but this software can also be used to find connections between a set of genes if you were on this path with your research. Your question is defined by the set of genes you input. If members of your gene list make up a protein complex, GeneMANIA will return more potential members of the protein complex. If you enter a gene list, GeneMANIA will return connections between your genes, within the selected datasets.

Navigate to GeneMania at <http://www.genemania.org/>. In the drop down menu on the left (Find genes in), select “**S. cerevisiae (baker’s yeast)**” and enter your gene name into the **related to** search box and click Go.

You will be routed to page that displays the known genetic and physical interactions, pathway, co-expression, co-localization and protein domain similarity for your gene of interest. Before we discuss how to navigate and mine for data let’s first look at what all of these categories mean. GeneMANIA searches many large, publicly available biological datasets to find related genes. Few if any genes/proteins in a cell perform their function alone or without some back-up mechanism in the cell; this may take the form of a protein complex which could be detected via protein interactions or a signaling pathway in which case deletions of multiple genes can be detected by increasing detrimental phenotypes – and are thus said to have a genetic interaction. GeneMania searches previously published data for these types of interactions as well as for protein-DNA interactions, reactions, gene and protein expression data, protein domains and phenotypic screening profiles. Networks names describe the data source and are either generated from the PubMed

Guide for Bioinformatics Project Module 2

entry associated with the data source (first author-last author-year), or simply the name of the data source (BioGRID, PathwayCommons-(original data source), Pfam).

Below you will find a list of the types of data that will be returned from a GeneMania search – READ through each category to understand what this type of information says about the relationship between your gene and the others it is linked to. For instance, a physical interaction may indicate a protein complex and similar functionality of two genes/proteins, however co-localization will give you information about where your protein functions in a cell but does not necessarily mean that the function is the same as all other proteins found in the same location.

Co-expression: Gene expression data. Two genes are linked if their expression levels are similar across conditions in a gene expression study. Most of these data are collected from the Gene Expression Omnibus (GEO); GeneMania only collects data associated with a publication.

Physical Interaction: Protein-protein interaction data. Two gene products are linked if they were found to interact in a protein-protein interaction study. These data are collected from primary studies found in protein interaction databases, including BioGRID and PathwayCommons.

Genetic Interaction: Genetic interaction data. Two genes are functionally associated if the effects of perturbing one gene were found to be modified by perturbations to a second gene. These data are collected from primary studies and BioGRID.

Shared protein domains: Protein domain data. Two gene products are linked if they have the same protein domain. These data are collected from domain databases, such as InterPro, SMART and Pfam.

Co-localization: Genes expressed in the same tissue, or proteins found in the same location. Two genes are linked if they are both expressed in the same tissue or if their gene products are both identified in the same cellular location.

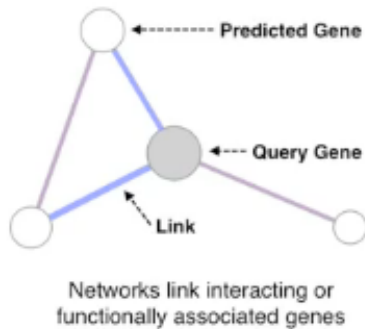
Pathway: Pathway data. Two gene products are linked if they participate in the same reaction within a pathway. These data are collected from various source databases, such as Reactome and BioCyc, via PathwayCommons.

Predicted: Predicted functional relationships between genes, often protein interactions. A major source of predicted data is mapping known functional relationships from another organism via orthology. For instance, two proteins are predicted to interact if their orthologs are known to interact in another organism. Also, GeneMania includes predicted functional associations from other groups that combine multiple data sources for a given organism, e.g., the entire YeastNet predicted network: Lee-Marcotte-2007 YeastNet; the genetic interaction data used to generate YeastNet: Lee-Marcotte-2007 Genetic interactions. In these cases, the network name indicates the original publication detailing the predicted network, and (in some cases) lists the individual network that was used to generate the entire predicted network (for latter example above).

Other: Networks that do not fit into any of the above categories. Examples include phenotype correlations from Ensembl, disease information from OMIM and chemical genomics data.

