Reverse Engineering of Gene Regulatory Networks for Discovery of Novel Interactions in Pathways Using Gene Expression Data

Tanwir Habib

University of Southern Mississippi

The University of Southern Mississippi

REVERSE ENGINEERING OF GENE REGULATORY NETWORKS FOR DISCOVERY OF NOVEL INTERACTIONS IN PATHWAYS USING GENE EXPRESSION DATA

by

Tanwir Habib

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

[redacted]

August 2009

The University of Southern Mississippi

REVERSE ENGINEERING OF GENE REGULATORY NETWORKS FOR DISCOVERY OF NOVEL INTERACTIONS IN PATHWAYS USING GENE EXPRESSION DATA

by

Tanwir Habib

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2009

ABSTRACT

REVERSE ENGINEERING OF GENE REGULATORY NETWORKS FOR DISCOVERY OF NOVEL INTERACTIONS IN PATHWAYS USING GENE EXPRESSION DATA

by Tanwir Habib

August 2009

A variety of chemicals in the environment have the potential to adversely affect the biological systems. We examined the responses of Rat (*Rattus norvegicus*) to the RDX exposure and female fathead minnows (FHM, *Pimephales promelas*) to a model aromatase inhibitor, fadrozole, using a transcriptional network inference approach. Rats were exposed to RDX and fish were exposed to 0 or 30mg/L fadrozole for 8 days. We analyzed gene expression changes using 8000 probes microarrays for rat experiment and 15,000 probe microarrays for fish. We used these changes to infer a transcriptional network. The central nervous system is remarkably plastic in its ability to recover from trauma. We examined recovery from chemicals in rats and fish through changes in transcriptional networks. Transcriptional networks from time series experiments provide a good basis for organizing and studying the dynamic behavior of biological processes. The goal of this work was to identify networks affected by chemical exposure and track changes in these networks as animals recover.

The top 1254 significantly changed genes based upon 1.5-fold change and P< 0.05 across all the time points from the fish data and 937 significantly changed genes from rat data were chosen for network modeling using either a Mutual Information

network (MIN) or a Graphical Gaussian Model (GGM) or a Dynamic Bayesian Network (DBN) approach. The top interacting genes were queried to find sub-networks, possible biological networks, biochemical pathways, and network topologies impacted after exposure to fadrozole. The methods were able to reconstruct transcriptional networks with few hub structures, some of which were found to be involved in major biological process and molecular function. The resulting network from rat experiment exhibited a clear hub (central in terms of connections and direction) connectivity structure. Genes such as Ania-7, Hnrpdl, Alad, Gapdh, etc. (all CNS related), GAT-2, Gabra6, Gabbr1, Gabbr2 (GABA, neurotransmitter transporters and receptors), SLC2A1 (glucose transporter), NCX3 (Na-Ca exchanger), Gnal (Olfactory related), skn-la were showed up in our network as the 'hub' genes while some of the known transcription factors Msx3, Cacng1, Brs3, NGF1 etc. were also matched with our network model. Aromatase in the fish experiment was a highly connected gene in a sub-network along with other genes involved in steroidogenesis. Many of the sub-networks were involved in fatty acid metabolism, gamma-hexachlorocyclohexane degradation, and phospholipase activating pathways. Aromatase was a highly connected gene in a sub-network along with the genes LDLR, StAR, KRT18, HER1, CEBPB, ESR2A, and ACVRL1. Many of the sub-networks were involved in fatty acid metabolism, gamma-hexachlorocyclohexane degradation, and phospholipase activating pathways.

A credible transcriptional network was recovered from both the time series data and the static data. The network included transcription factors and genes with roles in brain function, neurotransmission and sex hormone synthesis. Examination of the dynamic changes in expression within this network over time provided insight into recovery from traumas and chemical exposures.

ACKNOWLEDGEMENTS

I would like to thank my wonderful family for all of their support. Mom, Dad, sister Shagufta Habib and brother Ata Habib: You have all been there for me during this entire process, cheering me along and supporting all the other life events that came in the way. I could not have finished this work without your strong support. I love you all.

My sincere appreciation goes to my committee chair Dr Mohamed Elasri and to my entire committee, Drs Youping Deng, Robert Diehl, Joe Zhang, and Jonathan Sun, who have given me many wonderful opportunities and never gave up on me. I have learned so much from them. Thank you for your valuable suggestions and help with my research needs.

The most learning and joy came about because of lab partners and classmates Arun, Ying, Venkata, Mehdi, Imran, Suja and Sammy.

I owe a gratitude to Drs. Edward J. Perkins and Natalia G. Reyero for providing me the experimental data to work on, valuable suggestions throughout this work, and guidance on writing manuscripts.

Thank you all.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure

LIST OF TABLES

Table

CHAPTER I

INTRODUCTION AND BACKGROUND

The increasing popularity of microarray analysis has been partially fueled by the frustration of biologists with limited technological tools they have had at their disposal for gaining a comprehensive understanding of complex biological problems. Although still not finally known, the number of genes in the human genome is estimated to exceed 40,000. These genes and their protein products determine biological phenotypes, both normal and abnormal. Although the reductionist method has enabled researchers to delineate a number of signal transduction pathways, it cannot yield a comprehensive picture of the systems under study. By allowing simultaneous measurement of the expression of thousands of genes, microarray technologies have marked a defining moment in biological research. Gene expression microarray technology has been shown across a wide range of fields including but not limited to biomarker discovery, predicting disease outcomes and response to treatments, assessing co-regulation via time course and/or dose-response experiments, and detecting molecular mechanisms and/or pathways associated with a particular disease state. Several factors must be considered throughout the experimental process to ensure that the correct information is extracted. From a statistical point of view, these consist of (i) choosing an appropriate experimental design to answer the question of interest, (ii) implementing an appropriate normalization procedure that adjusts for experimental effects so that expression levels can be effectively compared across biological samples, (iii) assessing differential expression via statistical methods that are capable of distinguishing meaningful biological changes in protein expression from random noise, and (iv) using tools for clustering and classification [1].

Toxicogenomic data such as changes in gene expression, protein levels, or metabolite levels may be used in risk assessment. Toxicogenomics, resulting from the merge of conventional toxicology with functional genomics, is the scientific field studying the complex interactions between the cellular genome, toxic agents in the environment, organ dysfunction and disease state. When an organism is exposed to a toxic agent, the cells respond by altering the pattern of gene expression. Genes are transcribed into mRNA, which in turn is translated into proteins that serve in a variety of cellular functions. Toxicogenomics through microarray technology offers large-scale detection and quantification of mRNA transcripts, related to alterations in mRNA stability or gene regulation. This may prove advantageous in toxicological research.

Microarray Design

In general, a microarray experiment starts with the acquisition of biological materials from which RNA is isolated. However, for many experiments involving clinical tissues, the process is more complex and special attention must be paid to quality control. Central purpose of most microarray experiments is to map gene expression in biological samples. Microarray experiments, whether utilizing one-channel or two-channel technology, are comparative experiments involving populations of measurements, with the end goal being to compare abundance of targets in complex populations [2]. In most microarray facilities, there are two types of microarrays that are generally produced: cDNA microarrays, in which the PCR products of cDNA clones are printed, and long-oligonucleotide (oligo) arrays, in which oligos of a certain length are printed. Because all subsequent experiments and data generation rely on the quality of the microarray slides, their production is critically important and requires the maintenance of rigorous quality control. In the glass-based microarrays, the targets are labeled with fluorescent dye Cy3 and Cy5. A further advantage is that two different fluorescent dyes, such as Cy3 and Cy5, can be used simultaneously, which allows two different samples to be directly compared on a single microarray. As DNA microarray experiments are becoming larger, involving larger number of samples and conditions, it is important to design experiments in the most efficient way in order to obtain precise estimates with minimized unwanted variations of the biologically important parameters. The most commonly used design is the so-called reference design, where each condition of interest is compared with samples taken from some standard reference. This design allows an indirect comparison between the conditions of interest. This approach uses 50% of the hybridization resources to produce a control or common reference signal of no intrinsic interest to the biologists. In contrast, a loop design compares two conditions via a chain of other conditions, thereby removing the need for a reference sample. In a *n*-array loop design, if one array fails all the contrasts are still estimable, where as in the reference design, all the contrasts that involve the condition in the failed array are not estimable anymore. Vinciotti et al. (2005) have done a comparative analysis of the two models. The study was conducted to compare the variability of estimates and the differentially expressed genes between the two models. It was found that the percentage of significant genes when using loop design was higher than when the reference design was used. It was also found that the square root of the average estimated variance of the contrast estimates for the loop design was lower than that of the reference design [3].

Microarray Data Analysis

Data analysis typically represents the last stage of a microarray experiment. It is at this step that biologically relevant conclusions are typically made. In a microarray experiment, there are many sources of variation. For instance, samples to be compared are not always labeled with the same efficiency. Samples to be compared on an array are not always mixed in equal proportions prior to hybridizations. There are certain systematic sources of variation, usually due to specific features of the particular microarray technology that should be corrected prior to further analysis.

Microarray data preprocessing contains three phases: quality control, within-slide normalization, and between-slide normalization. Within-slide normalization aims to correct dye incorporation differences, which affects all the genes similarly, or affects genes with the same intensity similarly.

The process of removing or minimizing such systematic variability is called *normalization*, and it is an important aspect of quality control in microarray data analysis. One way to remove a systematic intensity-dependent bias is to smooth the data with a locally weighed regression method, such as *lowess.* This method is useful for smoothing scatter plots to reveal the underlying patterns or structure and for identifying nonlinear relationships between log intensity (M) and log ratio (A). Lowess-based intensity-dependent normalization consists of simply subtracting the smooth curve from the original log ratio data.

$$\log_2 R/G = \log_2 R/G - c(A) = \log_2 R/[k(A)G]$$

where c(A) is the lowess fit to the MA-plot. Lowess scatter plot smoother performs robust locally linear fits (Figure 1).

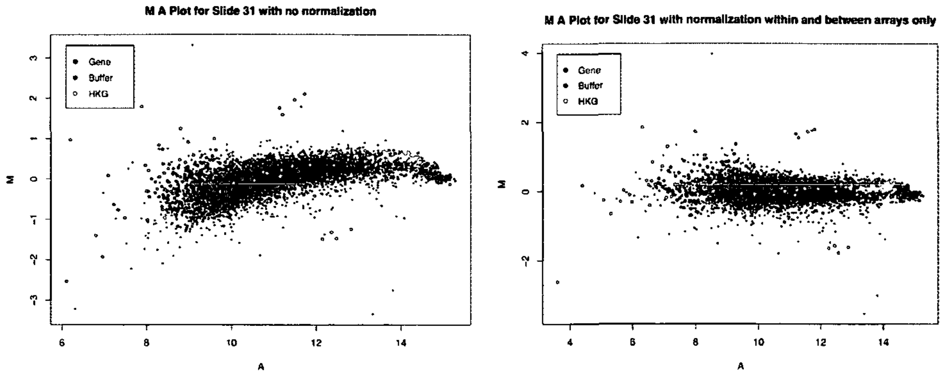


Figure 1: M vs A plot before and after the lowess normalization.

This normalization can be used in two-color array. An alternative approach is to use some subset of genes for the normalization, so-called housekeeping genes, such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH), b-actin (ACTB), tubulin a1 (TUBA1), and others. Housekeeping genes are expected to be expressed ubiquitously at stable levels under different biological conditions. Dye swap is another way to correct for systematic sources of variations within a two-color microarray. The idea behind this approach is to perform the same two-sample hybridization experiment twice, but with the dyes (CY3 and CY5) reversed. Dye swap experiments can result in significant increase in the cost and complexity of experiments, but they can potentially improve the quality of the results. Quantile normalization can also be used in single-color array to effectively remove the systematic biases.

Since a single microarray experiment represents an observation, multiple observations would be needed to compute a reliable estimate of the true transformed ratio values. Only a small number of replicate slides may be satisfactorily used to determine reliable estimates of true gene expression, and one study showed that three replicates suffice for significantly reducing experimental variability [4]. With the growing number of publicly available microarray data, conducting replicate experiments is becoming a popular solution to assess experimental errors and reduce noise bias in the measurements [5]. The advantages of replicate slides also greatly help the analysis of between-slide variability and help address formal statistical considerations when drawing biological conclusions.

Replicates

Replication is a basic principle in experimental design. It involves making independent observations under the same experimental conditions and carries tremendous implications for quality control. This issue is particularly relevant in the design of microarray experiments, where we can distinguish two different kinds of replication: intra-array and inter-array replication. Intra-array replication refers to measuring the same gene via several different spots on the same microarray. Inter-array replication refers to repeating the same hybridization experiment on several different microarrays. It is obvious that both inter- and intra-array replication can produce more consistent and reliable findings and increase the overall quality of the data analysis at the expense of increased cost and amount of biological material used. Intra-array replication is an important aspect of quality control since it can provide a more accurate estimation of the inherent variability in a microarray experiment and can also increase the probability of detecting differentially expressed genes, given that variability. According to the study conducted by Black and Doerge (2002), control-array was used, where sample is co-hybridized with itself using two different dyes, in order to obtain information about the sampling variation. The data are used in conjunction with ANOVA models in order to calculate the number of replicate spots necessary for detecting significant changes in expression with high probability. The residuals from a fitted ANOVA model can be used for power calculations [6]. The minimum number of intra-array replicates should depend on the minimum level of fold change required for detecting the differentially expressed genes with a high probability and make this determination using a power study framework for the particular microarray technology. One study suggest that at least two replicates are necessary for a reasonably high probability of detecting a threefold change on expression, while another study suggests that three replicates are necessary to ensure a high probability of detecting a twofold change [6]. Inter-array replication refers to repeating a microarray experiment more than once. Suppose that we are working with some cell lines and wish to perform microarray experiments under certain conditions. In order to produce replicated measurements, we could extract RNA from several different cell lines, cultured under as nearly identical conditions as possible, and perform microarray hybridizations using each of those different RNA samples. Or, we could extract RNA from one cell line, divide it into several parts, and perform hybridizations with each part. In the former approach, we will have to deal with additional experimental variability due to differences in cell lines and their respective RNA extraction steps. The latter approach is more informative about the particular cell line being used, but it cannot provide any knowledge of the population differences. If the replicates are concordant either in terms of intensities or log ratios and the genes are reliable, then we can simply combine the values of the replicates by averaging them to form a single estimate of the gene expression or log ratios.

Time Series Data

DNA microarray experiments are usually distinguished as static and time series experiments. In static expression experiments, a snapshot of the expression of genes in different samples are measured, while in time series expression experiments, a temporal process is measured. Another important difference between these two types of data is that while static data from a sample population are assumed to be independent and identically distributed, time series data exhibit a strong autocorrelation between successive points [7].

Gene expression is a temporal process. Different proteins are required (and synthesized) for different functions and under different conditions. Even under stable conditions, due to the degradation of proteins, mRNA is transcribed continuously and new proteins are generated. This process is highly regulated. One of the most important ways in which the cell regulates gene expression is by using a feedback loop. Taking a snapshot of the expression profile following a new condition can reveal some of the genes that are specifically expressed under the new condition. However, in order to determine the complete set of genes that are expressed under these conditions, and to determine the interaction between these genes, it is necessary to measure a time course of expression experiments. This allows us to determine not only the stable state following a new condition, but also the pathway and networks that were activated in order to arrive at this new state [7].

Microarray time series gene expression experiments are widely used to study a range of biological processes such as the cell cycle [8], development [9], and chemical exposure response. Experimental design is key to the success of any expression experiment. An important computational problem for designing time series expression experiments is the determination of sampling rates. If the experiment is under-sampled, the results might not correctly represent the activity of the genes in the duration of the experiments, and key events will be missed. On the other hand, over-sampling is expensive and time consuming. Since many experiments are limited by budget constraints, over-sampling will result in shorter experiment duration, which might lead to missing important genes that participate in the process at a later stage.

In following a microarray time series experiment, a key challenge is to extract the continuous representation of all genes throughout the course of the experiment. Such a representation enables us to overcome problems related to sampling rate differences and missing values. For instance, one would like to identify genes that have changed significantly after an experimental treatment or that differ between normal and diseased cells. In Bar-Joseph et al. (2004), a method for representing expression profiles by aligned continuous curves is described. Cubic splines are used to represent gene expression curves. Cubic splines are a set of piecewise cubic polynomials and are frequently used for fitting time series and other noisy data. Aach and Church (2001) used linear interpolation to estimate gene expression levels for unobserved time points. D'haeseleer et al. (1999) used spline interpolation on individual genes to interpolate missing time points. Zhao et al. (2001) fitted a statistical model to all genes in order to find those that are cell cycle regulated.

Network Models and Methods

The final analysis level is the networks level in which we focus on the interactions between genes and attempt to build descriptive and predictive models for different systems in the cell. Genomic technology permits large-scale experiments such as microarray experiments, perturbing the activity of many genes and assessing the effect of each perturbation on all other genes in a genome. Inferring how genes within a group of genes can influence the activity of each other is called the genetic network. The activity can be whether a gene is expressed or not, as mRNA or as protein. There is more to gene activity than just expression, for instance, post-translational regulations, phosphorylation, etc. A collection of regulatory proteins associated with genes across a genome can be described as a transcriptional regulatory network. In a genetic perturbation, gene activity is experimentally manipulating either by gene deletion or by inhibition of translation. When manipulating a gene and finding that this manipulation affects the activity of other genes, the question often arises as to whether this is caused by a direct or indirect interaction. A goal of systematic studies of genome regulation is to discover the network structures that control cellular functions at the transcriptional level. To understand the complex transcriptional regulatory networks, it is useful to identify the simplest units of commonly used network architecture. These simple units, or network motifs, provide specific regulatory capacities such as positive and negative feedback loops. The frequency with which cells use individual motifs reveals the regulatory strategies that they selected. These motifs can be assembled into network structures that help explain how a complex gene expression program is regulated. We assume that regulatory network motifs form building blocks that can be combined into larger network structures.

Biological networks are often represented as graphs, with "edges" connecting "nodes". In many applications of reverse engineering, researchers attempt to reconstruct the topology of the networks rather than the nature of the individual relationships (i.e., the type of interactions and its kinetic constants). In these cases a graph, possibly a directed one indicating the direction of influence, constitutes an adequate representation.

Reproducing the dynamic response of a network emphasizes different aspects of a model rather than capturing its steady-state behavior. A classic way to reverse-engineer cellular networks or to test their power as predictive models is to perturb the cellular system and observe its response. A large set of gene expression profiles, for instance, corresponding to distinct biochemical or environmental pertubations, can activate distinct pathways, forcing the cell to find new equilibrium points and thus providing much greater information about its dynamic response [10].

Most cellular processes involve many different molecules. The metabolism of a cell consists of many interlinked reactions. Products of one reaction will be educts of the next, thus forming the metabolic network. Similarly, signaling molecules forms the signaling network. And the same is true for regulatory networks between genes and their products. All these networks are closely related (i.e. the regulatory network is influenced by extra-cellular signals). Our main interest is in gene regulatory network and the role of transcription factors. High-throughput technologies and molecular biological methods allow studying aspects of gene regulatory networks on a large number of genes and proteins in parallel, enabling the study of larger gene networks. Gene networks are concerned with the control of transcription (i.e. how genes are up and down regulated in response to signals). Presence of regulatory sequences in the proximity of genes and the existence of proteins that are able to bind to those elements and to control the activity of genes by either activation or repression of transcription allows the formation of complex regulatory networks, including positive and negative feedbacks. Transcription factors that recognize the regulatory elements in the DNA binding site need to interact with other proteins in order to activate gene expression [11].

For regulatory networks, the components of such models are the genes (or their protein products) that are involved in a specific system, and the TFs that regulate the system. Such models provide a description of the process under investigation, and the interactions that take place during the activation of the system. Predictive models should also be able to address questions about different perturbations of the system. Models are useful for many applications. For example, in drug discovery, researchers are interested in identifying proteins that are at the root of a certain disease. Using these models, we can determine which genes are the causes and target them to prevent the spread of the disease. Another important application is to identify side effects of a certain treatment. Targeting a protein can cause a number of side effects that might be toxic to the cell. Using genetic interaction models, we can determine the most probable side effects in advance and target only those proteins for which these side effects are minimal.

Sampling rates and temporal aggregations can have a negative impact on our ability to correctly reconstruct temporal networks (Bay et al., 2003). Thus, solving problems at lower analysis levels is an important step toward reconstructing temporal interaction networks. We need to select an appropriate computational modeling framework for such systems. A generative model for various systems will be the ultimate goal; however, due to the large number of genes involved, the current amount of data cannot support such models on a large scale. One possible intermediate solution is to construct networks from gene modules—sets of genes that are assumed to share a common function or be involved in the same pathway. Developing algorithms to identify such modules and assembling them to temporal networks are an important first step toward modeling such systems. In addition, knowledge of the flow of information through cell in response to stimuli can be used to predict the effects of novel stimuli and to modulate the cell's response by altering the activities of specific members of a network. Understanding biology to this degree will require the complete determination of the interactions among genes, proteins and metabolites at many levels of regulation. The transcriptional portion of a cell's regulatory network is currently the most tractable given the availability of high-throughput gene expression data and the progress in sequence pattern analysis in the bioinformatics community [13].

Bayesian Networks

Bayesian networks are a class of graphical models that have been widely employed in the reverse engineering of cellular networks [14,15]. This approach represents a joint probability distribution as a directed acyclic graph whose vertices corresponds to random variables, and whose edges correspond to parent-child dependencies among variables. Given a set of microarray data, D, the inference task is to find a network that best matches these data. In general, Bayesian networks introduce a statistically motivated scoring function to evaluate the posterior probability of a graph given the data, P(G|D), and search for the graph that produces the highest score. The logarithm of the posterior probability is often used to simplify calculations, and by Bayes' rule:

$$\log P(G|D) = \log P(D|G) + \log P(G) - \log P(D).$$

where P(D) is independent of G, P(G) is the prior distribution of G, P(D|G) is the probability of the data given the network, G.

The Bayesian-Dirichlet equivalence (BDe) is a scoring criterion to capture the posterior probability. BDe is a Bayesian approach for penalizing complex models (i.e., models with many free parameters). Alternatively, the maximum likelihood parameters for the possible parameterization, θ, may be used to estimate P(D|G), while imposing an explicit penalty term that is a function of the complexity of the model. A common choice for this penalty is the Bayesian Information Criterion (BIC) [16]. The maximum likelihood estimate with the BIC penalty converges asymptotically to the BDe.

Once a Bayesian scoring metric has been defined, learning the most likely structure of a Bayesian Network reduces to searching the entire graph space for the highest-scoring model. This problem is known to be NP-hard [17] and can be written as:

$$\mathrm{S}\ (\mathrm{G{:}D}) = \sum_{i} \mathrm{ScoreContribution}(\mathrm{X_i,p_i,{:}D}),$$

where S(G:D) = logP(G|D). If uniform priors are used for P(G) and P(D) then the log likelihood logP(D|G) may be used as the scoring function [18].

Information Theoretic Methods

While graphical models provide a rich and flexible toolbox for probabilistic inference, they still rely on specification of a local probability distribution and the conditional independence. One information theoretic quantity, mutual information (MI), can capture arbitrary, nonlinear relationships between variables. MI computes the differential entropy between gene expression profiles (GEP), and for a pair of random variables, $\vec{X}_i$ and $\vec{X}_j$, is defined as:

$$I_{i,j} = \mathrm{S}(\vec{X}_i) + \mathrm{S}(\vec{X}_j) - \mathrm{S}(\vec{X}_i, \vec{X}_j)$$

where S(t) is the entropy of an arbitrary variable t. Like the Pearson correlation, MI measures the degree of statistical dependency between two variables. Several groups have developed network reconstruction algorithms based on MI. The first steps were taken by Butte and Kohane [19], using an approach that simply inferred edges to exist between gene pairs with MI above a certain statistical significance threshold, as calculated by permutation test. However, this approach will incur a large number of false positives, as many indirectly interacting genes will have significant MI scores, (e.g. those separated by one or more intermediaries in a transcriptional cascade). An extended approach was developed in the ARACNE [20] algorithm by applying the data-processing inequality (DPI), which states that if genes g1 and g3 interact only through a third gene g2, then

$$I(\text{g1,g3}) \leq \min[I\,(\text{g1,g2});\; I\,(\text{g2,g3})].$$

ARACNE starts with a network graph where each $I_{i,j} > I_o$ is represented by an edge (ij). The algorithm then examines each gene triplet for which all three MIs are greater than I_o and removes the edge with the smallest value. Provided that pair-wise interactions are the dominant interactions in the network and that MI can be estimated with no errors, ARACNE will model tree networks with zero error, as well as those that are locally tree like – that is, the shortest network paths dominate inter-node information transfer [20].

Partial Correlations

Another method proposed for the reconstruction of genetic network was aimed to identify correlations between variables that are not due to more distant network interactions (i.e., *A* correlates with *B* because one interacts with the other, rather than the two are correlated because *C* affects both) [21]. As with other correlation-based approaches, the number of data points presented should be large in order to allow for good statistical inference. The Pearson product moment correlation coefficient is a widely used measure of association between continuous random variables. A partial correlation coefficient quantifies the correlation between two variables when conditioning on one or several other variables. The order of the partial correlation coefficient is determined by the number of variables it is conditioned on. It can be calculated to any arbitrary order. For a calculation of 0-2 order:

zeroth-order correlation $r_{xy} = \frac{\mathrm{cov}(xy)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$

first-order correlation $r_{xy.z} = \frac{r_{xy} - r_{xy} r_{yz}}{\sqrt{(1 - r_{xz}^{2})(1 - r_{yz}^{2})}}$

second-order correlation $r_{xy.zq} = \frac{r_{xy.z} - r_{xq.z} r_{yq.z}}{\sqrt{(1 - r_{xq.z}^{2})(1 - r_{yq.z}^{2})}}$

Although partial correlation analysis still does not infer causal relationships, it excludes many of the possibilities, and thus is a step in the direction of causal inference [21].

Graphical Gaussian Model (GGM), also known as "covariance selection" or " concentration graph" model, has recently become a popular tool to study gene association networks. It is a multivariate analysis to infer or test a statistical model for the relationship among a plural of variables [22], where a partial correlation coefficient, instead of a correlation coefficient, is used as a measure to select the first type of interaction. The key idea behind GGMs is to use partial correlations as a measure of independence of any two genes. This makes it straightforward to distinguish direct from indirect interactions. There is a simple reason why GGMs should be preferred over relevance networks for identification of gene networks: the correlation coefficient is weak criterion for measuring dependence, as marginally (i.e. directly and indirectly), more or less all genes will be correlated. This implies that zero correlation is in fact a strong indicator for independence (i.e. the case of no edge in a network), but this is of course not what one usually wants to find out by building a relevance network. On the other hand, partial correlation coefficients do provide a strong measure of dependence and, correspondingly, offer only a weak criterion of independence [23].

Recently, a number of papers discussed the use of dynamic Bayesian networks (DBNs) for modeling time series expression data. DBNs are an extension of Bayesian networks (BNs), which have been successfully applied to model static expression data (Pe'er et al., 2001). The main advantage of DBNs for gene expression data is that unlike BNs, which are acyclic, DBNs allow for cycles, which are common in many biological systems. In addition, DBNs can also improve our ability to learn causal relationships by relying on the temporal nature of the data. Kim et al. (2003) used DBNs to model a 45 genes subnetwork of the cell cycle system in yeast. By comparing the resulting network with a previously determined network from the KEGG database, they have concluded that many of the edges can be correctly identified using DBNs. Perrin et al. (2003) presented a DBN model containing hidden variables (i.e. nodes for which we do not have direct observation) to overcome both biological and measurement noise. Their model uses an extension of the linear regression model with normally distributed noise. They applied their method to model the DNA repair network in *E. coli*, focusing on the eight main genes in that system. In general, they have found that their method was capable of extracting the main regulatory circuits for this system. As for prediction, they observed a very high correlation between the prediction of the generated network for the next time step and the actual values observed (0.97) and a somewhat lower correlation for similar prediction of multiple steps (0.65). In order to test the application of DBNs to gene expression data and to determine their accuracy, Husmeier (2003) performed a simulation-based analysis. Unlike with real biological data, with a simulation-based study we know what the correct network is, so it is possible to compare the resulting network and the true (underlying) network. This was done by selecting a significance threshold for each edge, and determining the true positive (how many correct edges were recovered) and false positive (how many recovered edges do not exist in the true network) rates. Husmeier et al. concludes that while the global network recovered by DBNs is not useful, local structures can be recovered to a certain extent.

Integrating Gene Ontology

The Gene Ontology (GO) is a structured vocabulary for describing biological processes, molecular functions, and cellular components of gene products [9]. GO classification helps gain biological insights from a set of identified genes of interest to determine which GO terms annotations are overrepresented among the genes in the set.

CHAPTER II

FATHEAD MINNOW OVARY EXPERIMENT

Background

The rapidly growing list of chemicals that have the potential of being released into the environment has generated much concern over environmental degradation. In response, both researchers and regulatory agencies are developing approaches to address this concern. The vast number of chemicals to be examined requires tiered screening strategies that incorporate bioassays and computational models to predict whether effects such as endocrine disruption are likely. Traditional methods such as bioassays still remain important tools to assess toxic effects. But lately, more efforts are being made to use sophisticated computational models to try to understand how chemicals can affect the hypothalamic-pituitary-gonadal (HPG) axes, and how changes in these axes can cause endocrine disruption by examining many parameters, such as potential binding relations, and creating dynamic models to predict biological changes (Breen et al. 2007; Watanabe et al., 2007).

The endocrine system regulates reproductive function through signalling molecules like estradiol and testosterone. A number of chemicals present in the environment have the potential to disrupt the endocrine system of the exposed organisms, and in turn, alter physiological functions. Disruption or interference of endocrine system, like dysregulated hormone release and inappropriate response to signaling, can lead to many abnormalities. Fadrozole is one of such chemical that has the potential to inhibit aromatase activities. Aromatase is a key enzyme that catalyzes the estrogen synthesis and converts androgens to estrogens. This aromatization is an important factor in sexual development. Inhibition of this enzyme leads to low estrogen levels.

Changes in the expression of genes that play fundamental roles in the development can signal for subsequent tissue-and organism-level effects. In this study, we analyzed *ex vivo* steroid production, plasma steroid levels, plasma vitellogenin concentrations in the ovary, and reverse engineered the transcriptional network from gene expression using 15,000 probe microarrays. Fathead minnow (*Pimeohales promelas*), a model species for endocrine disruption research (28), was used for this study.

FAD is a chemical that inhibits aromatase (CYP19A), a key enzyme that catalyzes the rate limiting conversion of testosterone (T) to 17b-estradiol (E2) (Miller 1998). Critical processes such as reproduction, metabolism, and development are maintained in the face of a multitude of chemical, physical, and biological stressors. The endocrine system is one such process, where an intricate network of organs, hormones, receptors, proteins and genes control reproduction, development, growth and metabolism. The system is highly conserved within vertebrates (Ankley and Johnson 2004). Nevertheless, little is known about the details of this system in terms of network structure and function. A detailed understanding of how biological pathways and networks function will be essential to developing predictive, mechanistic models that are useful in determining the impact of chemicals in the environment and wildlife.

At the networks analysis level, we focus on the interactions between genes and attempt to build descriptive and predictive models for different systems in the cell. For regulatory networks, the components of such models are the genes (or their protein products) that are involved in a specific system and the TFs that regulate the system. Such models provide a description of the process under investigation and the interactions that take place during the activation of the system. Predictive models should also be able to address questions about different perturbations of the system.

Recently, a number of papers discuss the use of dynamic Bayesian networks (DBNs) and Graphical gaussian Model (GGM) for modeling time series expression data. DBNs are an extension of Bayesian networks (BNs), which have been successfully applied to model static expression data. The main advantage of DBNs for gene expression data is that they allow for cycles, which are common in many biological systems. In addition, DBNs can also improve our ability to learn causal relationships by relying on the temporal nature of the data.

GGM has recently become a popular tool to study gene association networks. The key idea behind GGMs is to use partial correlations as a measure of independence of any two genes. This makes it straightforward to distinguish direct from indirect interactions. Also, in GGMs, missing edges indicate conditional independence.

Inferring regulator networks and pathways can be done by investigating the over-representation of the GO terms in the genes, but since fathead minnow has incomplete annotation information, this approach cannot be relied upon. Since clustering only indicates whether the genes are co-regulated with no fine resolution of interactions between them, reverse engineering methods were applied to reconstruct the interaction network. The advantage is that the dependencies among co-regulated genes are often much stronger and robust since genes encoding proteins that participate in the same pathway or are part of the same protein complex are often co-regulated. However, co-regulation does not necessarily imply that genes are functionally related (31). By exploiting the co-regulation dependency information, we may discover more regulatory patterns. Recent studies show that Graphical Gaussian Model (GGM) and Bayesian Network (BN) are two useful tools to reconstruct transcriptional networks (32). Generally, there are a small number of samples (n) than the number of genes (p) in microarray experiments. However, classic GGM theory cannot accommodate the data settings for p>>n. Recently, GGM has been developed to infer gene networks with a limited-order partial correlation function (33, 34).

In this work, we sought to understand the processes involved in response and adaptation of the model species FHM to chemical inhibition of steroidogenesis using a reverse engineering approach. With that purpose in mind, we reverse engineered a transcriptional network from gene expression changes in the ovaries of FHM exposed to fadrozole (FAD) over a period of eight days. We applied clustering methods to the gene expression data with the idea that co-expression is indicative of co-regulation, thus it may identify genes that have similar functions or are involved in related biological processes. We used this regularized GGM to reconstruct the network for 1254 differentially expressed genes for all time points and also for the individual time point genes.

Material and Methods

All chemical exposures and microarray experiment was conducted in the Environmental Protection Agency (EPA) labs and the Environmental Laboratories (EL) at the US Army Corps of Engineers, Vicksburg, MS, USA.

Fish Exposures

Fish exposures and sampling have been previously described in Villeneuve et al. (2009). Briefly, FAD was provided by Novartis, Inc. (Summit, NJ, USA). All fish used in the study were reproductively mature adult fathead minnows (5-6 months old) obtained from an on-site culture facility at the US EPA Mid-Continent Ecology Division (Duluth, MN). All laboratory procedures involving animals were reviewed and approved by the Animal Care and Use Committee in accordance with Animal Welfare Act and Interagency Research Animal Committee guidelines. Exposures were conducted in 20 L glass aquaria containing 10 L of UV treated, membrane filtered, Lake Superior water containing nominal concentrations of 0 or 30 mg/L FAD. All treatments were delivered as a continuous flow through at a rate of approximately 45 ml/min without the use of carrier solvents. Toxicant (and control water) delivery was initiated to 16 replicate tanks per treatment group approximately 48h prior to test initiation to ensure that stable water concentrations were achieved before adding fish. Exposures were then initiated by transferring random groups of 4 female FHM to each tank. After 24, 48, 96, and 192 h of exposure, fish from two replicate tanks per treatment group were sampled (a total of 8 females per treatment per time point). During each sampling period, the fish were euthanized in a buffered solution of tricaine methanesulfonate (MS-222; Finquel; Argent, Redmond, WA, USA). Blood was collected using heparinised microhematocrit tubes and plasma was separated by centrifugation. Plasma samples were stored at -80°C until extracted and analyzed. Liver, gonads, brain, and pituitary were removed, snap frozen in liquid nitrogen, and stored at -80°C until posterior use for RNA extraction.

RNA

Total RNA was isolated from 30-50 mg FHM ovary tissue with the RNA Stat-60 reagent (Tel-test, Friendswood, TX), as previously described (Garcia-Reyero et al. 2006). Total RNA was treated with DNase and the quality assessed with an Agilent 2100 BioAnalyzer (Agilent, Palo Alto, CA), and the quantity determined on a nanodrop spectrophotometer (Nanodrop Technologies, Wilmington, DE). RNA was stored at 80°C until further use.

Microarrays

Fathead minnow microarrays manufactured by Agilent (Palo Alto, CA) were designed at University of Florida. The arrays contain 15,000 genes in an 8 array per slide format. Array hybridizations were performed using a single color design. Due to sample quality, the total number of replicates per treatment was: 5 for control and 7 for treated day 1; 8 for both control and treated day 2; 8 for control and 7 for treatment day 4; and 7 for both control and treated day 8.

The cDNA synthesis, cRNA labeling and hybridizations were performed following the manufacturer's kits and protocols (One Color Microarray-based Gene Expression Analysis Quick Amp Labeling version 5.7; Agilent, Palo Alto, CA). Briefly, 500 ng of each sample was labeled with Cy_3. Once the labeling was complete, samples were hybridized to the microarray using conditions recommended by the manufacturer. After hybridizing for 17 h, microarrays were washed and then scanned with a laser-based detection system (Axon GenePix, Molecular Devices, Sunnyvale, CA, USA). Data was extracted using Feature Extraction (Agilent, Palo Alto, CA). Text versions of the Agilent raw data have been deposited at the Gene Expression Omnibus website (GEO: http://www.ncbi.nlm.nih.gov/geo/).

Microarray Data Analysis

Samples from all high exposure groups (0 and 30 mg fadrozole/L) were analysed to filter the most significantly expressed genes. Raw microarray data was first log-transformed to reduce skewness of the distribution, followed by quantile normalization was applied using Genespring GX10 (Agilent, Palo Alto, CA, USA). All biological replicates of a condition were averaged to reduce the complexity of the data set. To identify genes that are most variable between the control and the treatment, the one-way Analysis of Variance (ANOVA) test was performed, followed by pair-wise analysis for each time point (day 1, 2, 4 & 8) between matched control and treated samples. In order to get a modest number of genes, a cut-off threshold of 1.5 fold-change and $p<0.05$ was used to generate the lists of the most differentially expressed (DE) genes across all four time points. Lower value of alpha or higher value of fold-change reduced the number of DE genes to very low (Table 1b). It is reasonable to assume that threshold below this is unlikely to be of interest for any gene. Genes filtered with ANOVA test were also validated using the PCA for their variance.

Cluster Analysis

Our approach was motivated by an earlier work by Petti and Church (2005), which suggested that the biological networks are modular. These are groups of genes, proteins and other molecules involved within a common subcellular process. Clustering based on the co-regulation indicates they share functionality. DAVID database (35) and MeV (36) clustering tool was used to cluster the DE genes before modeling the gene regulatory networks.

Network Inference

Total DE genes were grouped into 4 clusters and used for network construction with GGM and BN. In the GGM approach, the correlation network was first transformed into a partial correlation network, essentially an undirected graph displaying only direct associations. The partial correlation is the correlation that remains between two random variables if the effect of the other variables or set of controlling variables are removed. The undirected graph was then converted into a partially directed graph. A partial correlation coefficient (pcor) was estimated for each pair of genes in the cluster using the shrinkage approach (37). All edges in the correlation graph with significance are directed in such a fashion that the direction of the arrow points from the node with the larger standardized partial variance to the node with the smaller standardized partial variance (37). The unequal time series aspect of the data was also taken into account by employing dynamic (partial) correlation estimation. Feature selection is a must for any data mining approach. That is because while building a data mining model, the dataset frequently contains more information than is needed to build the model. Removing unneeded data is important and feature selection helps solve this problem by calculating a score for each attribute and then selecting only the attributes that have the best scores. There are many ways to implement feature selection depending on the type of the data and the algorithm that we choose for analysis. In BN approach, a Bayesian–Dirichlet equivalence (BDe) (38) scoring criteria was used to learn optimal network from the data. The Dirichlet distribution is a multinomial distribution that describes the conditional probability of each variable in the network, and has many properties that are useful for learning. We limited the search space to at most 3 parents for each vertex to reduce the computational time for Bayesian and Dynamic Bayesian Network.

Network Properties

Highly connected regions in the network were extracted using the clustering algorithm “Molecular Complex Detection” (MCODE) (39). Recurring network motifs were searched, and the degree distribution of the network was calculated to determine fit to a power-law distribution. The networks were then imported and visualized using Cytoscape (40). BinGO (41) as a Cytoscape plug-in and the literature search was used to interpret our results for any available biological evidence.

Results and Discussion

Genes Affected by Fadrozole Exposure

The number of differentially expressed genes identified in ovaries of FHM females after exposure to 30 ug/L FAD increased with each time point (Table 1-a). The number of DE genes was the highest in 192 hr with more than 63% of the genes were up regulated (Table 1a). Very few genes were differentially expressed in more than two time intervals. A total of 1254 genes were found to be differentially expressed across all four-time points. Analysis of Gene Ontology (GO) overrepresentation determined that the genes had their role in signal transduction, developmental processes, lipid, fatty acid and steroid metabolism, immunity and defense, protein metabolism and modification, and cell communication. Their role in molecular function was mostly in the kinase activities, oxidoreductase activities, nucleic acid binding and as transcription factors. Many of the cytochrome P450 family members (cyp19a1a, cyp26a1, cyp3a65), estrogen receptors (beta a, beta b, estrogen receptor 1), streroid dehydrogenase (hsd3b7), steroidogenic acute regulatory protein (StAR), vitellogenin 3 (vtg3), ATPases, solute carrier family members were significantly expressed. Only 670 gene symbols could be found in the GO and DAVID databases. 322 genes with known GO terms were classified into 4 functional and co-regulated groups. An overall GO term distribution network for the known genes is described in the Figure 2a and Table 1c. Correlation between the samples were examined using the Principal Component Analysis. The first three component of PCA analysis contains more than 65% of the variance between the samples, and the treatment (Figure 2b) and the variance between the exposure and recovery samples (Figure 2c). Gene expression of some of the known biomarkers of aromatase inhibitors were matched with the qPCR results (Figure 2d).

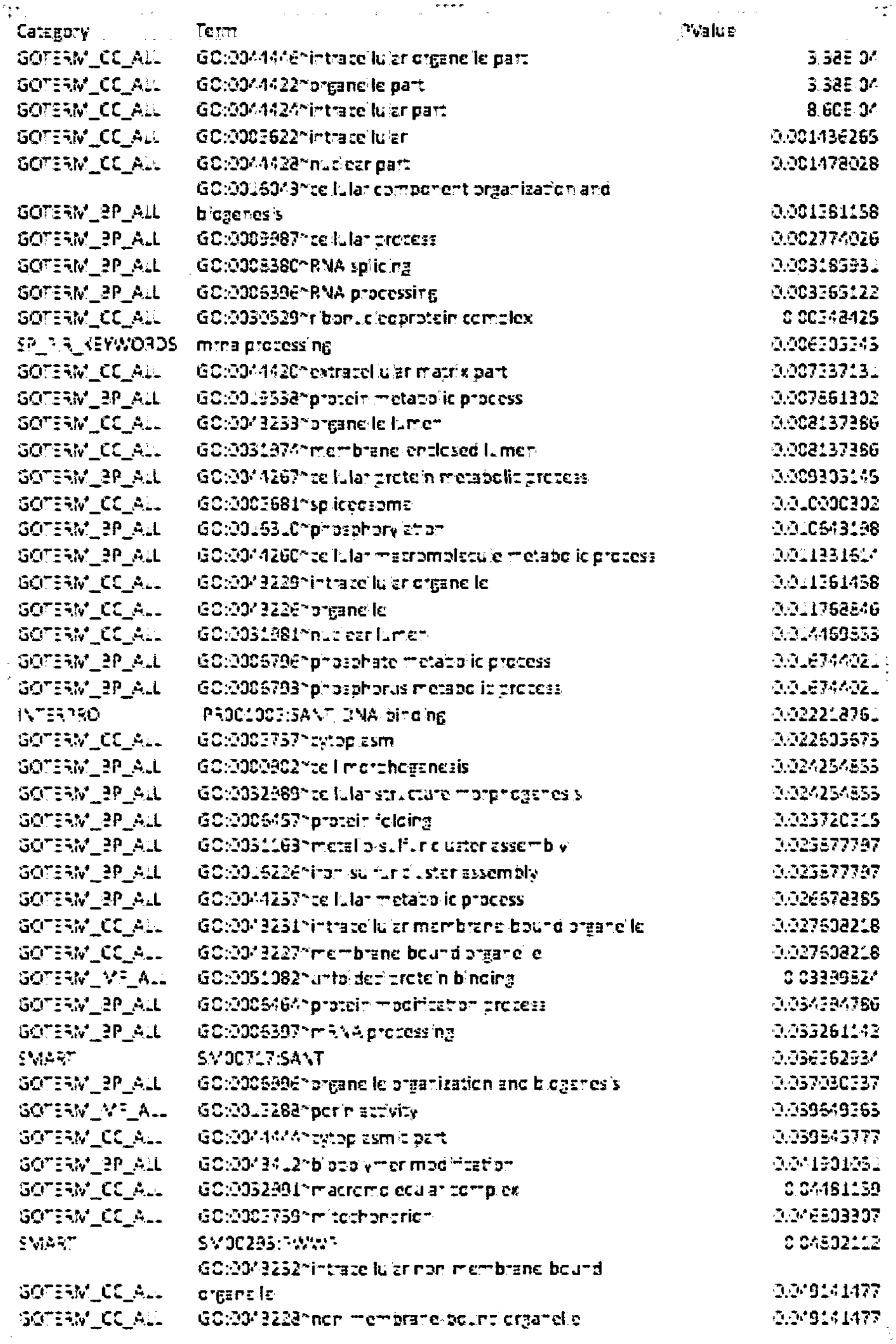
Table 1a: Differentially Expressed Gene set obtained using 1-way ANOVA with a threshold of p-value<0.05 and fold change >1.5.