Data presented by GeneMania is displayed with classical node and edge mapping. Each node (or circle) represents a gene, each edge (or connecting line) represents a connection and the color represents the type of interaction or data that connects these two genes/proteins. For instance, co-expression data is represented by pink lines and co-localization is represented by blue lines

Guide for Bioinformatics Project Module 2



You can change the network visualization in many ways, which will ease the ability to focus on specific types or networks and create a less complicated picture. In the diagram to the left your query gene is linked to the predicted gene by co-localization data.

These data networks are most easily viewed one at a time. To change this setting simply move to the right-hand portion of the screen and select ONLY ONE (for example, Co-expression) data box at a time. Once you have selected one and only one type of data you can also manipulate the network by clicking any given node and dragging it to a new location in the left-side box.

For each of the data sets, individually create the network, copy and paste it into your Module 2 Worksheet, by data type. [Note – exporting the file creates a vector image file (.svg) that is best opened with Adobe Illustrator – if you do not have access to this program, using the Snipping Tool (Window) or capturing the image with Grab (Mac) is the best way to save and insert these network graphics.]

To learn more about the genes in any given network you can do several things: 1) Click on the relevant node and gene information will pop up 2) Change to the Genes Tab in the right side panel and gene information is displayed (again clicking a gene will display additional information **NOTE: when you flip to the Genes Tab ALL genes found any network are displayed NOT just the ones relevant for that given network – make sure you look at which genes are represented in the node and edge diagram before just copying all the gene information.** 3) Look up the gene name in SGD 4) By clicking the arrow next to the type of data displayed in the right side box the papers or program from which the data were mined or generated is displayed, you can find the article or program specific for the interaction of interest and read/run through to get more information.

As you look at each type of data, i.e. co-expression or physical interactions look up the other gene's functions and relate this information based on the type of interaction to create a hypothesis of your gene/protein's function. NOTE: This means that how you interpret data needs to be relevant to what network you are investigating – if two proteins interact that could mean they form a complex together to function, however if two proteins co-localize that may not being they are doing the same job at that location. Include summary information of the relevant genes and their functions in your Module 2 Worksheet.

SPELL

As mentioned above and seen in GeneMania's co-expression data – information that shows that two genes are expressed in similar patterns in normal cells or in response to some stimuli can be helpful in predicting function and give us at least TWO key pieces of information. **1. Expression in similar patterns can tell us that these proteins are needed at the same time, either at a point in the cell cycle or in response to, for example, some toxin. This information may lead us to a hypothesis that our gene has similar functions to those with which it shares expression patterns.** **2. Change in expression in response to a stimulus tells us the cell is either upregulating or downregulating your protein due to some treatment. For example, if we expose to a toxin and expression of your protein goes up you could hypothesize that your protein functions in breakdown of the toxin or in repair of damage due to the toxin; if however expression goes down it could lead you to hypothesize that your protein may function as perhaps a cellular pore that the cell needs to turn off in order to prevent entry of more of the toxin into the cell.** Many other hypotheses are possible, and we will hunt down more information through this project that might lead us towards one or another hypothesis as being most probable. We will look at both of these types of data with the SPELL data system and you will craft hypotheses according to the data you find and analyze.