Time (hr)	DE genes	Up-regulated	Down-regulated
24	209	100	109
48	399	185	214
96	313	135	178
192	427	270	157
Union	1254		

Table 1b: Threshold selection for the optimal p-value and fold-change

	P all	P<0.05	P<0.02	P<0.01	P<0.0050	P<0.0010
FC all	209	209	86	44	27	3
FC > 1.1	209	209	86	44	27	3
FC > 1.5	209	209	86	44	27	3
FC > 2.0	88	88	36	22	14	2
FC > 3.0	23	23	7	5	2	0
Expected by chance		10	4	2	1	0

Table 1c: GO enrichment analysis was performed on all the ANOVA genes. Highly enriched terms found are: -Protein metabolic process, Phosphorylation, Phosphate metabolic process



Category	Term	PValue
GOTERM_CC_ALL	GO:0044446~intracellular organelle part	3.38E-04
GOTERM_CC_ALL	GO:0044422~organelle part	3.38E-04
GOTERM_CC_ALL	GO:0044424~intracellular part	8.60E-04
GOTERM_CC_ALL	GO:0005622~intracellular	0.001436265
GOTERM_CC_ALL	GO:0044428~nuclear part	0.001478028
GOTERM_BP_ALL	GO:0016043~cellular component organization and biogenesis	0.001381158
GOTERM_BP_ALL	GO:0009987~cellular process	0.002774026
GOTERM_BP_ALL	GO:0008380~RNA splicing	0.003185931
GOTERM_BP_ALL	GO:0006396~RNA processing	0.003365122
GOTERM_CC_ALL	GO:0030529~ribonucleoprotein complex	0.00348425
SP_PIR_KEYWORDS	mrna processing	0.006305345
GOTERM_CC_ALL	GO:0044420~extracellular matrix part	0.007337131
GOTERM_BP_ALL	GO:0019538~protein metabolic process	0.007861302
GOTERM_CC_ALL	GO:0043233~organelle lumen	0.008137386
GOTERM_CC_ALL	GO:0031974~membrane-enclosed lumen	0.008137386
GOTERM_BP_ALL	GO:0044267~cellular protein metabolic process	0.009305145
GOTERM_CC_ALL	GO:0005681~spliceosome	0.010000302
GOTERM_BP_ALL	GO:0016310~phosphorylation	0.010643198
GOTERM_BP_ALL	GO:0044260~cellular macromolecule metabolic process	0.011331614
GOTERM_CC_ALL	GO:0043229~intracellular organelle	0.011361458
GOTERM_CC_ALL	GO:0043226~organelle	0.011768846
GOTERM_CC_ALL	GO:0031981~nuclear lumen	0.014469553
GOTERM_BP_ALL	GO:0006796~phosphate metabolic process	0.016744021
GOTERM_BP_ALL	GO:0006793~phosphorus metabolic process	0.016744021
INTERPRO	IPR001005:SANT, DNA-binding	0.022218761
GOTERM_CC_ALL	GO:0005737~cytoplasm	0.022603675
GOTERM_BP_ALL	GO:0000902~cell morphogenesis	0.024254655
GOTERM_BP_ALL	GO:0032989~cellular structure morphogenesis	0.024254655
GOTERM_BP_ALL	GO:0006457~protein folding	0.025720215
GOTERM_BP_ALL	GO:0031163~metallo-sulfur cluster assembly	0.025877797
GOTERM_BP_ALL	GO:0016226~iron-sulfur cluster assembly	0.025877797
GOTERM_BP_ALL	GO:0044237~cellular metabolic process	0.026672865
GOTERM_CC_ALL	GO:0043231~intracellular membrane-bound organelle	0.027608218
GOTERM_CC_ALL	GO:0043227~membrane-bound organelle	0.027608218
GOTERM_MF_ALL	GO:0051082~unfolded protein binding	0.03339824
GOTERM_BP_ALL	GO:0006464~protein modification process	0.034384786
GOTERM_BP_ALL	GO:0006397~mRNA processing	0.035261142
SMART	SM00717:SANT	0.036362934
GOTERM_BP_ALL	GO:0006996~organelle organization and biogenesis	0.037030337
GOTERM_MF_ALL	GO:0015288~porin activity	0.039649365
GOTERM_CC_ALL	GO:0044444~cytoplasmic part	0.039843777
GOTERM_BP_ALL	GO:0043412~biopolymer modification	0.041301051
GOTERM_CC_ALL	GO:0032991~macromolecular complex	0.04481159
GOTERM_CC_ALL	GO:0005739~mitochondrion	0.046803307
SMART	SM00295:[illegible]	0.04802112
GOTERM_CC_ALL	GO:0043232~intracellular non-membrane-bound organelle	0.049141477
GOTERM_CC_ALL	GO:0043228~non-membrane-bound organelle	0.049141477

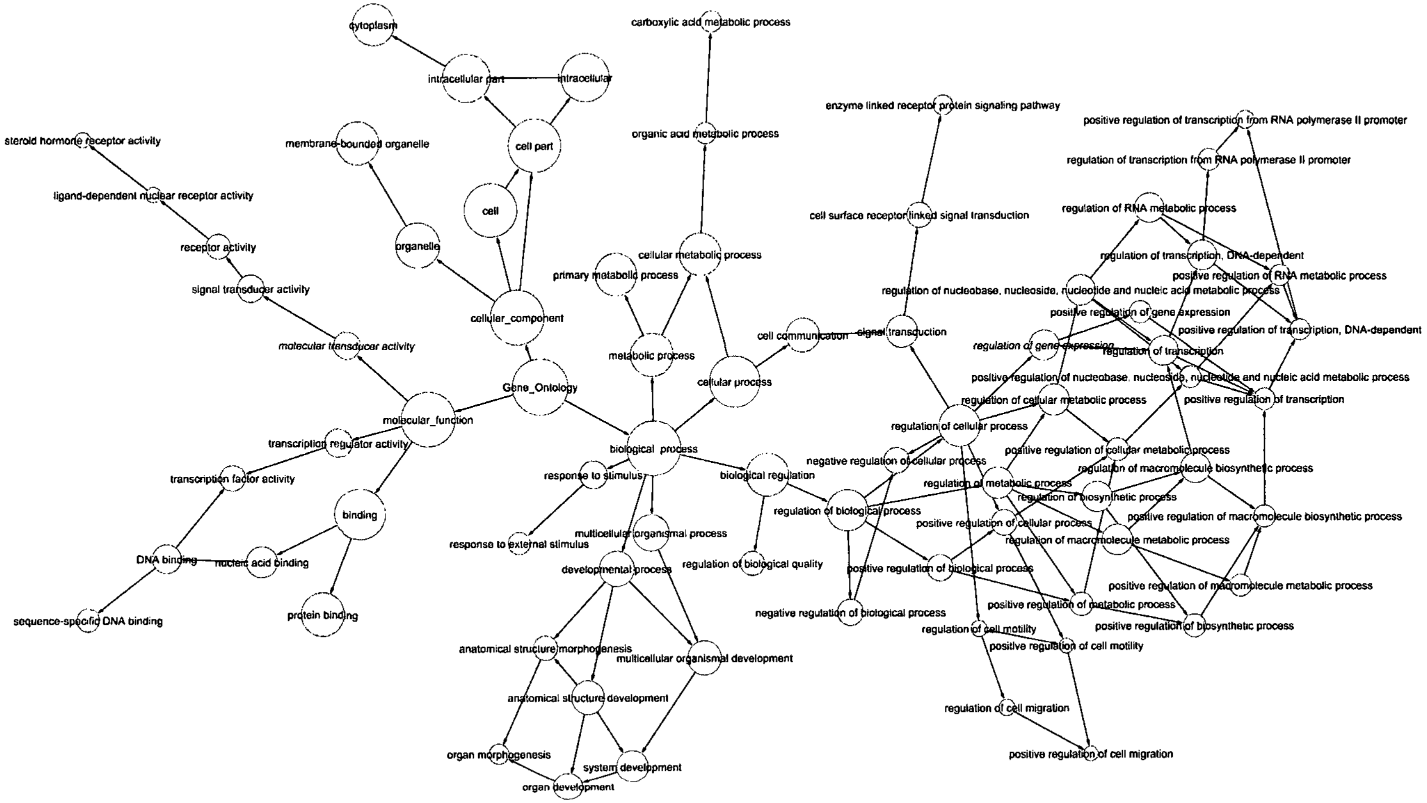

cytoplasm
intracellular part
intracellular
cell part
cell
membrane-bounded organelle
organelle
cellular_component
Gene_Ontology
steroid hormone receptor activity
ligand-dependent nuclear receptor activity
receptor activity
signal transducer activity
molecular transducer activity
molecular_function
transcription regulator activity
transcription factor activity
DNA binding
nucleic acid binding
binding
protein binding
sequence-specific DNA binding
carboxylic acid metabolic process
organic acid metabolic process
cellular metabolic process
primary metabolic process
metabolic process
cellular process
cell communication
biological process
response to stimulus
response to external stimulus
multicellular organismal process
developmental process
anatomical structure morphogenesis
multicellular organismal development
anatomical structure development
organ morphogenesis
system development
organ development
biological regulation
regulation of biological quality
regulation of biological process
negative regulation of biological process
enzyme linked receptor protein signaling pathway
cell surface receptor linked signal transduction
signal transduction
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
regulation of gene expression
positive regulation of gene expression
regulation of transcription
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
regulation of cellular metabolic process
regulation of cellular process
negative regulation of cellular process
regulation of metabolic process
positive regulation of cellular process
positive regulation of biological process
regulation of cell motility
positive regulation of cell motility
regulation of cell migration
positive regulation of cell migration
positive regulation of transcription from RNA polymerase II promoter
regulation of transcription from RNA polymerase II promoter
regulation of RNA metabolic process
regulation of transcription, DNA-dependent
positive regulation of RNA metabolic process
positive regulation of transcription, DNA-dependent
positive regulation of transcription
positive regulation of cellular metabolic process
regulation of macromolecule biosynthetic process
regulation of biosynthetic process
positive regulation of macromolecule biosynthetic process
regulation of macromolecule metabolic process
positive regulation of macromolecule metabolic process
positive regulation of metabolic process
positive regulation of biosynthetic process

Figure 2a - Enrichment of GO term network for 1256 DE genes. Any gene annotated to a certain GO category is also explicitly included all parental categories and then the statistical test was performed to reduce type II error. The most intensely colored nodes that farthest down the hierarchy are the most relevant ones.

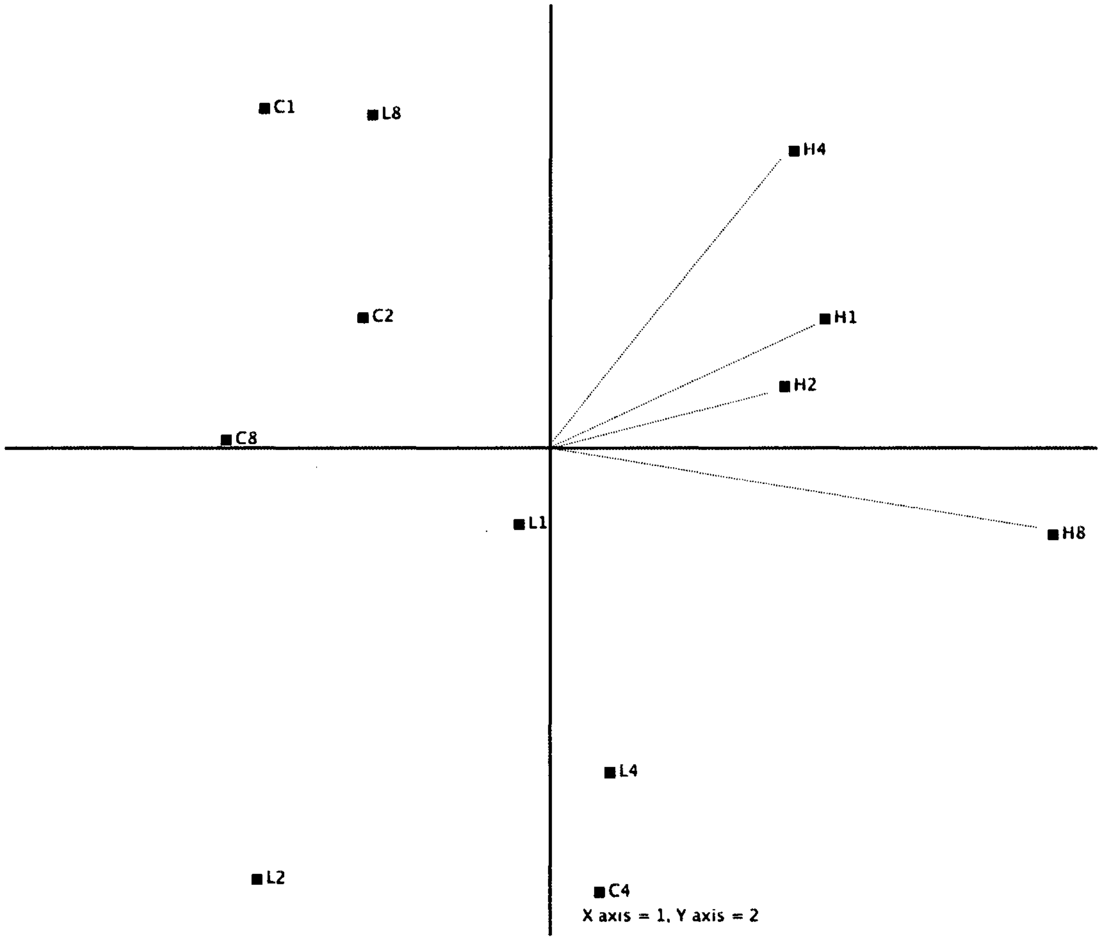


Figure 2b - PCA showing the variance in the samples. C stands for Control, L for Low dose, and H for High dose, 1,2,4, and 8 are the exposure time.

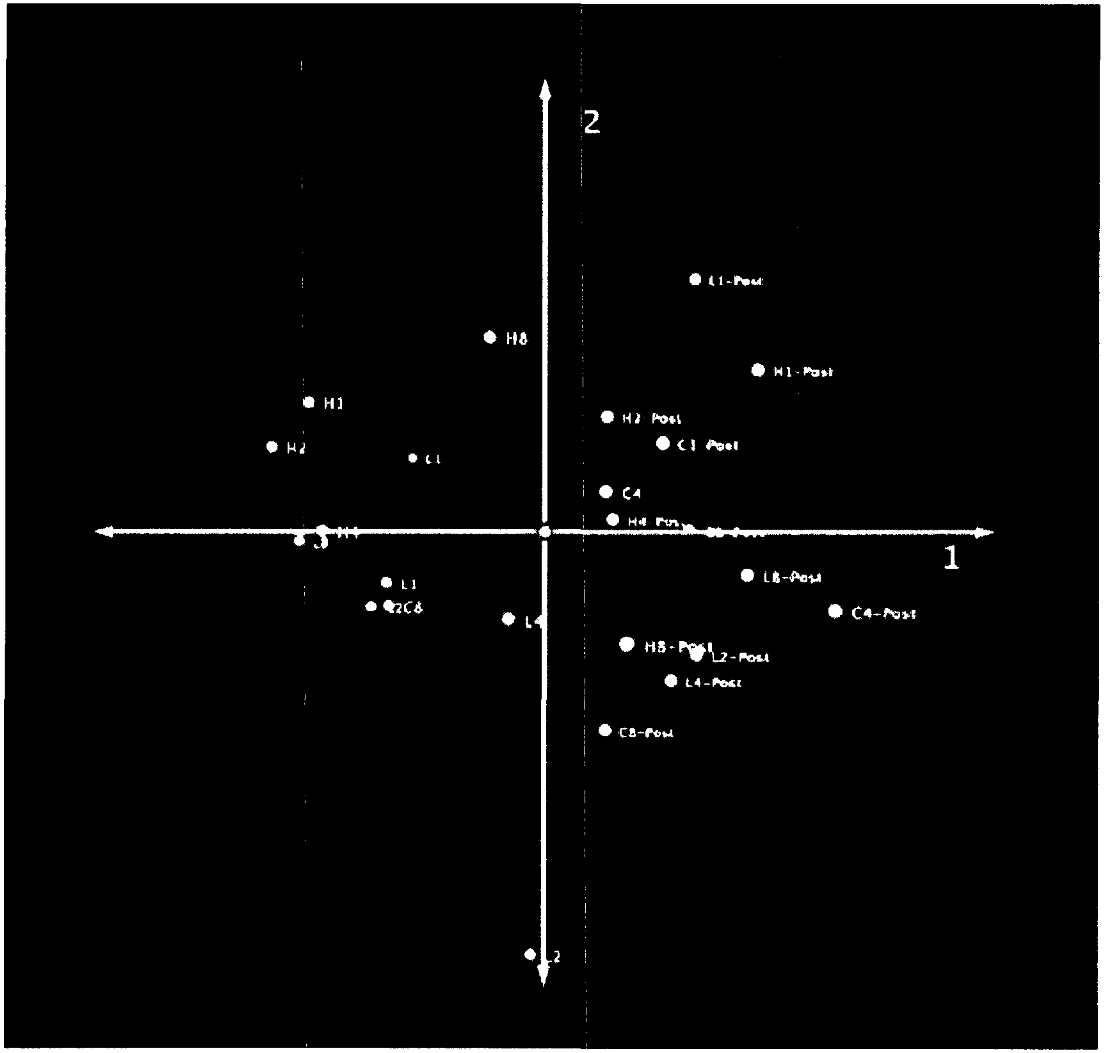


Figure 2c - PCA showing the variance in the exposure and recovery samples. Sample name-post stands for the recovery phase.

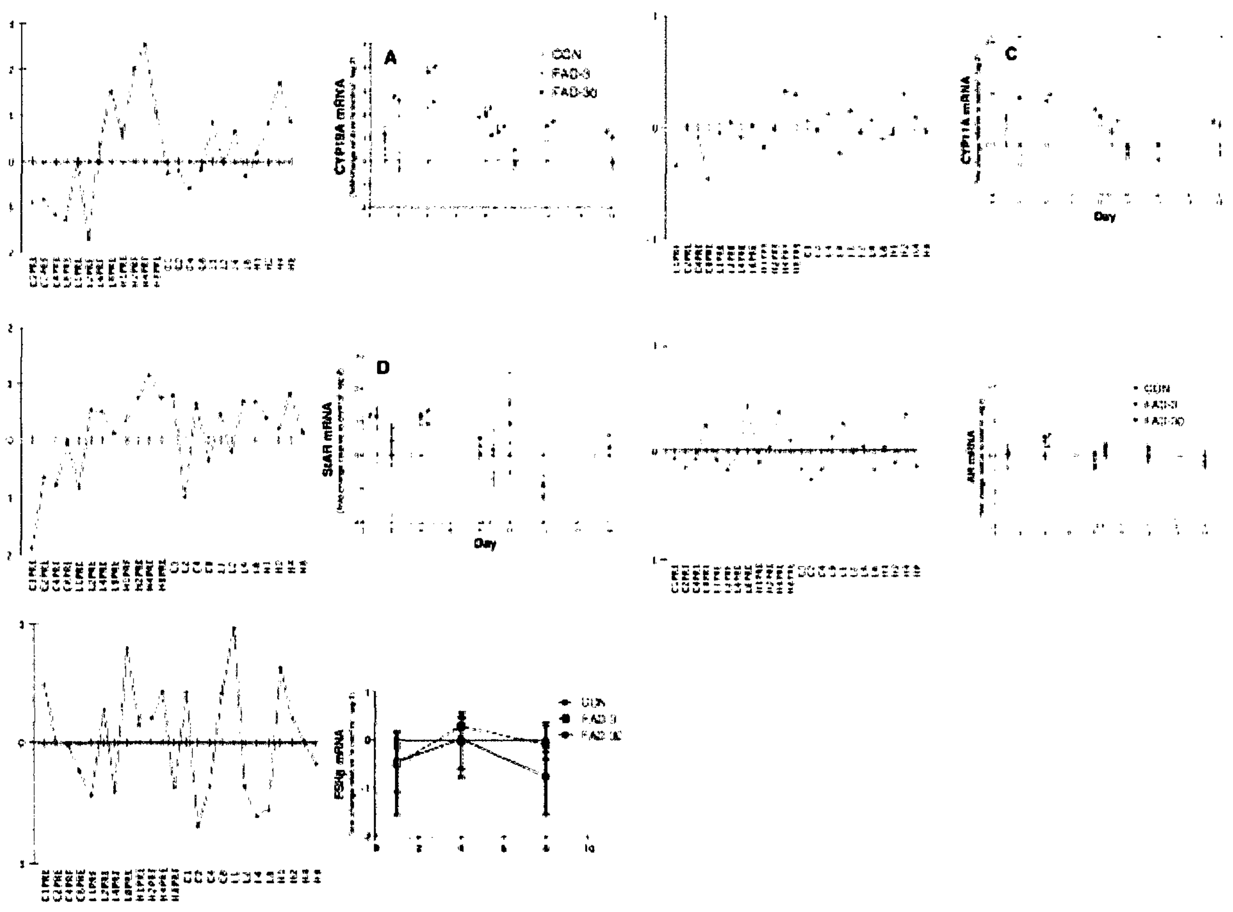


Figure 2d - Relative abundance of mRNA transcripts coding for aromatase (A isoform; CYP19A), follicle stimulating hormone receptor (FSHR), cytochrome P450 cholesterol side chain cleavage (CYP11A), and steroidogenic acute regulatory protein (StAR)

measured in ovary tissue from female fathead minnows exposed to 0 (CON), 3, or 30 μg fadrozole/L (FAD) and sampled on d 1, 2, 4, or 8 of the exposure period, or d 9, 10, 12, or 16 during the recovery period. Data are expressed as fold-change (log 2) relative to the control mean measured on a given day.

Dynamic Network Modeling

The genes identified as DE across all the time points (Table 1a) were used with two different reverse engineering approaches, Graphical Gaussian Model (GGM) and Dynamic Bayesian Network (DBN), to infer the transcriptional networks. Genes were clustered (k=4) based on fuzzy clustering approach, and GGM and DBN were used with each clusters separately for network modeling. The fuzzy clustering procedure allows the genes to participate in more than one cluster. The use of this method in grouping related genes better reflects the nature of biology that a given gene may be associated with more than one functional group of genes.

In order to test the significance of the correlations, the 'local fdr' network search was employed as proposed by Schäfer and Strimmer (2005) for GGM model. A total of 200 significant edges were selected from each cluster to infer the relationship between the nodes. Network obtained from DBN with maximum of 3 parents size network and the network obtained from GGM were very similar, but GGM networks were mostly undirected compared to the dynamic Bayesian network (Figure 5a-d). Genes with more than 10 interactions in each cluster and high correlation values in the network models were found to be transcription factors (e.g. NR3C1, LHX5, COE2, TFAP2C), steroid hormone receptors (NR3C1, ESR2A), and aromatase (CYP19A1A), involved in metabolic processes and trans-membrane movement of ions. The model also showed some other important hubs or interacting genes such as cyp19a1a, estrogen receptor beta a, ATPase Na+/K+ beta 1, ATPase Na+/K+ alpha 1, low-density protein LDLR, signal transducer and activator of transcription 4 STAT4, a number solute carrier family members, OPRL, EDG1, macromolecule metabolic processes (CASP2, ACVRL1, FGFR3, MMP2), transcription factor encoding genes (i.e. CEBPB & CEBPA, HER1, MMP9, LIMK2, and homeobox HOXC8A) (Figure 3 a-d). Interactions involving cyp19a1a and esr2a were highly similar in both GGM and DBN based models. Number of gene interactions such as estrogen receptor with vitamin D receptor, estrogen receptor with ATPase Na+/K_ beta 1a, cytochome P450 19A with estrogen receptor, and glucocorticoid receptor (nr3c1) with ATPase Na+/K+ alpha 1 were confirming other reports (Colin et al. 2003; Martyniuk et al. 2007; Filby et al. 2006; Kolla et al. 2002). Three genes, cyp19a1a, zgc:103585, and zgc:55389, were associated with tryptophan metabolism, and cyp19a1a and cyp26a1 were involved in gamma-Hexachlorocyclohexane degradation. Based on the biological knowledge, some missing edges were subsequently added to the network for sex hormone metabolism. Figure 5 a-b shows the interaction between the steroid hormone receptors (nr3c1, nr2f1a) and cytochrome P450, family 26, subfamily A (cyp26a1), and cyp19a1a interacting with estrogen receptor (esr2a). These interactions were consistent with GGM, DBN methods.

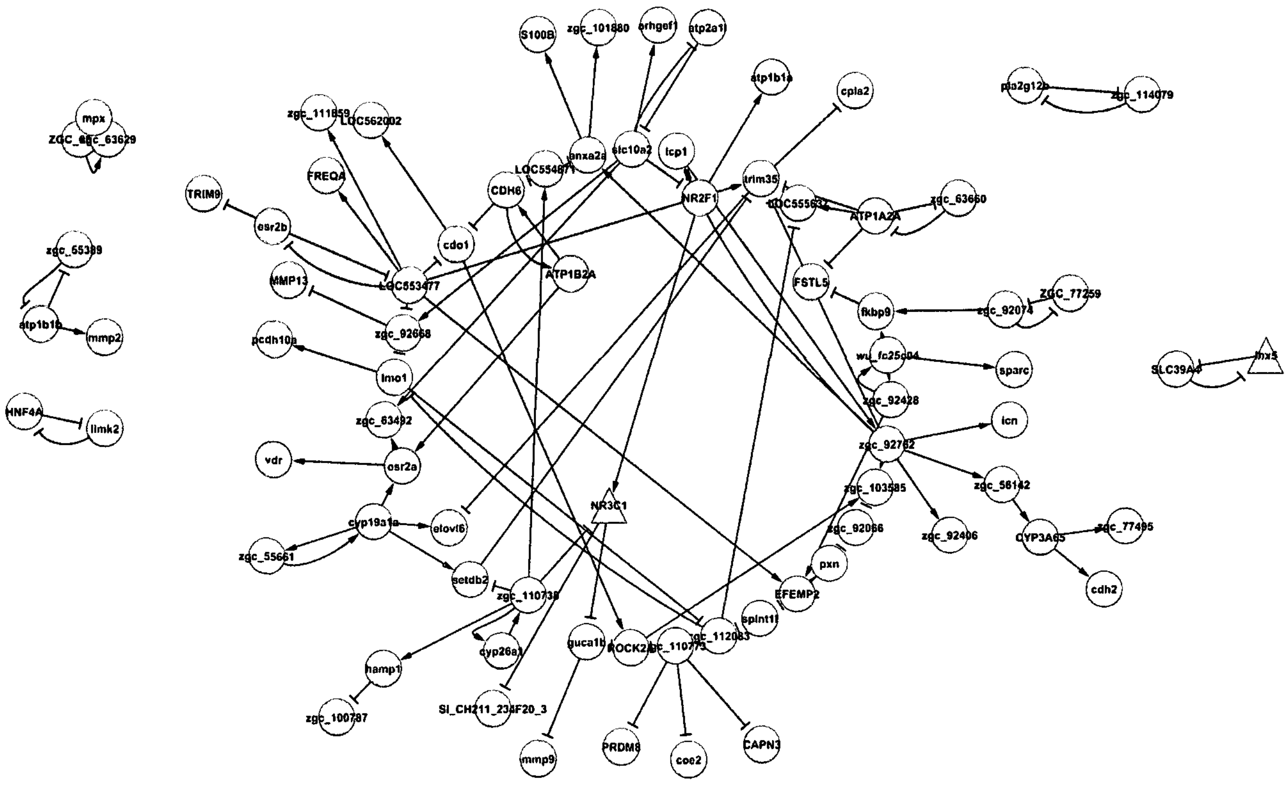


Figure 3: Network models for the cluster approach. (a) Dynamic Bayesian Network for cluster 1: Nodes filled with yellow are the steroid receptors, grey are the ATPases, and green are the Cytochrome P450 members in a circular network layout. Edges with arrow head denotes the positive correlation.

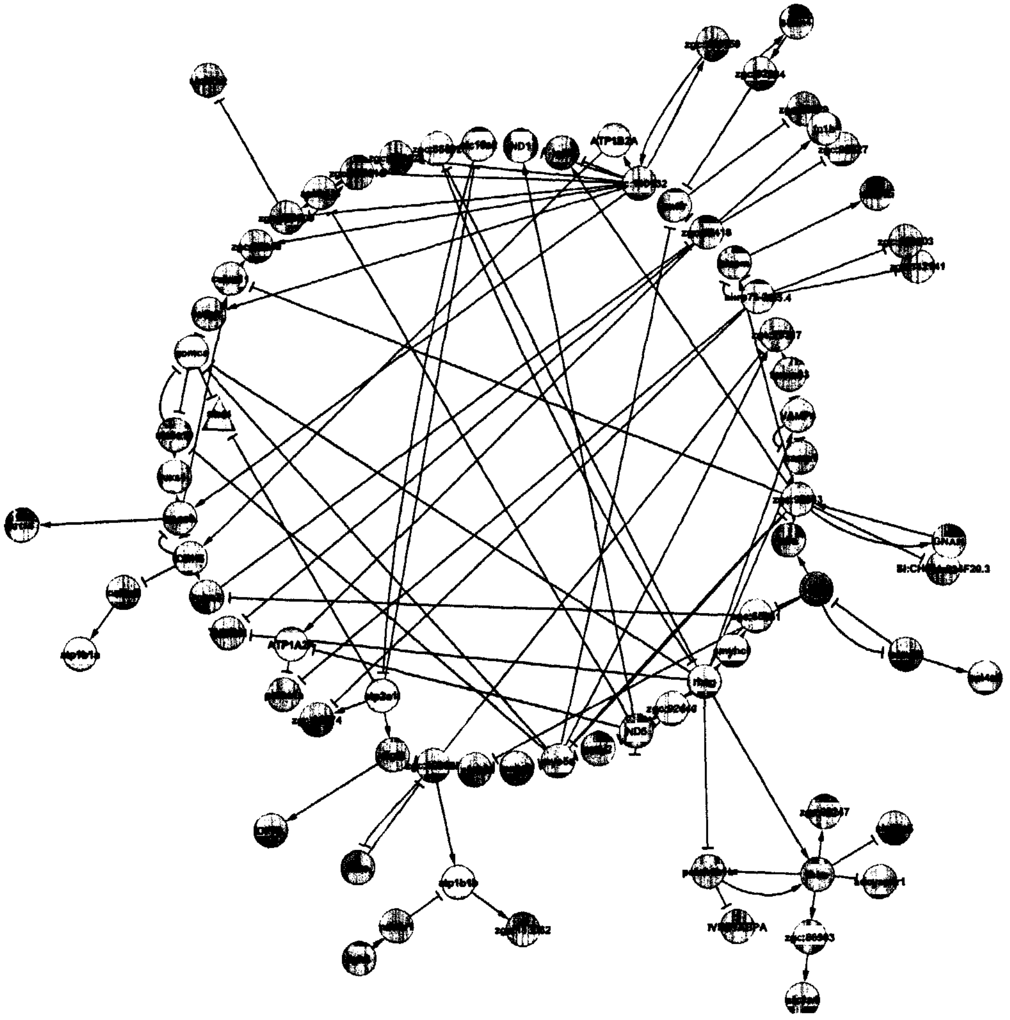


Figure 2b: DBN for cluster 2: Nodes filled green and grey are the LDLR and the ATPases respectively.

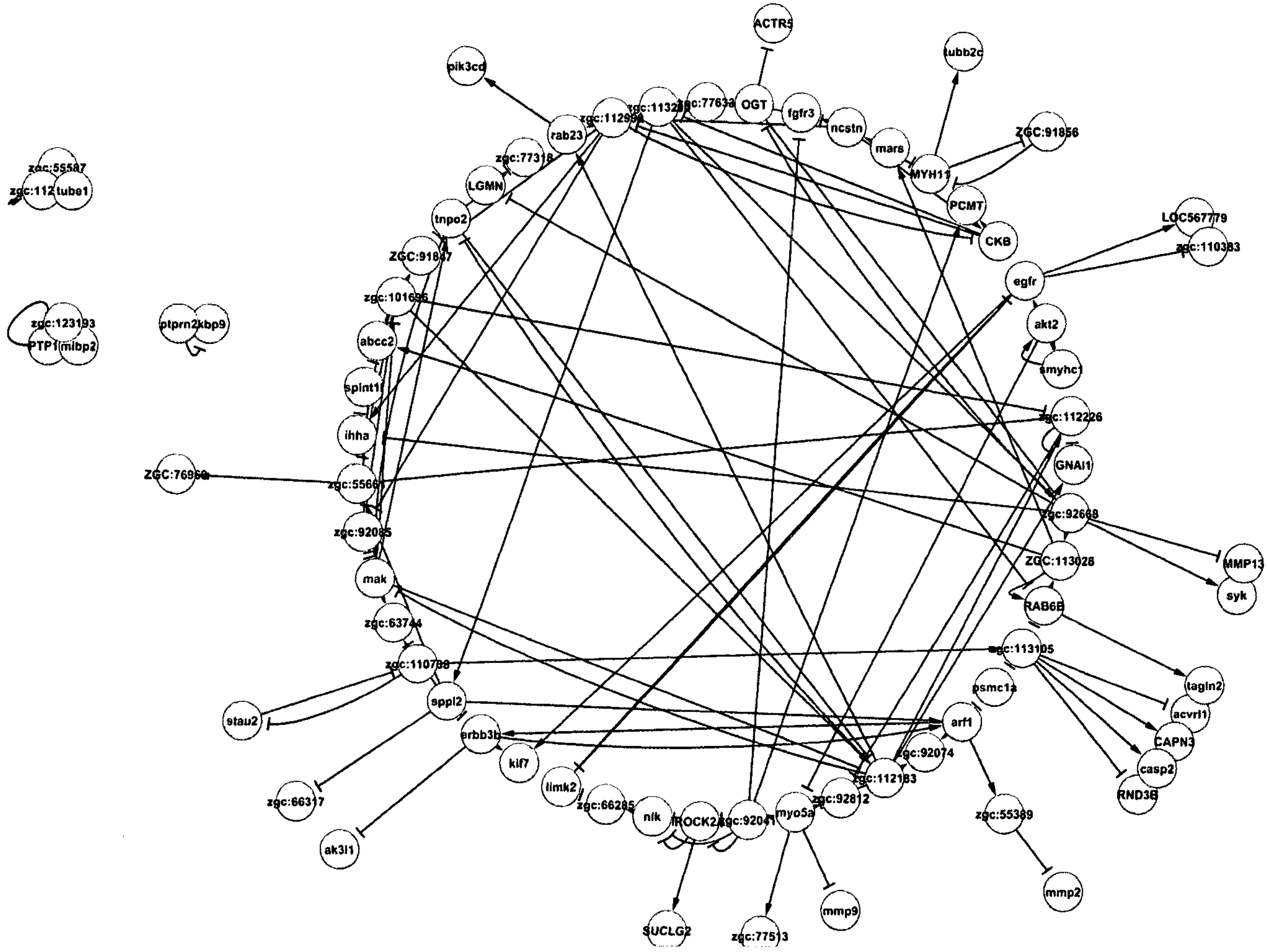


Figure 3c: DBN for cluster 3. Most of the genes in this cluster had no interaction information found.

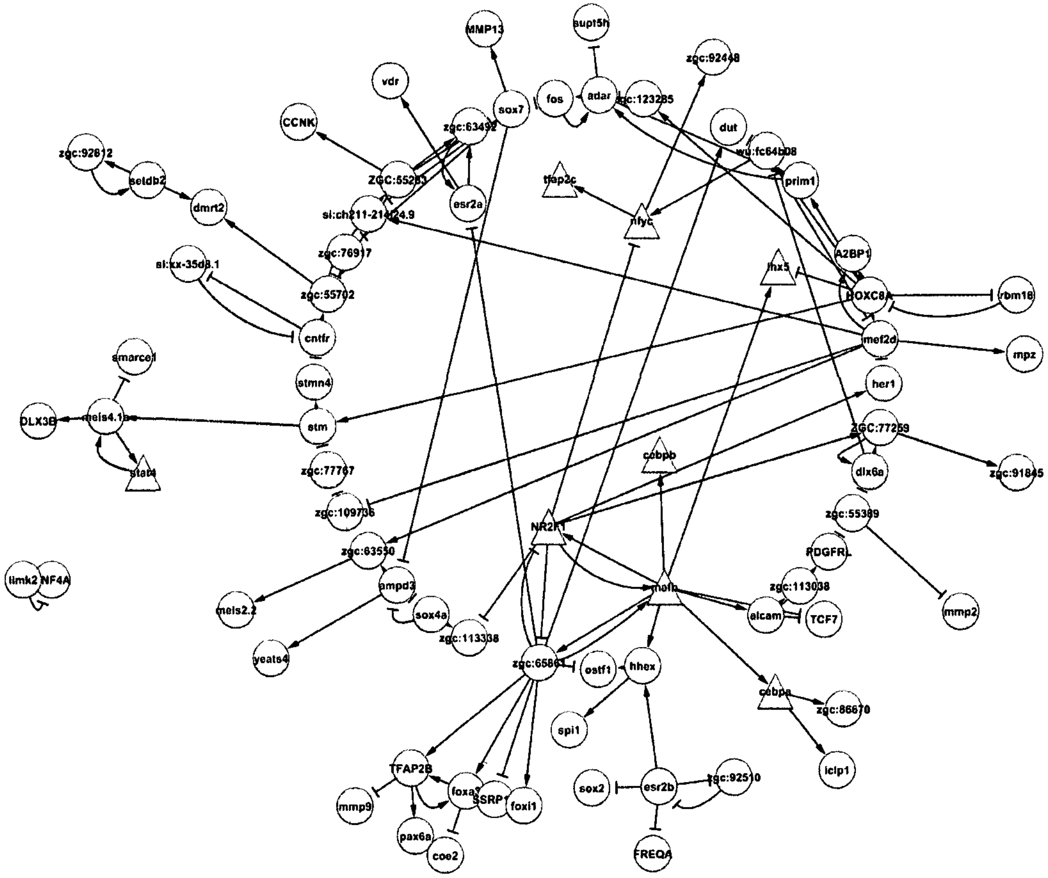


Figure 3d: Network for cluster 4: Nodes in triangle shapes are the Transcription factors, yellow are the steroid receptors. Edges with arrow head denotes the positive correlation.

Steroidogenesis Network

We have constructed a model for the 92 genes on the microarray that has been previously identified as their involvement in gonadal steroidogenesis (34) (Figure 4). With average 2.8 numbers of neighbours per gene, StAR, CYP19A, LDLR showed up as highly connected node. Interactions of StAR with aromatase, estrogen receptor was within one node distance, while interactions with follistatin, vitellogenin were in a very close interaction with StAR. Gene expression heatmap and the hierarchial clustering of 92 genes matched well with the results reported earlier by Villenenue et al. (2007). Literature evidence was found for many other interactions such as activin regulation of 17beta-hydroxysteroid, LDLR interaction with cytochrome P450 (Bak B et al. 2009). We observed that CYP19A1A and StAR were highly expressed during 96hr of exposure. Many variant of follistatins were present in our selective genes of interest, but follistatin 5 (FSTL5) expression was the highest at 24hr exposure, which also matched with the previously reported results.

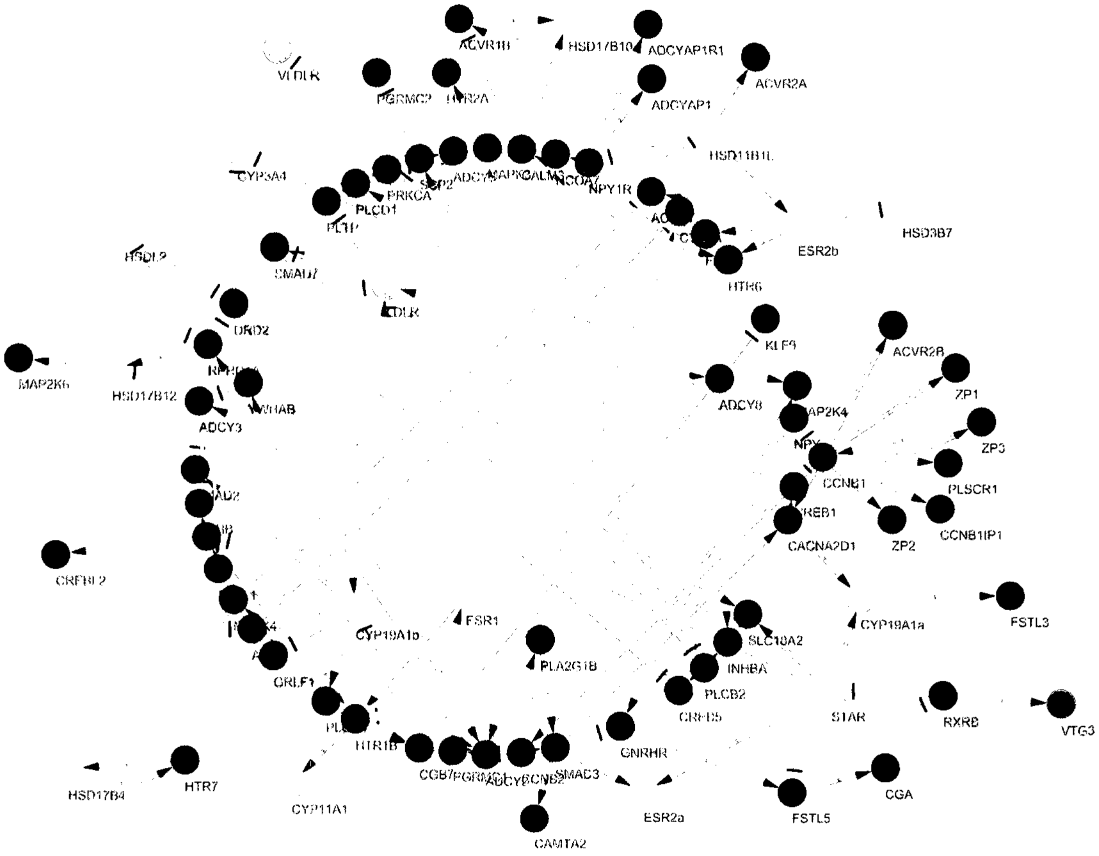


Figure 4: Network model for the genes known to be involved in steroidogenesis such as Cytochrome P450 members (in green), estrogen receptors (in yellow), StAR (in orange), vitellogenin (in red), LDLR (in blue), hydroxy-steroid dehydrogenises (in grey) are shown. Many of the interactions between them was reported earlier.

Bayesian and GGM for Individual Time Exposure

The goal of this approach was to capture the relationships between the individual time point network and compare the network constructed from the cluster approach. By looking at the relationships at the cluster level, we reduce the number of relationships to be estimated in the network, making the networks more tractable when considering large sets of genes from the individual time points. Both GGM and BN were used to construct the models for all the time points.

Inference of a transcriptional network at 24h exposure

Bayesian network and GGM was used to construct the transcriptional network for the 209 significant genes at 24hr. The top 500 significant edges from GGM and the utmost 3 parents size network search from BN were filtered for further investigation. Genes appeared as highly interacting nodes in both the network models were involved in protein kinase activities (e.g. LOC559341, dZ122B7.1), female gonad development (cdh6), regulator of estrogen receptor (smarce1), low density lipoprotein receptor (ldlr), transcription factor (tfap2c), statmin-like 2b gene (stmn2b), and homeo box genes (HOXC8A, DLX3b) (Figure 5). Model based on GGM had few additional highly interacting genes such as cytochrome C1, sex determining region box (sox4a), proteasome subunit, and ATPase 1a (psmc1a). The interaction of LDLR with the junctional adhesion molecule jam2 was confirming other report (Yang et al. 2008). GO enrichment analysis for the genes interacting with the hubs such as statmin-like 2b gene (zgc:110132), homeo box, and low density lipo protein receptor were associated mostly with the metabolic process, ion binding, and intracellular membrane-bound organelle.

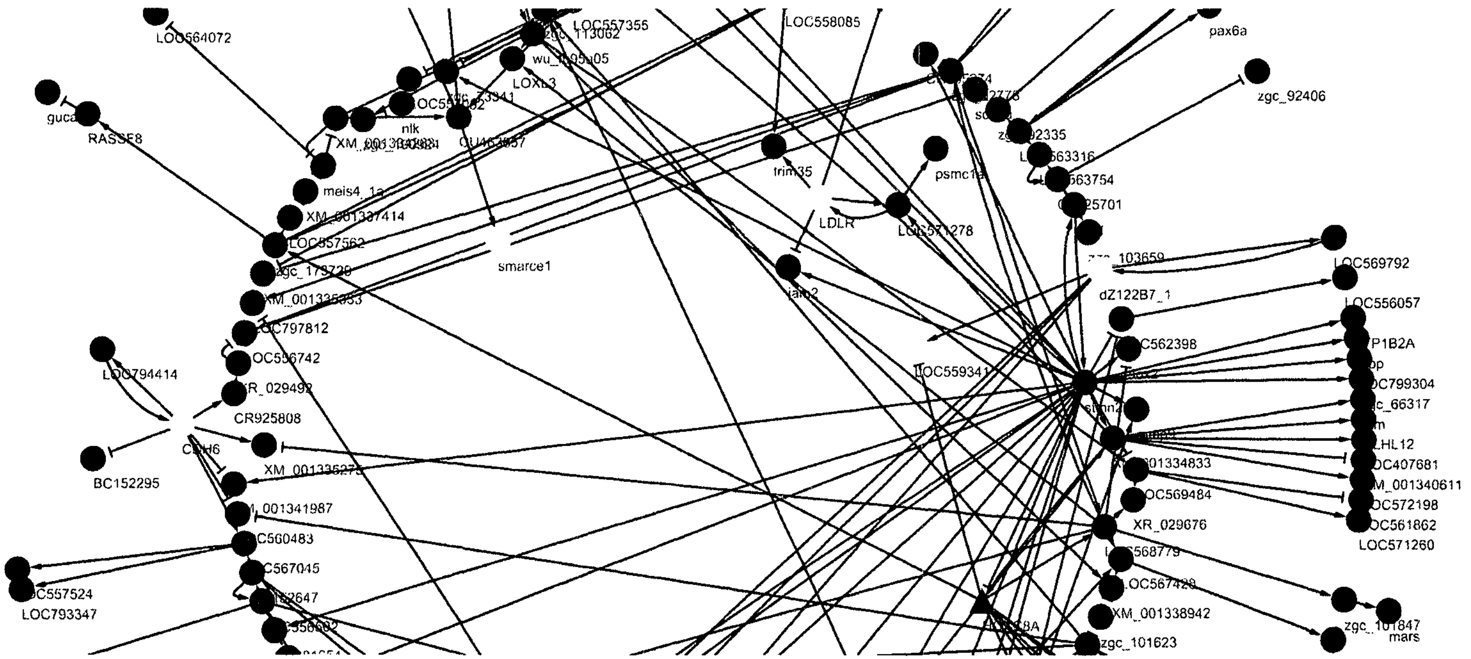


Figure 5: Transcriptional network at 24h exposure. Genes highlighted yellow are LDLR, CDH6, LOC559341, dZ122B7.1, SMARCE1. They are know to be involved in hormonal signalling and steroidogenesis. Hub node stathmin-like 2b was interacting with 29 other genes, mostly involved in cellular metabolic process.

Inference of a transcriptional network at 48h exposure

GGM and BN were used to construct the network for 398 DE genes, of which GGM resulted in 158 genes not interacting with other genes and were filtered out, while BN resulted in only 3 genes not participating in the interaction network. Some potential biomarkers of fadrozole effect such as cytochrome P450 family members (CYP19A1A, CYP26a1, and CYP3A65), estrogen receptor (ESR2b), and hydroxy-delta-5-steroid dehydrogenase (hsd3b7) were few differentially expressed in 48h. Some of these genes from this time exposure were also present in the cluster 1, but there was very less similarity between the two network, with very few interactions matched. The top 10 interacting genes with greater than 10 interactions were found to be involved in protein kinase activities (zgc_110383, fgfr3, LOC564064, dZ122B7.1), male germ cell-associated kinase (MAK), transcription factor interacting proteins (LOC563463, NFYC), chemokine receptor (CXCR4b), CNGA5, GCLM, Cyp3a65, rhag, and CREB1 (Figure 6a). One of the hub genes, Rh-associated glycoprotein (rhag), was interacting with at least 29 neighbours within 1 distance. GO enrichment analysis of this sub-network showed that few of them are associated with macromolecule complex, and majority of them had no annotations available. GO term analysis also showed genes such as Rhag and VAMP4 associated with membrane attack complex, a protein complex produced by sequentially activated components of the complement cascade inserted into a target cell membrane and forming a pore leading to cell lysis via ion and water flow. Another highly interacting gene LOC560805, which is similar to epidermal growth factor like domain (EGF9), was found interacting with estrogen receptor beta 2, epidermal growth factor receptor (egfr), and LIM-homeodomain transcription factor (lhx3) (Figure 8b). Reports confirming the interaction of EGF-receptor activating estrogen receptor, and the role of lim homeobox in gonad formation (Jonathan et al. 2009; Oshima et al. 2007). GO enrichment analysis for another sub-network showed genes involved in oxido-reductase activity, male germ-cell associated kinase, and cytochrome P450 family members (Figure 6b). GO over-representation comparison between 24hr and 48 hr shows an increase in the number of gene involvement in terms such as lipid metabolism, cell surface receptor linked signal transduction, steroid metabolism, spermatogenesis, physiological process, regulation of transcription, and metabolism. Pathway analysis with Fisher Exact P-Value of 0.1 and multiple test correction using Benjamini and Hochberg found four genes, CYP19A1A, CYP26A1, CYP3A65, and HSD3B7 involved with two significant pathways, gamma-Hexachlorocyclohexane degradation (KEGG 00361) and Linoleic acid metabolism (KEGG 00591).

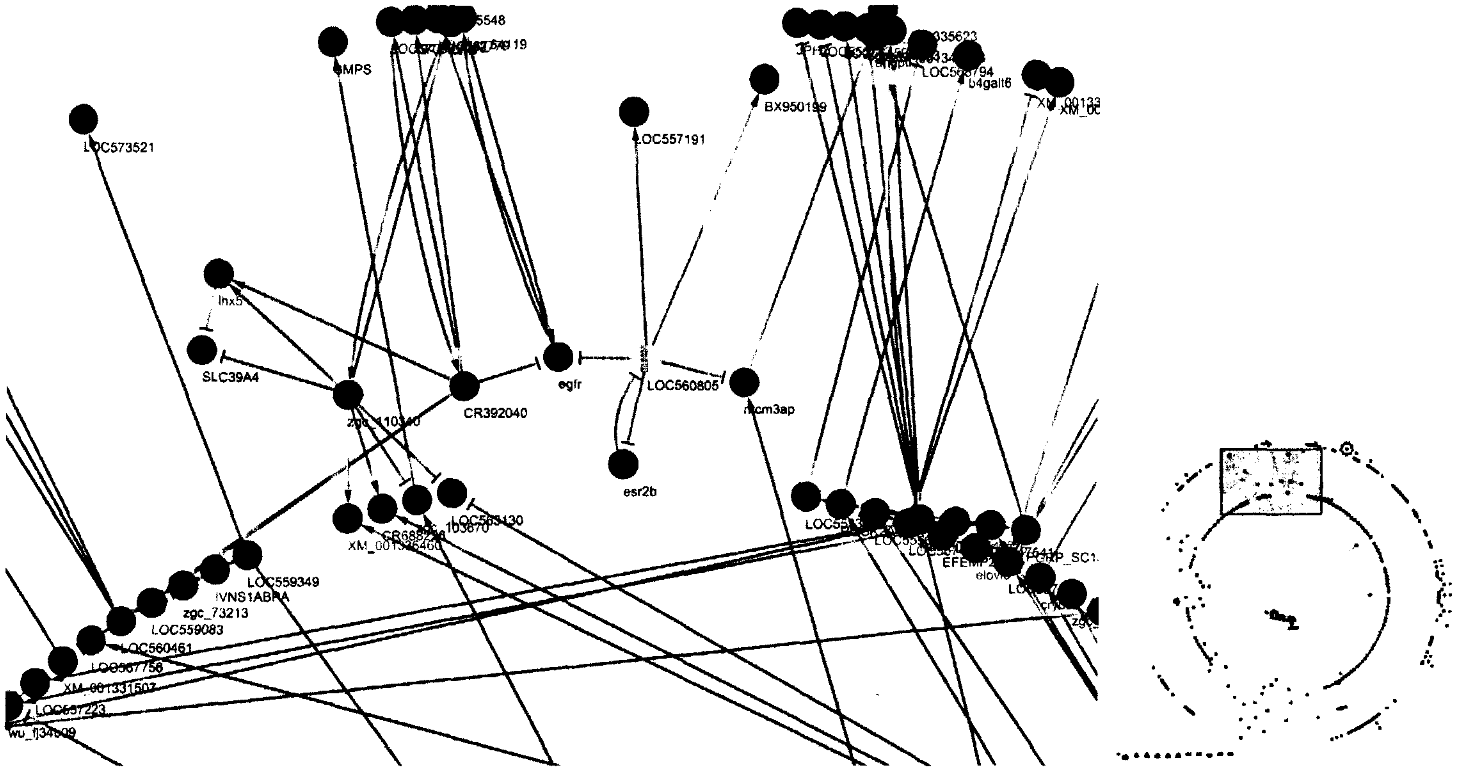


Figure 6(a): Highlighted box showing Interaction of LOC560805 (EGF9) (in yellow) with estrogen receptor (ESR2b) and epidermal growth factor (EGFR). Known genes and their interactions were found in the literatures, confirming our results.