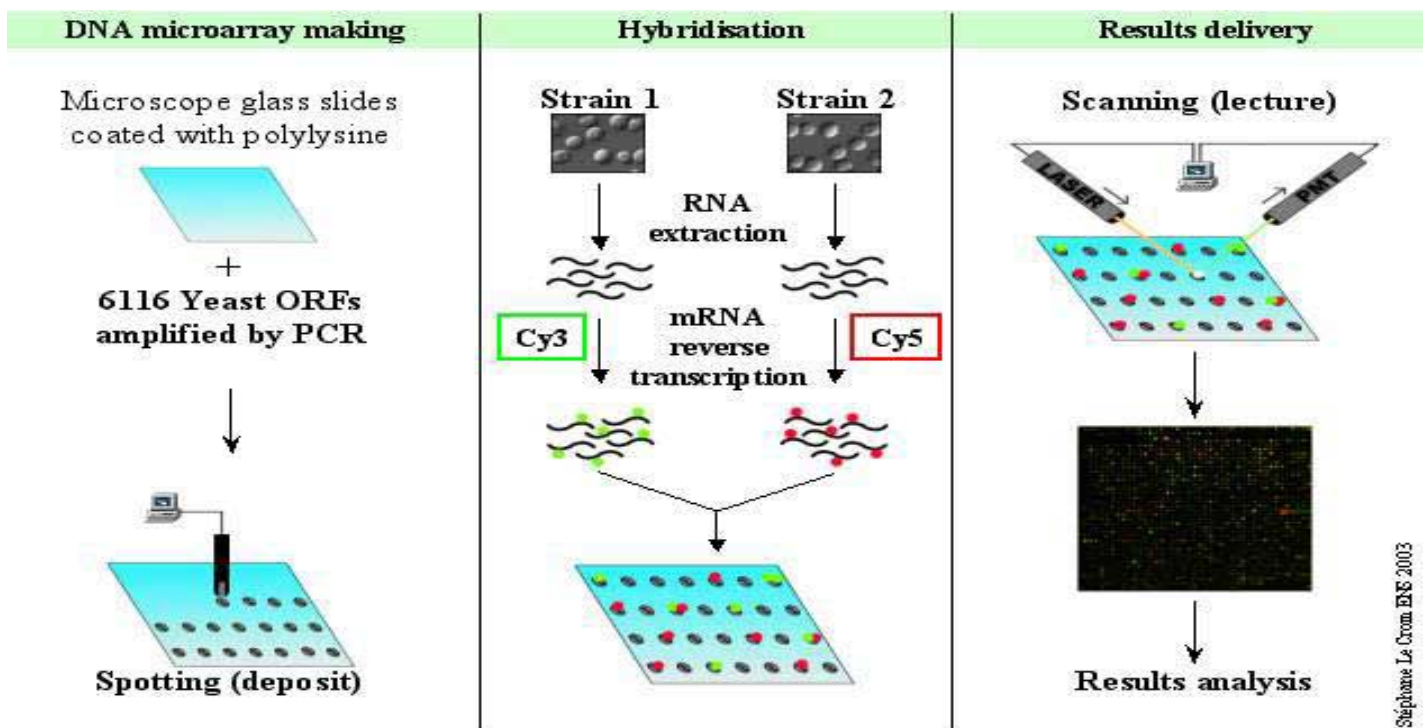
Guide for Bioinformatics Project Module 2

SPELL (Serial Pattern of Expression Levels Locator) is a query-driven search engine for large gene expression microarray compendia. Given a small set of query genes (even just one), SPELL identifies which datasets are most informative for these genes, then within those datasets additional genes are identified with expression profiles most similar to the query set.

Many of the experiments that are mined here are based on DNA microarray technology. This type of microarray is composed of a collection of DNA spots attached to a solid surface, usually some sort of glass slide. The microarrays (small arrays of DNA spots) can then be utilized to simultaneously measure the expression level of all of the genes in any given organism. This is achieved because each DNA spot (also known as a probe) corresponds to a gene or DNA element and can hybridize to cDNA or anti-sense RNA (also known as the target) that is expressed at any given moment in the cell. Cells can therefore be treated in a certain way, for example in the presence of a stimulus or chemical and then the cDNA and/or anti-sense RNA is harvested and hybridized to the probes on the chip. Probe-target interactions are then visualized by fluorophores or chemiluminescence such that any cDNA that is expressed and hybridizes to the microarray causes a signal to be released, the signal is detected and we can measure that this cDNA is expressed in response to the given stimuli. Furthermore relative abundance can be measured and in this manner tell us if the cDNA (corresponding to a gene) is expressed at high (**RED**), low (**GREEN**) or normal (**BLACK**) levels under the given conditions. A cartoon is included below to help explain how this data is generated and how the microarray process works.