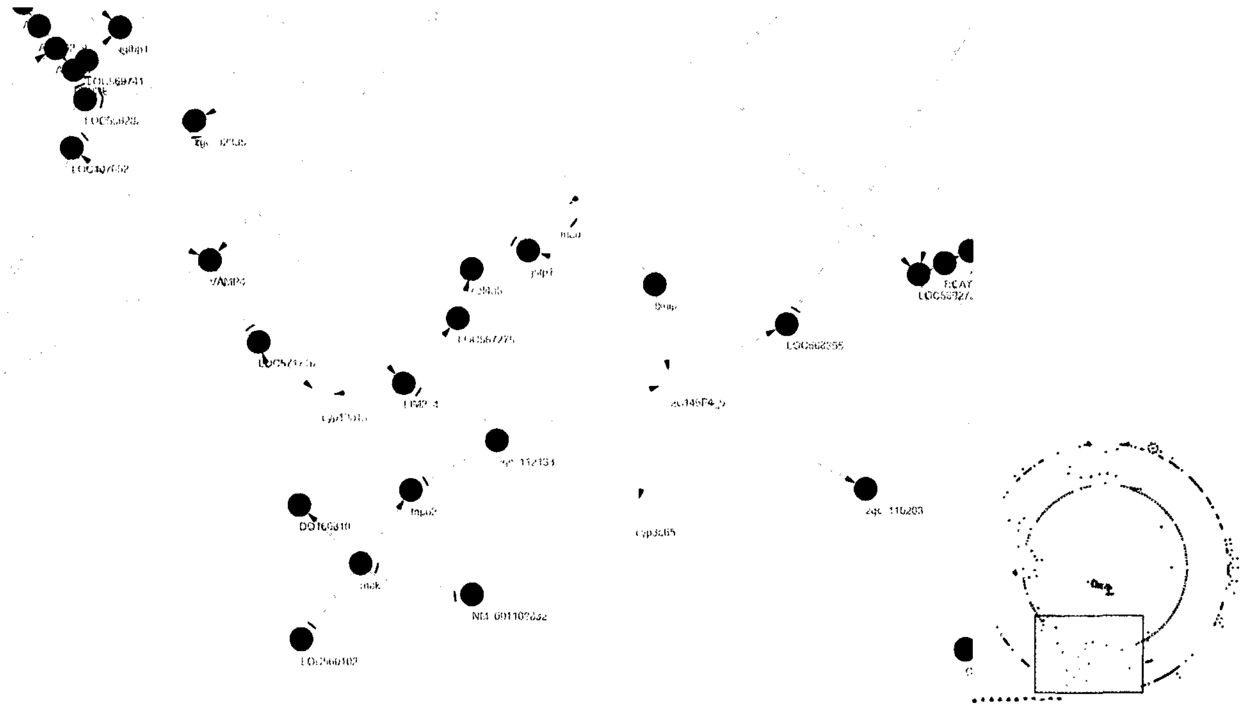


Figure 6(b): Highlighted box showing the sub-network that involves two cytochrome P450 members (yellow), oxidoreductase activity (blue), male-germ-cell associated kinase.

Inference of a transcriptional network at 96h exposure

KEGG pathway analysis with Fisher Exact P-Value of 0.1 and multiple test correction found that two significant pathways were involved during this time point: Tryptophan metabolism (KEGG 00380): CYP19a1a, zgc:103585, zgc:55389; and Oxidative phosphorylation (KEGG 00190): CYTB, IND5, ND1. Cytochome b and NADH dehydrogenase subunit 5 were found interacting in the network model with few other, mostly unknown genes. Also, genes such as StAR, CYP46a1, ESR2a, CREB313, STAT4 were forming a sub-network structure, similar to the steroidogenic pathway modeled by Ananko et al. (2002). Compared to the model from Anako et al., our model had few additional nodes that were extending out of this sub-network structure such as G-protein signalling regulator, vitellogenin, member from solute carrier family. Interaction of signal transducer and activator of transcription (STAT) with StAR, estrogen receptor alpha 2 with phosphodiesterase 1c (LOC799748) was reported earlier (Kanzaki et al. 1999; Etingof et al. 1984). A number of nodes such as nuclear receptor subfamily (NR2F1), keratine 18, ADP ribosylation factor (arf5), and transgelin 2 (tagln2) were interacting in a sub-networks. Some of the known sub-network structures were very consistent in the networks obtained by both GGM and Bayesian Network. StAR was 4-fold down-regulated, Cyp19a1a was 13-fold down-regulated, and Vitellogenin expression was 7-fold up-regulated during this exposure time (Figure 7).

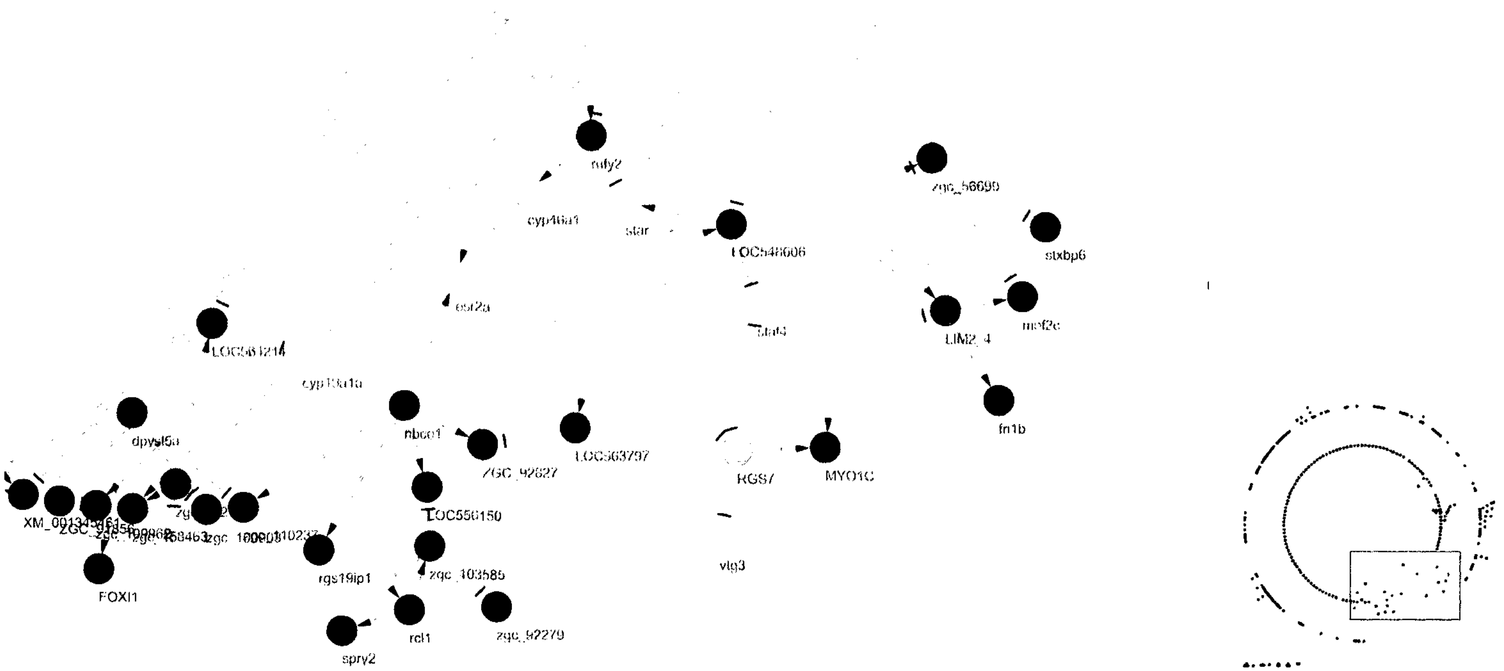


Figure 7: Transcriptional network at 96h exposure. Sub-network where StAR plays a central role, interacting with estrogen receptor alpha 2 and vitellogenin (in yellow), two cytochrome P450 members (in green), a transcription factor (triangular shape) . A number of interactions from this network were reported earlier.

Inference of a transcriptional network at 192h exposure

Genes with functions such as steroid metabolism and steroid hormone receptor, aromatase, estrogen related receptor (alpha), camp reposnive elements, and clusterin were found differentially expressed during this time exposure. The gene CLU (clusterin) which expresses in many tissues like testis, ovary and liver and is believed to be involved in many functions (e.g. lipid transportation, inhibitor of apoptosis, chaperon activity) displayed a central interacting role with genes related to steroidogenesis and hormonal signaling, including cyp19a1a, follistatin-like 1b (fstlb), transcription factor (tfap2b), zgc:162977, estrogen-related receptor (esrrap2), and frizzled related protein (frzb) (Figure 8). Expression of follistatin regulates follicular growth and may result in reduced FSH, impaired ovarian follicle development, and augmented ovarian androgen production (51). Frizzled related proteins are the modulators of Wnt-Frizzled signals. Overexpression of Wnt regulates the expression of many genes including aromatase and may perform important functions in the adult ovary (52, 53). Three significant pathways, gamma-Hexachlorocyclohexane degradation, Inositol phosphate metabolism, and Phosphatidylinositol signaling system, were found active during this exposure.

The network model from all four clusters were searched for biological information using data mining tools. The literature evidence confirmed many interactions in cluster 1 and 2. The network for each day was very densely organized. Querying these network against the known databases, we found coherent small subnetworks, some of which were consistent with known biological information. Genes dominating interactions in the subnetworks were related to steroid metabolism, estrogen receptor, transcription factor activities, and anatomical structure development. In the network model obtained from cluster approach, aromatase CYP19A1A showed strong correlation with elongation of long chain fatty acids (ELOVL6), estrogen receptor (ESR2A), vitamin D receptor (vdr), and follistatin-like 5 protein (FSTL5). Expression of FSTL5 may result in reduced FSH, impaired ovarian follicle development and augmented ovarian androgen production (10). GO overrepresentation with Bonferroni Family-Wise Error Rate correction ($p<0.05$) showed involvement of estrogen receptors along with vitamin D receptor (VDR), transcription factor nr2f1, two other genes in the steroid hormone receptor activity and interaction of cadherin with ATPase. The data suggests that VDR plays an important role in endocrine function (54). A previous study showed that cadherin is an instructive inducer of Na"/K"ATPases distribution.

Grouping algorithm with significance test performed on the genes for their GO term showed some distinct pattern of GO enrichments. During the initial exposure time of 24hr, genes with molecular function such as transcription factor activity and transcription regulator activity were significantly active. A number of genes were active in the developmental process and grouped in intracellular or membrane bound organelle. During 48hr of exposure a majority of genes were active in protein kinase activity, phosphorylation and protein metabolic process and were integral to membrane. Cytochome P450 members were also active during this period of exposure. During 96hr of exposure, a significant number of genes were involved in localization, cofactor biosynthesis, metal ion transport, and ion binding and were grouped as non-membrane bound organelles and cytoskeleton. Genes in 192hr were active in developmental process, cellular morphogenesis.

Each time point had a unique set of differentially expressed genes. Many steroid hormone receptors and their family members of nuclear transcription factors that are critical to the reproduction and differentiation had increased activities during the initial hours of the exposure. Exposure longer than 24h caused a significant increase in many cytochrome P450 family members, estrogen receptors, hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 7, and glutathione S-transferase pi and their involvement in many pathways such as metabolism of xenobiotics by cytochrome P450, tryptophan metabolism, Linoleic acid metabolism, and ovary infertility (55). Most of them were highly expressed during 96hr of time exposure. During the time exposure of 192hr the expression for estrogen receptors, cytochrome P450 members except aromatase, and StAR was very few or absent. Increased expression of aromatase, vitellogenin, CYP11A, and StAR are associated with the fadrozole exposure and can be used as a potential molecular marker.

Also, many of the network interactions recovered in our model, especially the interaction between the known biomarkers of fadrozole exposure, were reported earlier in the literature. Some additional perturbation in the co-expressed group of genes fine tuned and improved some of the sub-network structures. Since fathead minnow is not well annotated, removing the unknown genes from the network also improved the readability of the network. Majority of the interactions obtained by shrinkage approach of GGM were undirected while the interactions obtained from Bayesian Network were directed. Results from both GGM and Bayesian Network strongly agreed with established biological information.

Results from our study strongly suggest that fadrozole interferes with the estrogen activity in fathead minnow and that an exposure longer than 24hr may be required for aromatase inhibitor to respond and interact with the endocrine system.

Both GGM and BN methods had quite similar results, except that GGM was much faster, had more edges and showed node-distribution similar to power-law distribution; however there was insufficient evidence as to which method performed better. Both network models had several novel interactions with genes not previously found associated with any module or biochemical pathways. Incorporating prior knowledge and annotation information can be useful in reconstructing some known relationships and proposing some novel interactions.

Overall Network Using Mutual Information Theory

Mutual information is used to measure the nonlinear relationship between the expressions of two genes. Since these metrics are computed from a finite number of samples, a threshold is often imposed so that two nodes are connected if the computed metric between the two nodes is above the threshold. An efficient estimators is required that can accurately compute mutual information from the data. Mutual information network inference proceeds in two steps. The first step is the computation of the mutual information matrix (MIM). The second step is the computation of an edge score for each pair of nodes by an inference algorithm that takes the MIM matrix as input. Each mutual information calculus demands the estimation of three entropy terms. A fast entropy estimation is therefore essential for an effective network inference based on MI. In this study we have used the Miller-Madow estimator as described by Meyer et al. (2008).

A total of 8600 annotated genes were taken for mutual information (MI) calculation. MI was calculated for each pair of genes, and the weighted adjacency matrix with values was generated, where the higher the weight is; the higher is the evidence that a gene-gene interaction exists. Significant gene – gene weight matrix was imported into Cytoscape using the force-directed layout algorithm, which resulted in a modular MI network. Differentially expressed genes from individual time points were then mapped into this overall network to investigate further using the GO enrichment analysis. Within the cluster, some hub genes are: Sodium ion transport, nucleosome assembly, chloride transport, defense response. GO tern enrichment for the bigger cluster relates to steroid hormone receptor, inflammatory response, and immune receptor. Only greens present in the cluster suggests that 192hr exposure may not have recovered yet compared to other exposure samples (Figure 8 a-d).

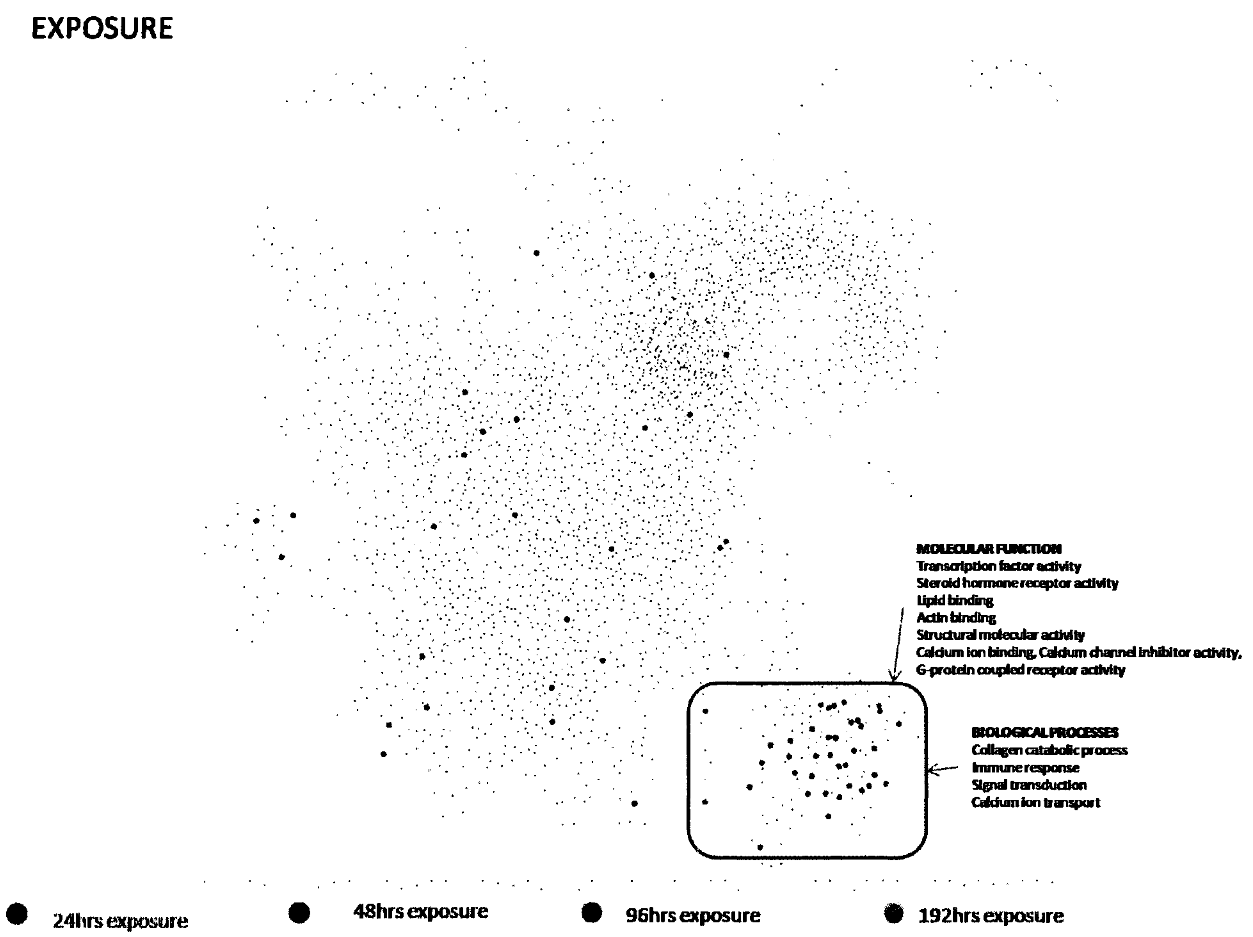

EXPOSURE
MOLECULAR FUNCTION
Transcription factor activity
Steroid hormone receptor activity
Lipid binding
Actin binding
Structural molecular activity
Calcium ion binding, Calcium channel inhibitor activity,
G-protein coupled receptor activity
BIOLOGICAL PROCESSES
Collagen catabolic process
Immune response
Signal transduction
Calcium ion transport
● 24hrs exposure
● 48hrs exposure
● 96hrs exposure
● 192hrs exposure

Figure 8(a): All four exposure samples were mapped into the overall network model. Genes from the day 2,4 and 8 samples tends to map in a subnetwork, which is involved in lipid biosynthesis, steroid hormone receptor activity, and signal transduction.

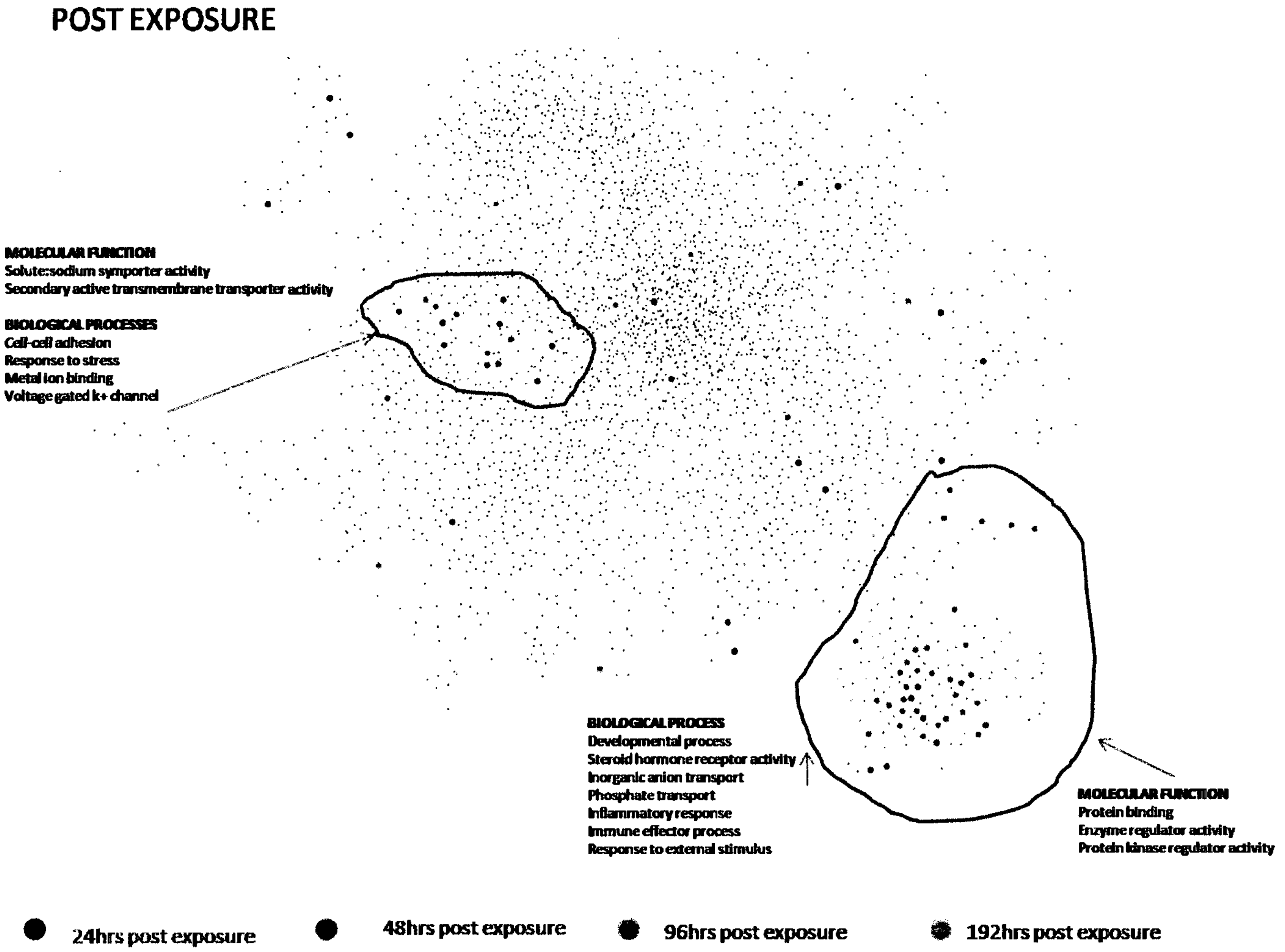


Figure 8(b): Most of the genes from day 8 were clustered together in the subnetwork, genes from day 4 were clustered separately.

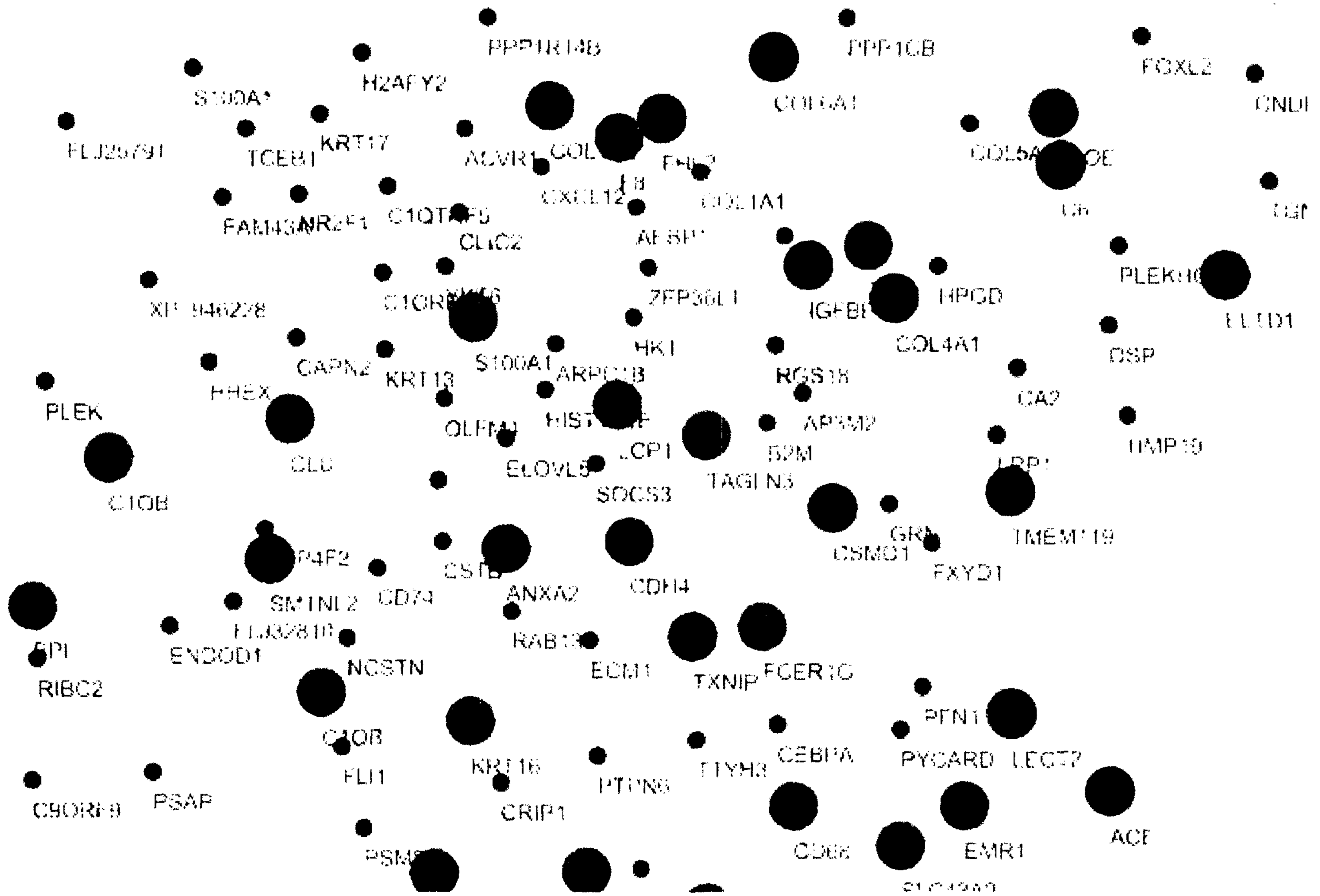


Figure 8(c): A close-up view of the subnetwork and the gene interactions,

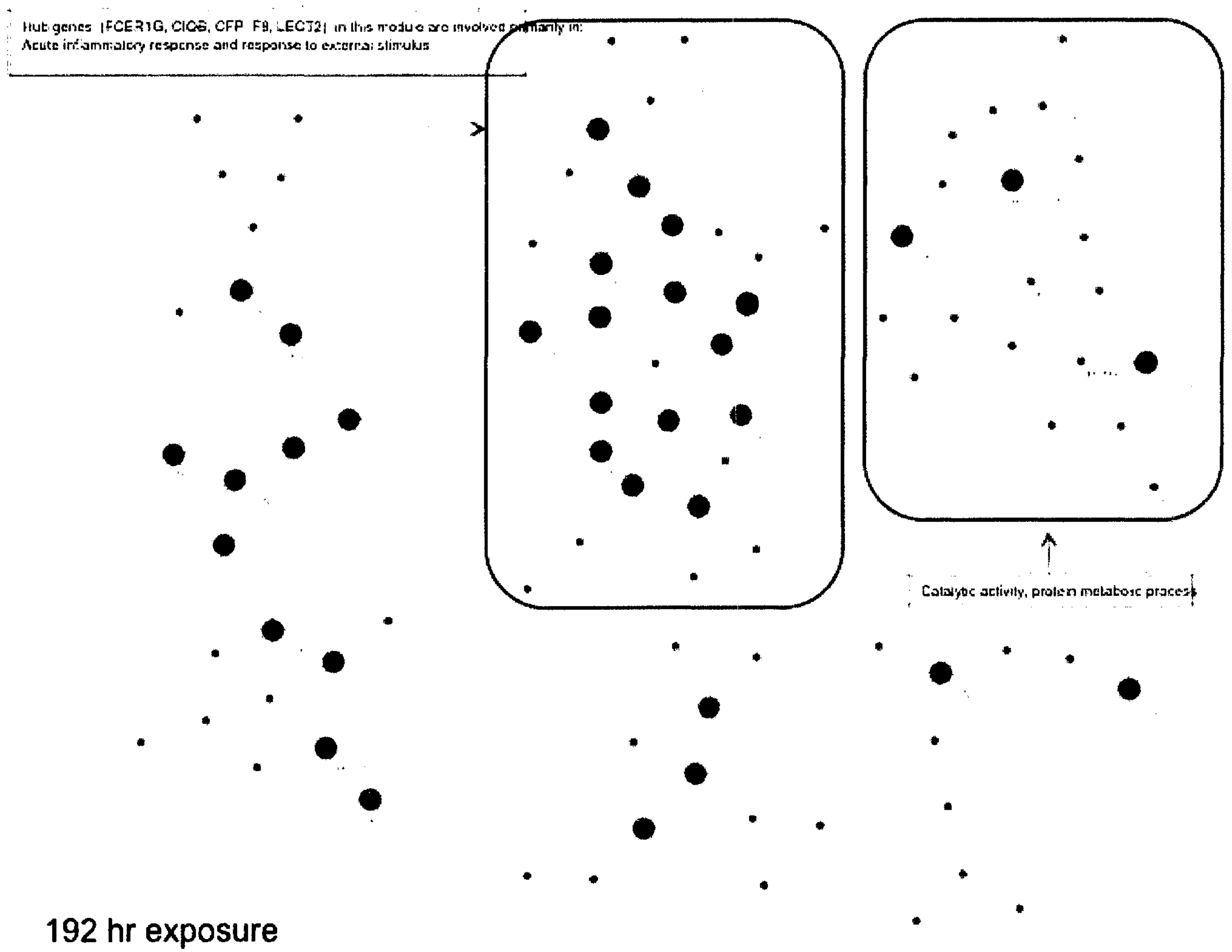

Hub genes (FCER1G, C1QB, CFP F9, LECT2) in this module are involved primarily in:
Acute inflammatory response and response to external stimulus
Catalytic activity, protein metabolic process
192 hr exposure

Figure 8(d): Network model for 192hr exposure shows the modular behavior. Many nodes had no annotation found in the public database. Interactions such as FCER1G and C1QB, C1QC have been reported earlier.

CHAPTER III

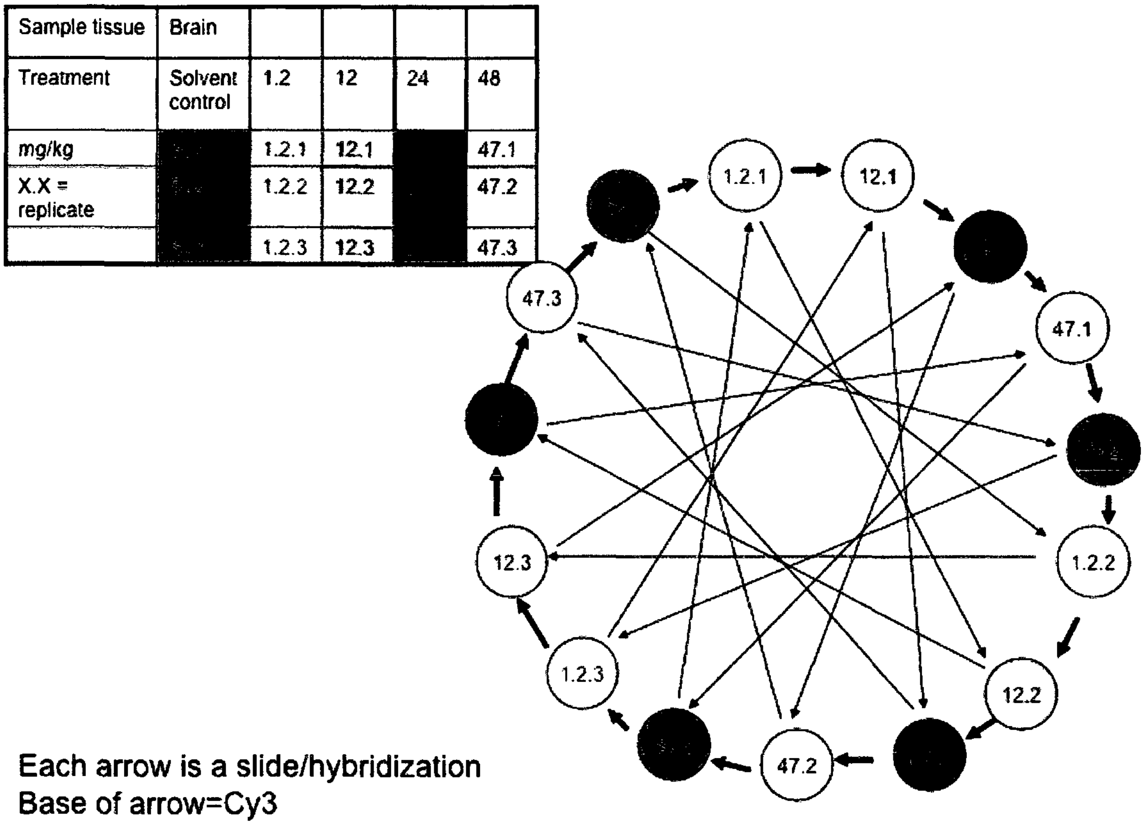
RAT BRAIN EXPERIMENT

Background

Over most of the last century, manufacturing, processing, and storage of the explosives, 2,4,6-trinitrotoluene (TNT) and hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX), have been responsible for extensive contamination of soil, as well as ground and surface water, throughout the U.S. and Europe. Unlike many other organic compounds possessing nitro- moieties, such as pesticides and various feedstock chemicals, these explosives are highly resistant to biological degradation, and are thus able to persist in the environment for long periods of time.

Hexahydro-1,3,5-trinitrotriazine (RDX) is the energetic compound used in military high explosives [56]. Residues of this compound are deposited onto the surface during live-fire training and are a common environmental pollutant from military exercises. Although poorly soluble in water, RDX and its metabolites were identified in water sources, including underground water resources [57]. They have raised health concerns and have already been reported to affect central nervous system in mammals by reversible seizure activity [58]. U.S. Environmental Protection Agency has classified RDX as a class C potential human carcinogen.

In this study, we examined recovery from 1,3,5-trinitroperhydro-1,3,5-triazine (RDX) induced seizures in *Rattus norvegicus* through changes in transcriptional networks. The goal of this work was to identify networks affected by chemical exposure and track changes in these networks as animals recover. We examined brain microarray data from *R. norvegicus* treated with 0, 1.2, 12, 24, and 47 mgRDX/kg body weight at different time points after exposure (24hr, 48hr, 7d, 14d, 28d and 90d). We focused mainly on the temporal gene expression data obtained from the 8k, two-color cDNA microarray using the extended loop design experiment (Figure 9). The experiment includes four technical replicates and three biological replicates.



Sample tissue	Brain				
Treatment	Solvent control	1.2	12	24	48
mg/kg	■	1.2.1	12.1	■	47.1
X.X = replicate	■	1.2.2	12.2	■	47.2
	■	1.2.3	12.3	■	47.3

Each arrow is a slide/hybridization
Base of arrow=Cy3
Arrowhead=Alexa647
All slides are labeled with sample replicate Cy3 vs A647

Figure 9: Interwoven loop design cDNA microarray

The analysis of genetic regulatory networks has received a major impetus from huge amount of data such as cDNA microarray. To fully understand the regulatory structures, different analysis tools will have to be used. To infer gene regulatory network, one general strategy is to learn functional associations among the genes, called 'guilt-by-association' strategy [59]. The advantage of taking this approach is that many of the available functional genomics data naturally describe relationships between genes, rather than directly correlate with functions.

The choice of the algorithm depends upon the model architecture as well as the quantity of the measured data [60]. Model selection is the most crucial step. A best model selection technique would be the one with a balance of goodness of fit and the complexity. To identify the structure of the network, an overall model fit measure is needed to assess how well a genetic network fits the data and to compare the merits of alternative network structure. The model fit measure allows us to rank genetic networks according to their ability to fit the observed data. Score-based approach in principal is more powerful. However, even a simple model can produce an incredible number of possible graphs [61]. It is nearly impossible to explore all the graph models to determine the network consistency; therefore, it is essential to include biological constraints to narrow down the complexity of network inference. Over the past years, many modeling methods have been proposed, and Bayesian network in particular has become very popular. Some methods exploit the prior knowledge on the network structure, while some focus on the conditional dependencies between the genes. Several approaches for gene regulatory modeling (GRN) have been proposed in the literatures and can be used. A majority of those modeling approaches describe graph mathematically as: Bayesian Networks (BN), Boolean Networks, Differential Equations, or Graphical Gaussian Models (GGMs). We have used BN in this study. There is a limitation with BN that it cannot capture feedback loops, which are essential in genetic networks. Dynamic Bayesian networks (DBN) can be used for time-series data, since DBN can capture the dynamic behavior of the network, and more importantly, they can describe the feedback mechanisms in the networks.

Materials and Methods

All chemical exposures and microarray experiment were conducted in the Environmental Laboratories (EL) at the US Army Corps of Engineers, Vicksburg, MS, USA.

Experimental Design

Female Sprague-Dawley rats (175-225 grams) used were from the in-house breeding colony (College of Pharmacy, University of Louisiana at Monroe [ULM]) and treated in accordance with the *Guide for Use and Care of Animals* (National Research Council 1996). Food was removed the night before dosing, which occurred the next morning between 8 and 11 AM. Rats were weighed then and were randomly assigned to treatment. Doses, which were administered by oral gavage, consisted of control (5% v/v DMSO in corn oil), RDX ranging from 1.2 to 47 mg/kg in 5% DMSO in corn oil emulsion. Animals were monitored for seizure activity after dosing and were euthanized using CO_2 if moribund as stated by OECD criteria (OECD 2000). Brains were flash frozen with liquid nitrogen, crushed with mortar and pestle over liquid nitrogen, placed in RNA Ice overnight, and then frozen at -80 degrees C. RNA was then extracted.

Total RNA from three biological replicates at each dose was compared using the two-color interwoven loop design microarray experiment [62]. cDNA from 1mg total RNA was synthesized, hybridized to arrays, and detected by secondary hybridization to Alex647 and Cy3 dendrimer oligonucleotides using an Array900 detection kit per manufacturer's instruction. cDNA was hybridized to 8k Sigma/Compugen rat 70-mer oligonucleotide libraries arrayed on glass slides (http://www.cag.icph.or/). After hybridization, slides were scanned using a 5-micron ChipReader microarray reader (BioRad Hercules, CA) [62].

Microarray Data Preprocessing and Analysis

Raw intensity data was obtained from image analysis program GenePix and was imported into R package "Bioconductor". Print tip group loess was applied within the array and Quantile normalization was used between the arrays (Figure 10).

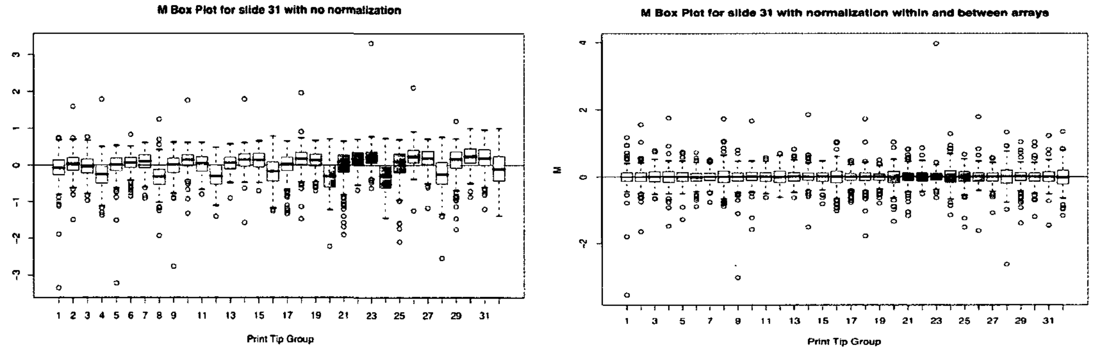


Figure 10: Box plot after the normalization

The normalized log (intensity) and log (ratio) values were exported, and missing values were estimated using least square principle [63] and introduced to Bayesian Analysis of Gene Expression (BAGEL) [64] model for identifying the differentially expressed genes (DEG).

Transcriptional Network Modeling

We compared two algorithms to check the consistency in the network models and also to find some interesting interactions from the networks from the two methods.

Overall DEG and the DEG from each time were modeled with GGM and Mutual Information Relevance Network.

A more promising machine learning method is given by GGMs. These models are based on the assumption that the data are distributed according to a multivariate Gaussian distribution N(m, Σ). But to avoid the shortcomings of relevance networks based on Pearson correlation coefficients, partial correlations are considered in Gaussian graphical models. That is, the strength of a direct association between two nodes X_i and X_j is measured by the partial correlation coefficient $p_{i,j}$, which describes the correlation between these two nodes that is conditional on all the other nodes in the system. From the theory of normal distributions, it is known that the partial correlation coefficient $p_{i,j}$ can be computed from the inverse $\Omega = \Sigma^{-1}$ of the covariance matrix Σ via

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

where $\omega_{i,j}$ are the elements of the matrix Ω.

The disadvantage of this procedure is that the empirical covariance matrix can only be inverted if the number of observations exceeds the number of nodes in the network, that is, if the matrix is nonsingular.

To learn a Gaussian graphical model from such data, Schafer and Strimmer [37] have proposed the application of a shrinkage covariance estimator. The shrinkage estimator $\sum'$ is guaranteed to be nonsingular so that it can be inverted to obtain a new estimator $\Omega' = (\sum')^{-1}$ for the matrix Ω. In order to test the significance of the correlations, the "local fdr" network search was employ as proposed by Schafer et al. The local fdr is an empirical Bayes estimator of the false discovery rate. This method computes the posterior probability for an edge to be present or absent, and takes account of the multiplicity in the simultaneous testing of edges.

The method of relevance networks, proposed in Butte et al. [19], is exclusively based on pairwise association scores and therefore represents a very simple machine learning approach to reverse engineering regulatory networks. An association score is computed for all pairs of variables Xi and Xj (i, j 2 ∈ {1, . . ., n}) from the data. For continuous data, the Pearson correlation coefficient can be used:

$$corr\ (x, y) = \frac{(\frac{1}{m}) \sum_{i=1}^{m} (xi - x)(yi - y)}{\sqrt{(\frac{1}{m}) \sum_{i=1}^{m} (xi - x)2} \sqrt{(\frac{1}{m}) \sum_{i=1}^{m} (yi - y)2}}$$

where x=(x_1, ...,x_n) and y=(y_1, ...,y_m) are the m-dimentional observations of two different variables with empirical means x and y. Interpreting the variables as the nodes of a network, the pairwise association scores are compared with a predefined threshold value, and the nodes whose pairwise association scores exceed this threshold are linked by an undirected edge.

In a relevance network, the interactions are not inferred within the context of the whole system, that is, there is no distinction between direct and indirect interactions. Not rarely does a high correlation coefficient between two nodes indicate only a pseudo-association, for example, if both nodes depend on a common regulator. Hence, a high correlation coefficient between two nodes does not necessarily indicate a direct association, and with regard to the graphical representation of the network, only the direct interactions are of interest.

In statistical terminology, a relevance network based on the Pearson correlation is referred to as a covariance graph. The threshold value can be estimated by a randomization test so as to keep the number of false positive edges below an a priori specified tolerance level. Alternatively, instead of the Pearson correlation, the mutual information can be used to compute the pair-wise association scores in relevance networks. Mutual information scores can be computed for discrete variables only, so that continuous data have to be discretized; this incurs a certain information loss. But an advantage is that this score can deal with nonlinear associations [19].

Network was fine-tuned based on the node scoring and density cutoffs using the cytoscape plug-in MCODE [39]. Sub-networks were generated from the high cluster seed. Sub-networks created by this method are easier to investigate the interactions and compare them with the biological information. Networks were searched for any set(s) of recurring regulatory pattern called network motifs, and degree distribution and other network properties were calculated.

Biological Network and GO classification

The growing database of biological data includes information discovered by methods such as direct and experimental while others are indirect, predictive, and computational. Instances of such interactions are the observed or predicted relationships between genes and proteins, and they can be represented as networks of functional association [67]. Published databases such as BIND [68], KEGG [69], Predictome [70], and STRING [71] provide the conceptual platforms on which software for leveraging the full content of the interactome could operate. STRING uses conserved genomic neighborhood arrangements of genes to infer functional linkage. It is more error tolerant when assembling conserved neighborhoods, ignoring short, partially overlapping genes on the antisense strand that are likely to be spurious predictions.

Exploring Gene Ontology annotations is a common and widespread practice to get first insights into the potential biological meaning of the experiment in structured and controlled classifications. The Gene Ontology Consortium defines GO as an international standard to annotate genes [72]. Exploring all the three domains, biological process, molecular function, and cellular component in a term-to-term approach has a drawback that it does not respect dependencies between the GO terms that are caused by overlapping annotations. A parent–child approach is a statistical analysis of GO term overrepresentation that examines each term in the context of its parent terms could be used [73].

Results and Discussion

The regulatory networks based on mathematical models and biological networks based on existing biological information were obtained as described in the Method. The network models were represented as directed and undirected graphs with edges between them representing the mode of activation, repression, or unknown.

The differentially expressed genes were identified (see Table 2) and were further investigated for their Gene Ontology over-representation, biochemical pathways, and transcriptional network models. Dose response shows that more genes were expressed when treated with 24mg of RDX at all time points. With a threshold of p-value<0.05 and fold change>=1.5, a total of 937 genes were found significantly expressed across all six-time points. The number of DE genes was very high at day 7 and lowest in day 14, but clustering on conditions revealed a strong batch effect on day 7 (Figure 11). The data was adjusted to remove batch effects using the COMBAT software developed by Johnson et al. [74].

Table 2: Summary table of the differentially expressed (DE) genes obtained using the Bayesian analysis and comparison between each time point.

Day (DE genes)	2	7	14	28	90
1 (167)	47	28	2	23	23
2 (260)		41	1	28	36
7 (494)			1	18	66
14 (10)				1	1
28 (97)					19
90 (320)					

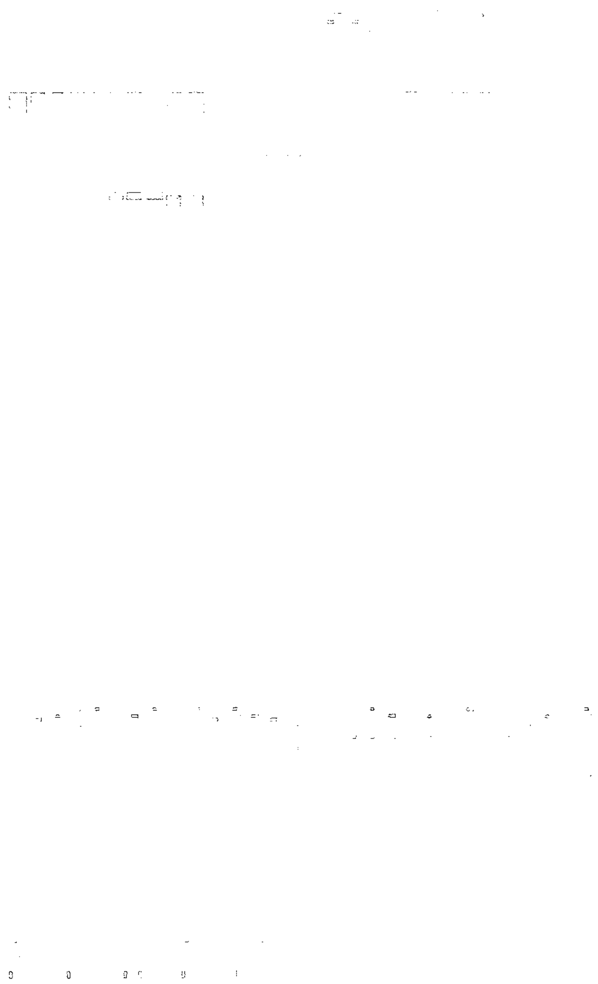


Figure 11: Condition tree across all samples shows a batch effect on Day 7.