To view an animation of a microarray setup (which may help you to understand the data) go to:

<http://learn.genetics.utah.edu/content/labs/microarray/>



<http://www.transcriptome.ens.fr/sgdb/presentation/principe.php>

Navigate to SPELL – *S. cerevisiae* at <http://spell.yeastgenome.org/> and enter your standard gene name in the Search box and submit the Search. (Leave the # Results at the default of 20)

Guide for Bioinformatics Project Module 2

You will be directed to a new page where microarray data is displayed for 10 experiments at a time (listed horizontally across the top of the chart). In this display the published dataset is hyperlinked at the top of each column, underneath of which the type of experiments is briefly displayed (Tags) and below this there is a description of the type genes that were found or the type of treatment that was applied (i.e ploidy regulation of gene expression or mitochondrial dysfunction). You will then see a red, green and black graphic chart with 20 gene names listed down the left side. Your gene of interest should be listed at the top, left position. As described above microarrays can be used to measure increases, decreases and no change results of expression of a given gene – these are represented as green (decrease), red (increase) or black (no change) colorimetric data in the microarray chart. If you click on any given colored bar – the numerical data and changes to stimuli will appear in a new window.

The first analysis we will do will search for genes that show similar expression profiles across different treatment parameters. Using the expression data shown for your gene across the top row, SPELL has searched and found 20 genes that show similar expression profiles to those seen for your gene, some may have similar profiles over just a few of the 10 experiments and other may be similar across all datasets displayed. ***As these data are collected across various experiments, from various researchers and the expression patterns have been observed to be consistent this repetition means this is strong data that your gene could be functioning in the same, similar or parallel ways as the other genes listed.*** Clicking on a gene name in the list will link you to that gene's SGD page to find out more description data on that gene's function.

Furthermore, below the microarray chart, SPELL has curated and is displaying the GO (gene ontology) data for this list of genes. This is a program that links proteins together by similarities in function or location – if a group, or GO Term, is displayed than you have an abundance of proteins that function in that category. These will once again be helpful in predicting the type of processes that the other genes, and therefore possibly your gene, are involved in.

[Note: If no datasets are found to contain a significant signal for a given query then SPELL is unable to assign per-dataset weights for the search. In this case a warning message is displayed, and all datasets are equally weighted for that query. If no genes are related to the query set at a reliable confidence level, then a warning message is displayed and the confidence level is weakened until results can be obtained. Both of these cases typically only occur when the query genes are either largely unrelated, or highly unique. Neither of these cases occurs very often.]

In your Module 2 Worksheet copy and paste the page with microarray data including the Dataset, Tags, Description data and the chart for all 20 genes displayed. Also copy and paste the GO Term Enrichment Data that is located at the bottom of this page into your Module 2 Worksheet. [Note – Snipping Tool (Window) or capturing the image with Grab (Mac) may be the best way to save and insert the table.]

Look at the known functions of the other genes in this list that show similar expression profiles. What processes or reactions are these genes consistently involved in? Are there multiple pathways? Which genes group together? Create a hypothesis (or hypotheses) about your gene function based on this data and include it in your Module 2 Worksheet.

The second analysis we can perform with this same data set involves researching the type of treatments that were performed that led to an increase or decrease in the expression profile of your protein. Treatments that lead to a shift to a green color mean your gene was downregulated and those with a red color mean your gene was upregulated. Take note of the 10 treatments that showed changes in the expression profile, which one's caused increases, which ones caused decreases? Research what the treatments were and what they are known to cause in the cell – for instance if you protein is said to increase in expression due to MMS treatment you will need to look up what the drug MMS is and how it functions. Your professor will be happy to help with this research portion of the lab if you don't know where to start, how to look up data or how to link the data together.

Guide for Bioinformatics Project Module 2

In your Module 2 Worksheet list the treatments that cause increases and those that cause decreases in the amount of your protein. Then search for links between the treatments, i.e. do all of the treatments that lead to an increase in protein production cause DNA replication stress? Use this data to formulate or reformulate a hypothesis (or hypotheses) about the function of your protein. Include this in your Module 2 Worksheet.