Majority of the DE genes were related to GABA receptors, glutamate receptors, dopamine receptors, cholinergic receptors, Na/K ATPase exchanger, cytochrome P450 family members, solute carrier family members. Genes from individual time points were also analyzed separately to find the functional differences between the early exposures versus late exposure genes. Genes with functions such as Glutamate aspartate transporters, GABA receptors, Na/K transporting ATPases, Ca ATPase, cholinergic receptors, calmodulin dependent protein kinases, interleukin, and heat-shock proteins were differentially expressed during the early exposure. Genes involved in the first two exposures were very similar in terms of the gene-functions such as calcium binding, voltage gated channel members, cholinergic receptors, dopamine receptors, calmodulin dependent kinases. No neurotransmitter gamma-aminobutyric acid (GABA) was expressed in the second day exposure. Day 7 had many genes related to glycoprotein hormone, gap junction membrane channel proteins, dopamine receptors, thyroid stimulating hormone, and follicle stimulating hormone, and luteinizing hormone, somatostanin receptors, chemokine receptors, olfactory receptors family members, protocadherin family members, and few cytochrome P450 members were expressed. Dopamine receptors, chemokine receptors, calmodulin dependent protein kinases, and many cytochrome P450 members were expressed on day 14, 28 and 90.

Many ion channels, Na/K and Ca ATPases, neurotransmitter inhibitors such as Gabrg3, Gabrr1, Gabrr3, Gabra4, Gabra6, glutamate receptors, and dopamine receptors were differentially expressed during the first two time points (day 1 and day 2). More of chemokine receptors, calmodulin receptors, and cytochrome P450 members such as Cyp3A, Cyp2B, Cyp24B, Cyp11A, and Cyp11B were differentially expressed in the late time points (day 7, 28, and day 90). Exposures of RDX to the rat brain and our result from early time point suggest that RDX might trigger freeze messengers, also called inhibitory neurotransmitters, such as the GABA receptors. Inhibitory neurotransmitters allow chloride to enter the ion channel, which freezes the next neuron and makes it harder to excite. Excitatory neurotransmitters allow sodium to enter the ion channel, which excites the neuron and makes it pass the message. CYPs appear to have specific functions in brain (e.g. regulation of levels of endogenous GABA receptors). The role of

CYPs in the brain, a highly heterogenous and complex organ, is a relatively unexplored field of scientific enquiry. It holds promise for furthering our understanding of inter-individual variability in response to centrally acting drugs as well as risk for neurological diseases.

GO Classification and Pathway Analysis

All the differentially expressed genes were analyzed for the GO associations. In order to find the GO terms that are statistically significant within the group, a control set of genes needs to be used to obtain a total count of occurrences of each GO term. We used *Rattus norvegicus* database as the background using DAVID tool. For each GO term, a p-value was calculated representing the probability that the observed numbers of counts could have resulted from randomly distributing this GO term between the tested group and the background or reference group. The Benjamini and Hochberg correction method was used to control the false discovery rate. Functional annotation clustering revealed that the large number of genes involved in the GO terms such as development processes, cell to cell signaling and communication, signal transduction, transmission of nerve impulse, synaptic transmission, neurological system processes, and calcium ion homeostasis (Figure 12).

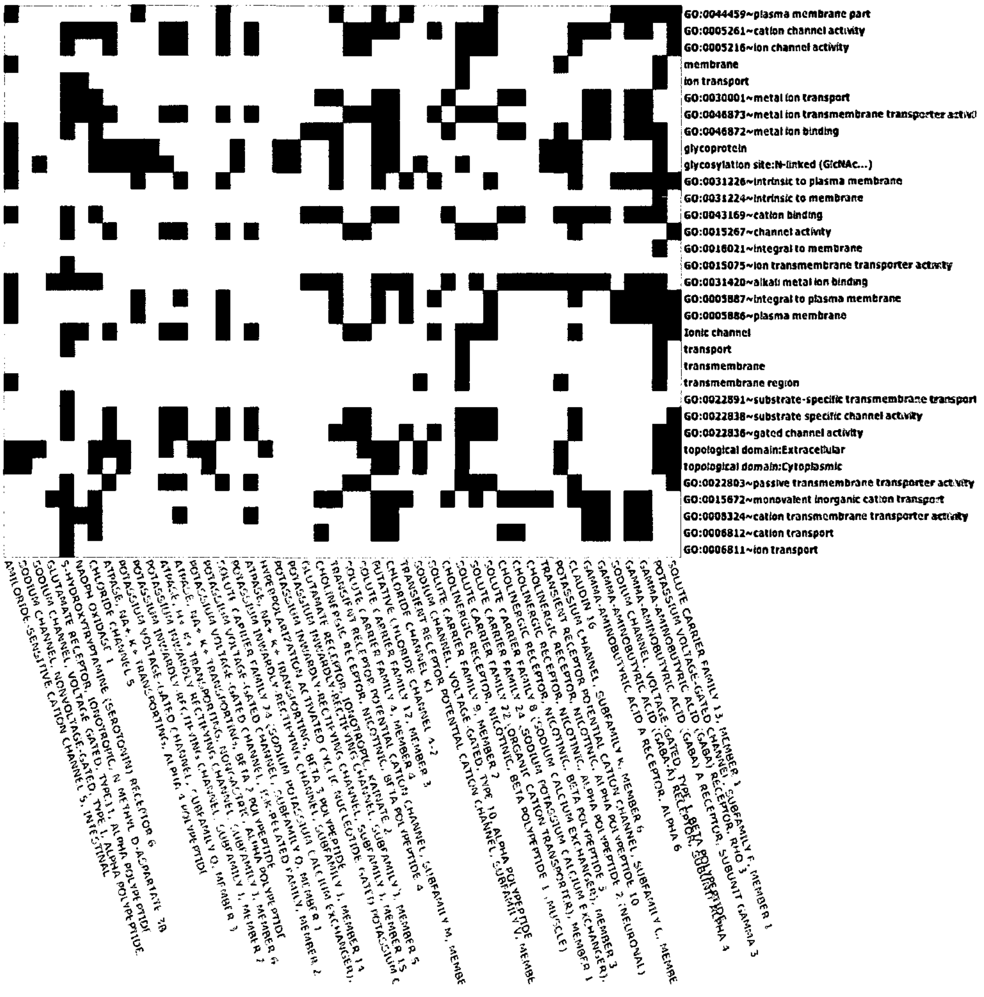


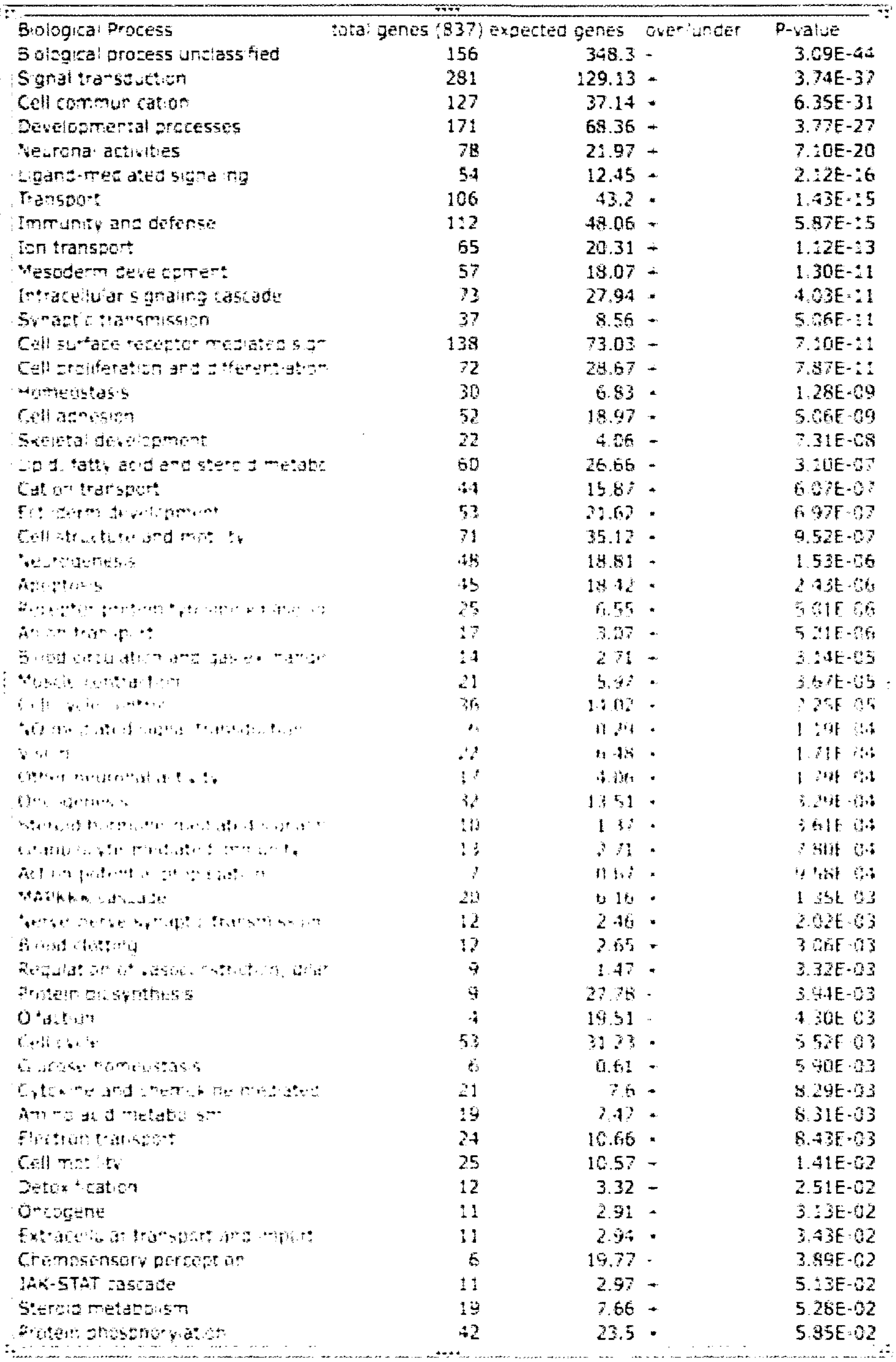
Figure 12: Best cluster from a GO enrichment analysis shows the majority of the DE gene in this cluster involved in ion channel activity.

Signal transduction events include cell communication, cell surface receptor mediated signal transduction, intracellular signaling cascade, and steroid hormone-mediated signaling. Nearly 80 genes were classified in the neuronal activities, which include action potential propagation and synaptic transmission (See Table 3 a-b for Biological Processes and Molecular Function enrichment (p<=0.05)). Overrepresented genes in the molecular functions were receptors, signaling molecules, Ion-channel, and kinase activities. A total of 129 genes classified as receptors such as cytokine, G-protein coupled, Immunoglobin, Ligand-gated ion channel, nuclear hormone, and protein kinase receptors.

Nearly 56% of the total DE genes on day 1 were up-regulated. A threshold of minimum 2 genes and "Benjamini" multiple testing correction ($p<0.05$) resulted in 23 genes from day 1 involved in 4 different pathways (Figure 13a-b): Neuroactive ligand-receptor interaction, Neurodegenerative diseases, GnRH signaling pathway, and Huntington's disease. Two neurotransmitters gamma-aminobutyric acid A receptor, subunit alpha 6 (Gabra6), gamma-aminobutyric acid B receptor 1 (Gabbr1), and neurohormone receptor (GnRHR) were up-regulated and found interacting with other 8 genes in GPCRs. The gonadotropins induce ovulation and stimulate estradiol and progesterone production, which in turn, bind to specific amygdaloid hormone receptors and influence neural activity including epileptiform discharges. Amyloid beta precursor protein (APP) was interacting with glutamate receptor (GluR), glutamate-cysteine ligase (Gclm), Snap25, and Camk2a. Calcium signaling is crucial for several aspects of plasticity at glutamatergic synapses. Increased MAPK3 and GABA receptor activities in neuron can be correlated with neuronal seizures. GO network also revealed the involvement of genes in terms such as "transmission of nerve impulse", "neurological system process", "synaptic transmission", etc. (Figure 14).

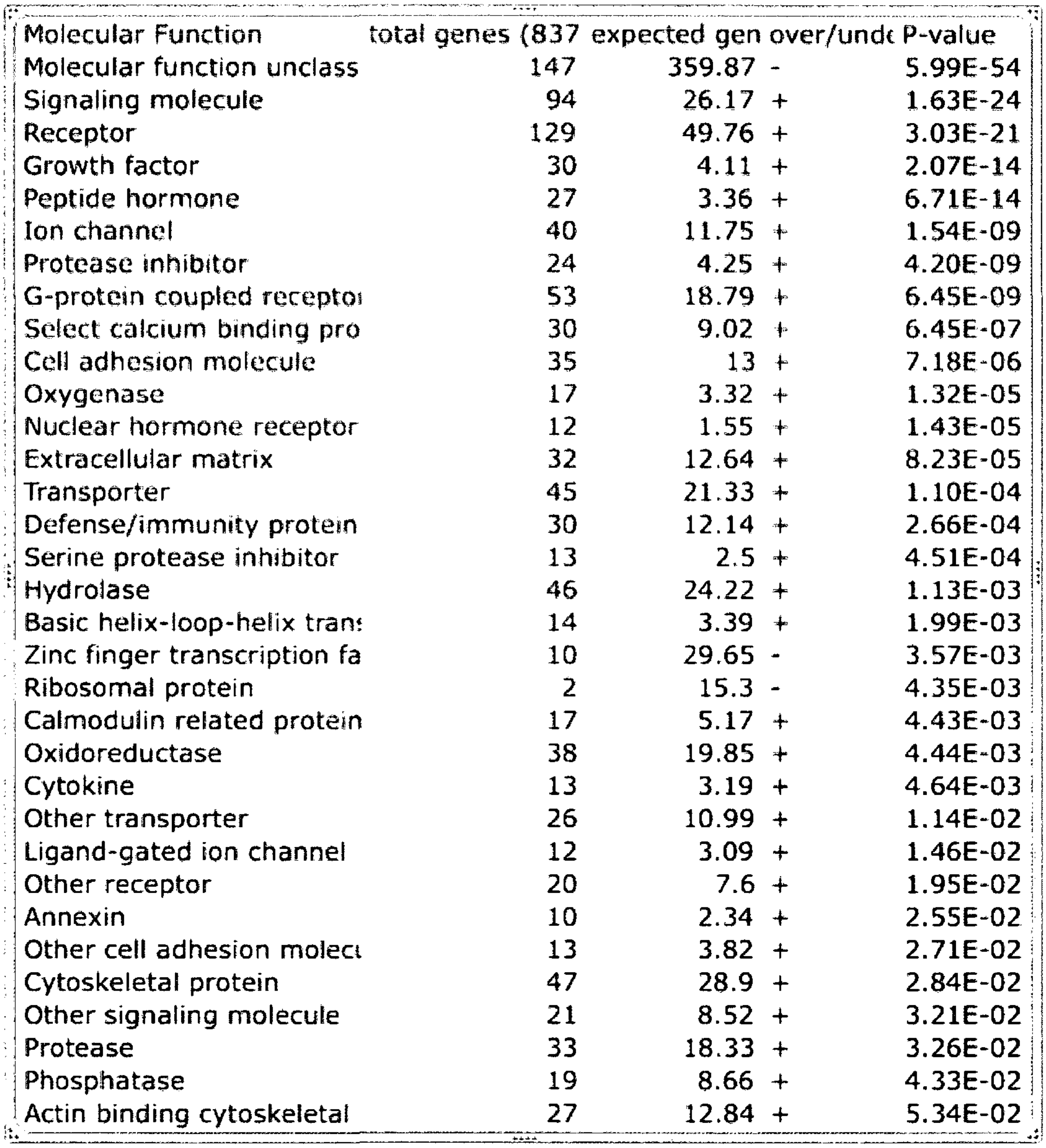
Approximately the same % of genes were up-regulated on day 2 but no neurotransmitters were differentially expressed, which confirms our observation that we did not see rats getting seizures after day 1. More genes were found involved in two significant pathways we found significant on day 1: Neuroactive-ligand receptor interaction and neurodegenerative diseases. Overrepresented GO terms for gene sets from day 28 and 90 were found very similar. They were mostly enriched as signal transduction, neuronal activities, neurogenesis, and sensory perception. Molecular functions for day 90 were involved in protease inhibitor, signaling molecules, and ion channels.

Table 3a: GO Biological Process enrichment for the DE genes shows which Gene Ontology (GO) terms are significantly overrepresented (+) in a set of genes.



Biological Process	total genes (837)	expected genes	over/under	P-value
Biological process unclassified	156	348.3	-	3.09E-44
Signal transduction	281	129.13	+	3.74E-37
Cell communication	127	37.14	+	6.35E-31
Developmental processes	171	68.36	+	3.77E-27
Neuronal activities	78	21.97	+	7.10E-20
Ligand-mediated signaling	54	12.45	+	2.12E-16
Transport	106	43.2	+	1.43E-15
Immunity and defense	112	48.06	+	5.87E-15
Ion transport	65	20.31	+	1.12E-13
Mesoderm development	57	18.07	+	1.30E-11
Intracellular signaling cascade	73	27.94	+	4.03E-11
Synaptic transmission	37	8.56	+	5.06E-11
Cell surface receptor mediated sign	138	73.03	+	7.10E-11
Cell proliferation and differentiation	72	28.67	+	7.87E-11
Homeostasis	30	6.83	+	1.28E-09
Cell adhesion	52	18.97	+	5.06E-09
Skeletal development	22	4.06	+	7.31E-08
Lipid, fatty acid and steroid metabo	60	26.66	+	3.10E-07
Cation transport	44	15.87	+	6.07E-07
Ectoderm development	53	21.62	+	6.97E-07
Cell structure and motility	71	35.12	+	9.52E-07
Neurogenesis	48	18.81	+	1.53E-06
Apoptosis	45	18.42	+	2.43E-06
Receptor protein tyrosine kinase sig	25	6.55	+	5.01E-06
Anion transport	17	3.07	+	5.21E-06
Blood circulation and gas exchange	14	2.71	+	3.14E-05
Muscle contraction	21	5.97	+	3.67E-05
Cell cycle control	36	14.02	+	7.25E-05
NO mediated signal transduction	6	0.24	+	1.29E-04
Vision	22	6.48	+	1.71E-04
Other neuronal activity	17	4.06	+	1.79E-04
Oncogenesis	32	13.51	+	3.29E-04
Steroid hormone-mediated signaling	10	1.37	+	3.61E-04
Calcium mediated signaling	13	3.71	+	7.80E-04
Action potential propagation	7	0.67	+	9.58E-04
MAPKKK cascade	20	6.16	+	1.35E-03
Nerve-nerve synaptic transmission	12	2.46	+	2.02E-03
Blood clotting	12	2.65	+	3.06E-03
Regulation of vasoconstriction, dilat	9	1.47	+	3.32E-03
Protein biosynthesis	9	27.78	-	3.94E-03
Olfaction	4	19.51	-	4.30E-03
Cell cycle	53	31.23	+	5.52E-03
Glucose homeostasis	6	0.61	+	5.90E-03
Cytokine and chemokine mediated	21	7.6	+	8.29E-03
Amino acid metabolism	19	7.47	+	8.31E-03
Electron transport	24	10.66	+	8.43E-03
Cell motility	25	10.57	+	1.41E-02
Detoxification	12	3.32	+	2.51E-02
Oncogene	11	2.91	+	3.13E-02
Extracellular transport and import	11	2.94	+	3.43E-02
Chemosensory perception	6	19.77	-	3.89E-02
JAK-STAT cascade	11	2.97	+	5.13E-02
Steroid metabolism	19	7.66	+	5.26E-02
Protein phosphorylation	42	23.5	+	5.85E-02

Table 3b: GO Molecular Function enrichment for the DE genes shows which Gene Ontology (GO) terms are significantly overrepresented (+) in a set of genes.



Molecular Function	total genes (837	expected gen	over/und	P-value
Molecular function unclass	147	359.87	-	5.99E-54
Signaling molecule	94	26.17	+	1.63E-24
Receptor	129	49.76	+	3.03E-21
Growth factor	30	4.11	+	2.07E-14
Peptide hormone	27	3.36	+	6.71E-14
Ion channel	40	11.75	+	1.54E-09
Protease inhibitor	24	4.25	+	4.20E-09
G-protein coupled recepto	53	18.79	+	6.45E-09
Select calcium binding pro	30	9.02	+	6.45E-07
Cell adhesion molecule	35	13	+	7.18E-06
Oxygenase	17	3.32	+	1.32E-05
Nuclear hormone receptor	12	1.55	+	1.43E-05
Extracellular matrix	32	12.64	+	8.23E-05
Transporter	45	21.33	+	1.10E-04
Defense/immunity protein	30	12.14	+	2.66E-04
Serine protease inhibitor	13	2.5	+	4.51E-04
Hydrolase	46	24.22	+	1.13E-03
Basic helix-loop-helix tran	14	3.39	+	1.99E-03
Zinc finger transcription fa	10	29.65	-	3.57E-03
Ribosomal protein	2	15.3	-	4.35E-03
Calmodulin related protein	17	5.17	+	4.43E-03
Oxidoreductase	38	19.85	+	4.44E-03
Cytokine	13	3.19	+	4.64E-03
Other transporter	26	10.99	+	1.14E-02
Ligand-gated ion channel	12	3.09	+	1.46E-02
Other receptor	20	7.6	+	1.95E-02
Annexin	10	2.34	+	2.55E-02
Other cell adhesion molec	13	3.82	+	2.71E-02
Cytoskeletal protein	47	28.9	+	2.84E-02
Other signaling molecule	21	8.52	+	3.21E-02
Protease	33	18.33	+	3.26E-02
Phosphatase	19	8.66	+	4.33E-02
Actin binding cytoskeletal	27	12.84	+	5.34E-02

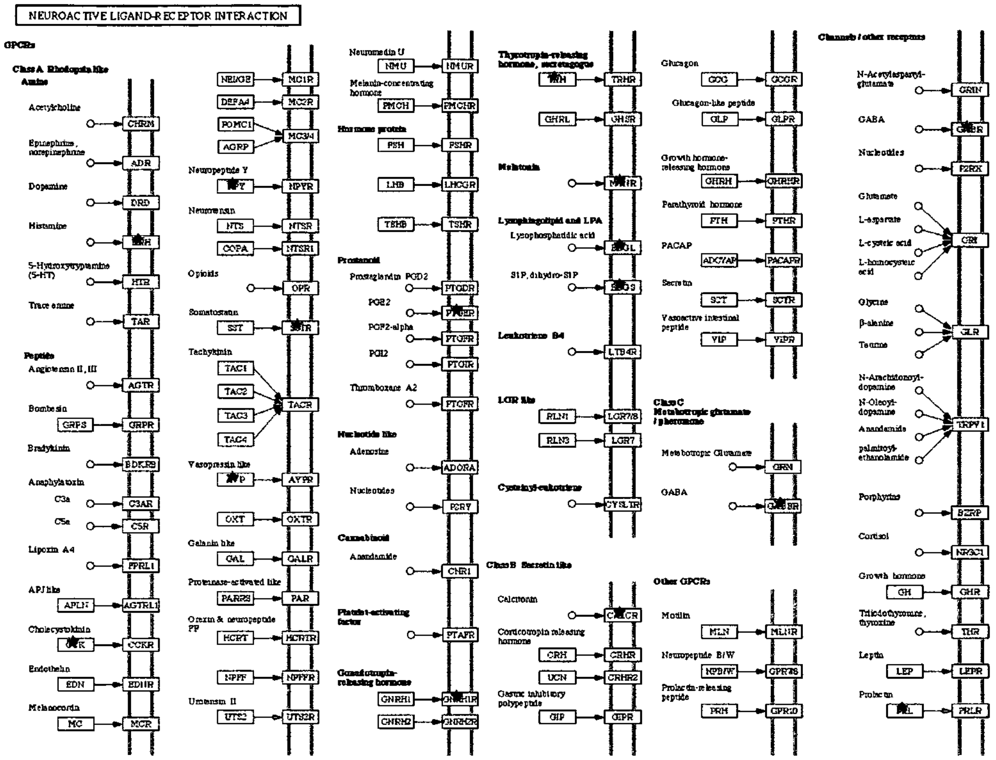


Figure 13-a: Total of 11 genes enriched (red stars) in the G-coupled receptors (p<0.05)

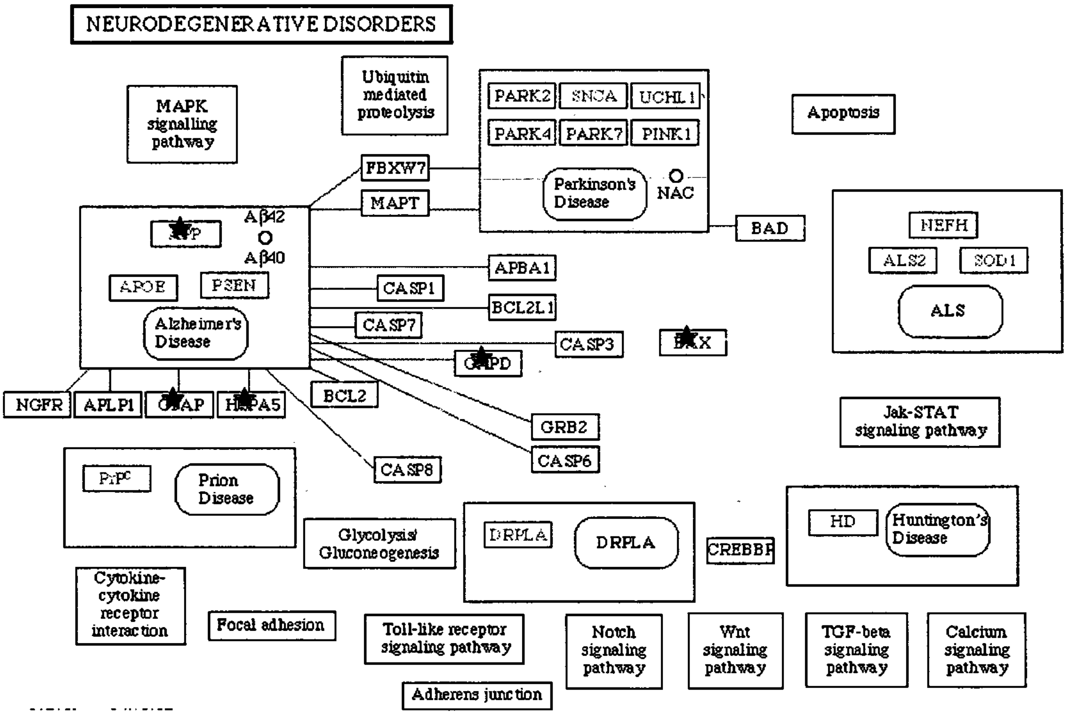


Figure 13-b: Gene enrichment in neurodegenerative disorder (red stars)

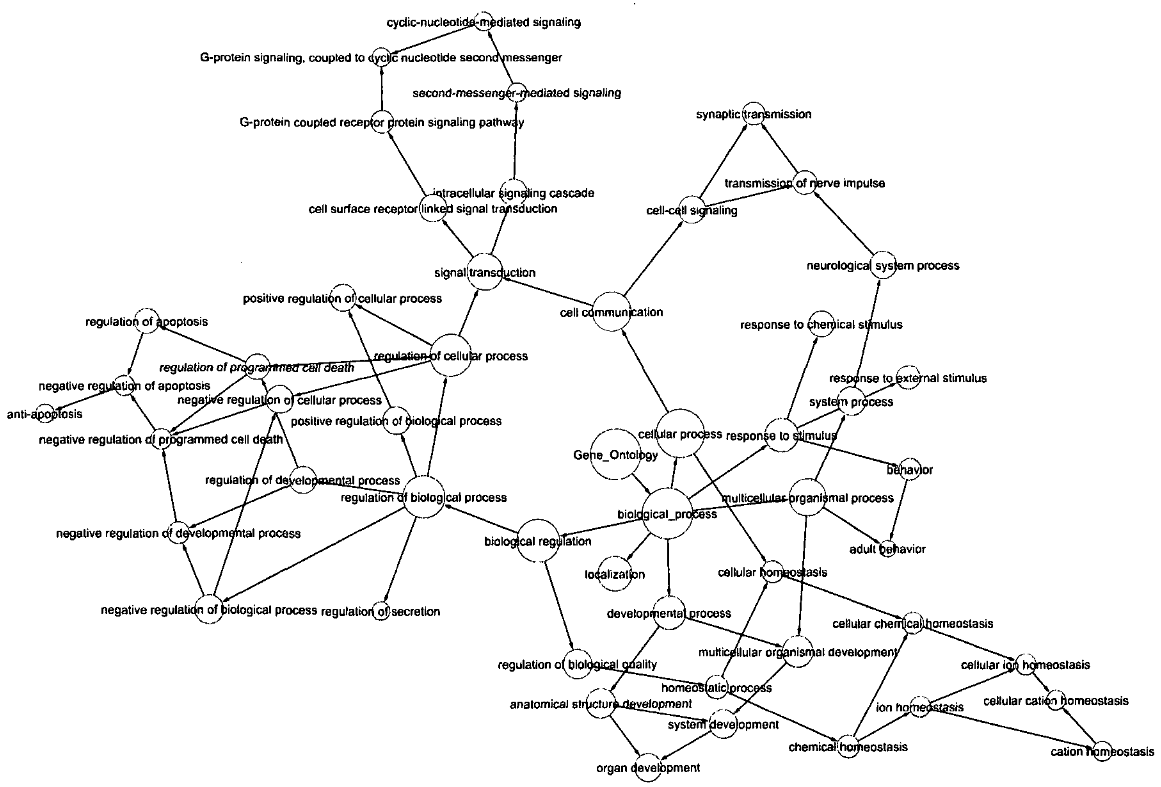

cyclic-nucleotide-mediated signaling
G-protein signaling, coupled to cyclic nucleotide second messenger
second-messenger-mediated signaling
G-protein coupled receptor protein signaling pathway
intracellular signaling cascade
cell surface receptor linked signal transduction
signal transduction
synaptic transmission
transmission of nerve impulse
cell-cell signaling
neurological system process
cell communication
positive regulation of cellular process
regulation of apoptosis
regulation of programmed cell death
regulation of cellular process
negative regulation of apoptosis
negative regulation of cellular process
anti-apoptosis
positive regulation of biological process
negative regulation of programmed cell death
regulation of developmental process
regulation of biological process
negative regulation of developmental process
negative regulation of biological process
regulation of secretion
response to chemical stimulus
response to external stimulus
system process
cellular process
response to stimulus
Gene_Ontology
behavior
multicellular organismal process
biological_process
biological regulation
adult behavior
localization
cellular homeostasis
developmental process
cellular chemical homeostasis
multicellular organismal development
regulation of biological quality
cellular ion homeostasis
homeostatic process
cellular cation homeostasis
anatomical structure development
ion homeostasis
system development
chemical homeostasis
cation homeostasis
organ development

Figure 14: Network representation of the genes overrepresented in the GO classification system. Each gene can have several associated GO terms, and due to the hierarchical structure of the GOs. Each GO term can be connected to several other GO terms higher in the hierarchy and therefore associated with the gene as well. Darker the color of the node in the above figure, more number of genes associated with that term.

Network Reconstruction

We used two alternative approaches for the evaluation of the reconstructed network model, based on mutual information and gene association network.

Mutual information relevance network. A relevance network is a group of genes whose expression profiles are highly predictive of one another. Each pair of genes related by a correlation coefficient larger than a minimum threshold and smaller than a maximum threshold is connected by a line. Groups of genes connected to one another are referred to as networks. The correlation coefficient between genes is calculated by comparing the expression pattern of each gene to that of every other gene. The ability of each gene to predict the expression of each other gene is measured as a correlation coefficient. Genes are represented as nodes in a network and edges are drawn between them if their correlation coefficient falls between the minimum and maximum thresholds specified in the initialization dialog. The system developed makes no prior assumptions about the underlying models linking gene expression but develops functionally relevant groupings of genes across the conditions.

We used the dataset containing 938 differentially expressed genes at 6 time points for pairwise calculations of the mutual information between them. Measurements of all genes were compared against each other, resulting in 271,183 total pairwise calculations of mutual information, ranging from 0.1 to 0.97. The number and size of the relevance networks increases with the decrease in the mutual threshold (Table 4). We set the threshold to 0.90, which produced 14 subnetworks with 565 interactions using a total of 192 genes (Figure 15a). Subnetworks were ranked based on the number of linked genes in the clusters. The associations between the genes in the networks were validated using the biological literatures. Subnetwork #1 was found to link with 159 other genes. GO Molecular Function enrichment analysis with significance cutoff <0.05 showed a large subset of genes involved in functions such as neuro-transmitting receptors, glutamate receptor, potassium voltage-gated receptors, and olfactory receptors. Top hub genes and their interacting genes were found to be involved in signal transducer activity and receptor activity. Laminin (Lama5), lectin-galactoside binding protein (Lgals4), Serine-proteinase inhibitor (serpina3m & serpina10), hemochromatosis (Hfe), forkhead boxes (Foxd4 & Foxe3), glutamate receptor onotropic, kainate 2 (Grik2), and GABA (A) receptors (Gabrg3, Gabra4, Gabrr3) appeared as some of the hub genes in the network. The major excitatory and inhibitory neurotransmitters in the brain, glutamate and GABA, activate both ionotropic (ligand-gated ion channels) and metabotropic (G protein-coupled) receptor, and are generally associated with neuronal communication in the mature brain. Biological literature also suggests that elevated expression of laminin may play a role in the development of epileptic seizures in patients with intractable epilepsy. KEGG pathway enrichment analysis also suggests their involvement in pathways such as neuro-active ligand receptor interaction and hematopoietic cell lineage and in cytokine-cytokine receptor interaction. The linked genes in the network #2 were involved in structural molecular activity. Network 5 connected Ache, Akap9, and Mylk2, which are known to be involved in catalytic activities. Network 6 linked Clcn5, chloride channel 5, Nox1, nadph oxidase 1 and Bsn, a presynaptic cytomatrix protein. This exact interaction has been reported in the literature as chloride channel prevents Nox-induced accumulation of negative charges in the endosomal lumen. Few networks contained various types of links, including a few associations not presently explained in the biological literature. The genes such as GABBR1, gastrulation brain homeobox 2(TF), calcium/calmodulin dependent protein kinase (CAMK2A), cholinergic receptors, and glutamate cysteine ligase modifier subunits formed a sub-network.

MI calculation using shrinkage entropy estimator. Shrink estimator combines two different estimators, one with low variance and the other with low bias by using a weighting factor l $\in$ [0,1]. Shrinkage is a general technique to improve an estimator for a small sample size. As the value of l tends to one, the estimated entropy is moved toward the maximal entropy, whereas when l is zero the estimated entropy tends to the value of the empirical one (Hausser J, 2006).

Table 4: Mutual Information threshold search for relevance network.

Mutual Information	Links	Subnetworks
0.97	0	0
0.95	8	8
0.95	9	8
0.93	116	14
0.93	64	19
0.91	343	14
0.90	454	13
0.90	**565**	**14**
0.85	3105	9
0.75	15297	18
0.1	251222	1

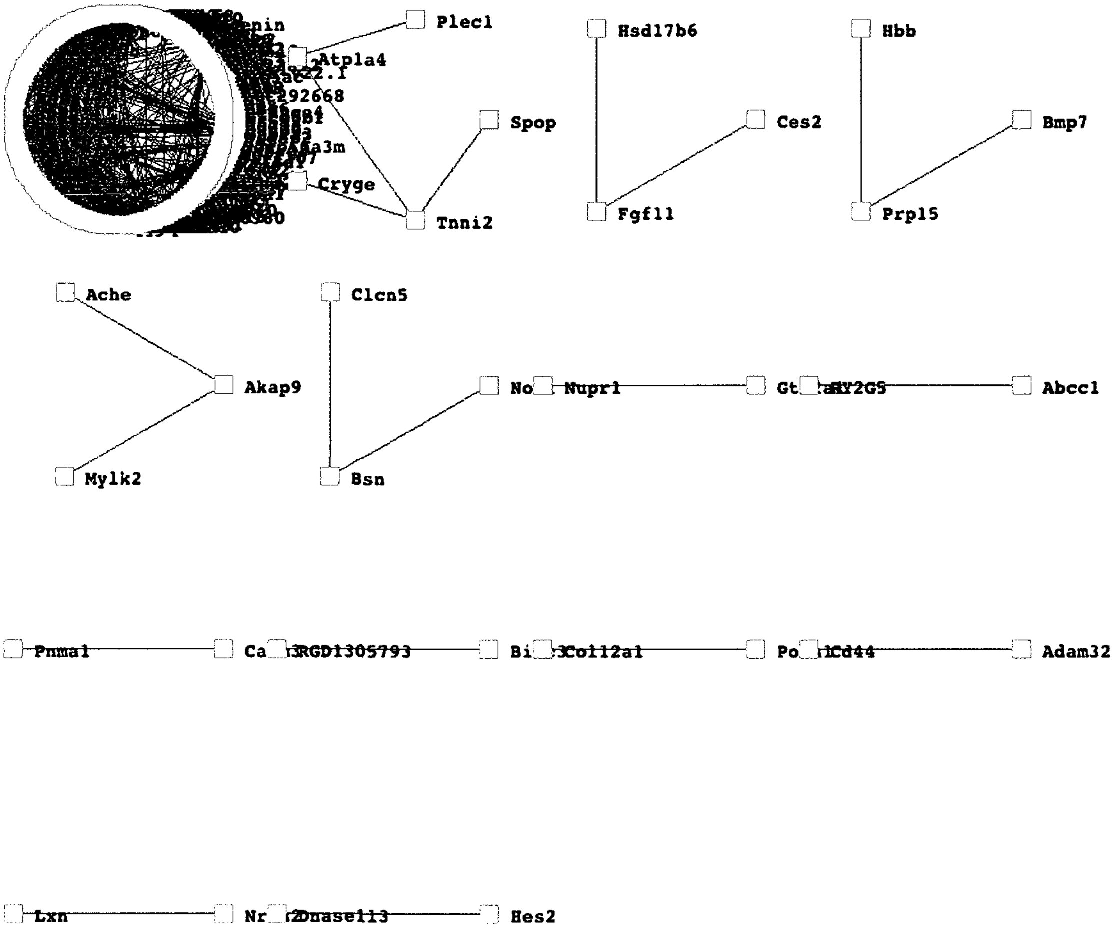


Figure 15a: Total of 192 genes interacting in 14 relevance networks created with mutual threshold of 0.90. Node labels represent gene symbols.

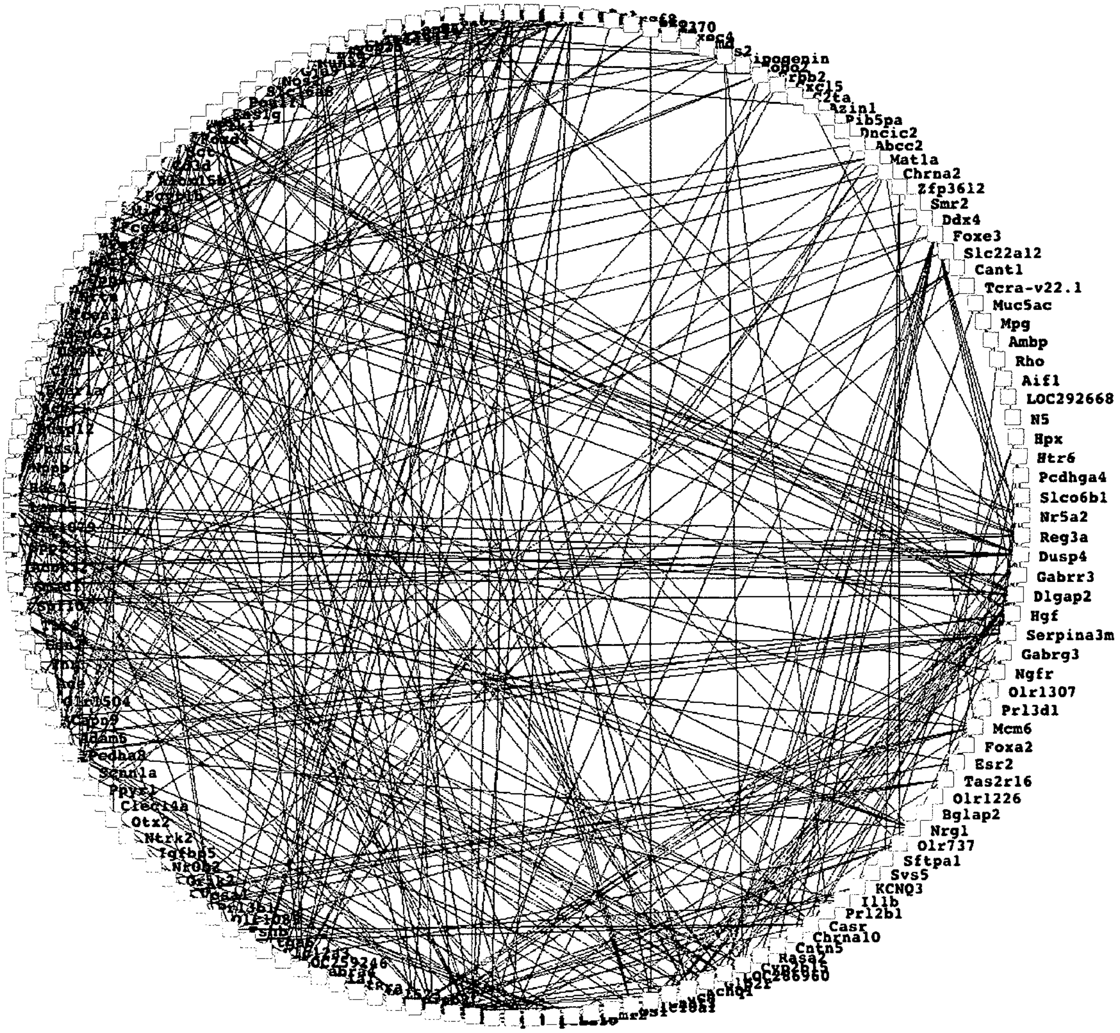


Figure 15b: Subnetwork #1 has 159 genes linked together.

A total of 6400 well-annotated genes from the array were considered for the mutual information network estimation. Top 25% of the gene pairs based on the weight was considered for network modeling. MI matrix was then normalized using x-min(x))/(max(x-min(x))), so that the network's weighted adjacency matrix was between zero and 1, where x is the matrix of the MI data to be normalized. The resulting network was very modular, with the two large modules, made of 1900 genes and 410 genes and many small modules (Figure 16a). Network was color-coded based on the time and found that day 7 and 90 were mostly together in the small modules (Figure 16b). GO molecular function and biological process enrichment analysis suggest their involvement in metal ion binding activity and metabolic process. We also found that one of the two large module was mostly involved in protein binding and negative regulation of biological process (Figure 16c). We also found that these modules were functionally similar.

Many receptors such as acetylcholine receptors, glutamate receptors, GABA receptors, adrenergic receptors, protein tyrosine phosphatase receptors were distributed all over the network, some of them were also acting as hub nodes in the network. Early exposure sample had many glutamate receptors and GABA receptors interacting in the network. B-crystallin (CRYAB) and Voltage gated channel (KCNH7), and few other receptors were playing a central or hub gene role (Figure 16d). Literature evidence suggests the role of these receptors in seizures, neurodegenerative diseases, and brain toxicity.

Graphical Gaussian Model (GGM). The differentially expressed genes across all the time points were selected for the network construction using the graphical Gaussian model. Partial correlation (pcor) was estimated for every gene pair from the 938 DE genes using the package developed by Schafer et al. (2006). For assessing the significance of edges, a two-sided p-values for the test of non-zero correlation, posterior probabilities (1-local fdr), as well as tail based q-values, were computed. A cutoff of 0.1 local fdr was used to determine and extract the significant partial correlation (edges), which resulted in a network with 747 nodes and 15310 edges. Total of 191 unconnected nodes were dropped from the network. Another network was constructed using the 500 most significant edges, which involved 237 genes in the network. Highly interacting genes in the network were insulin like growth factor binding proteins, calcium binding proteins, dopamine receptors, clock homologes, caldindin, and mitigen activated protein kinases. Genes such as prostaglandin (PTGER1), gastrulation homeobox (GBX2), heatshock protein (HSPA1L, HSPB1), neurotransmistter inhibitor (GABRA6), glutamate cystein and ammonia ligase (GCLM and GLUL), clock homologe (CLOCK), and prolactin (PRL) were found as hub genes in the network for the early exposure samples.

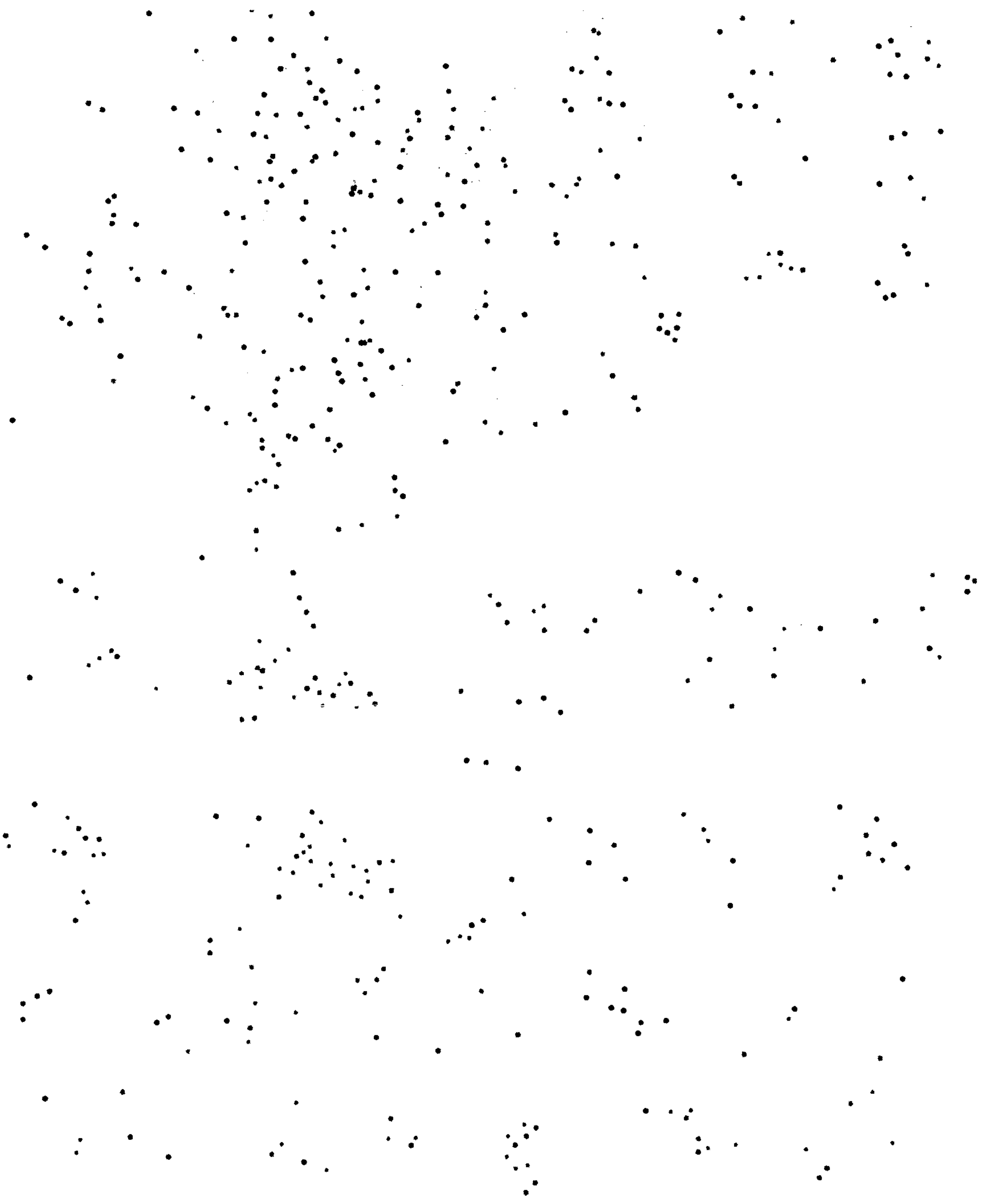


Figure 16a: Modular network constructed using the mutual information theory.

Figure 16b: Genes from each time points are color-coded. Day 7 and 90 are forming many small sub-networks (circled), suggesting the gene activity or expression are similar and distinctly different that day 1 and 2.

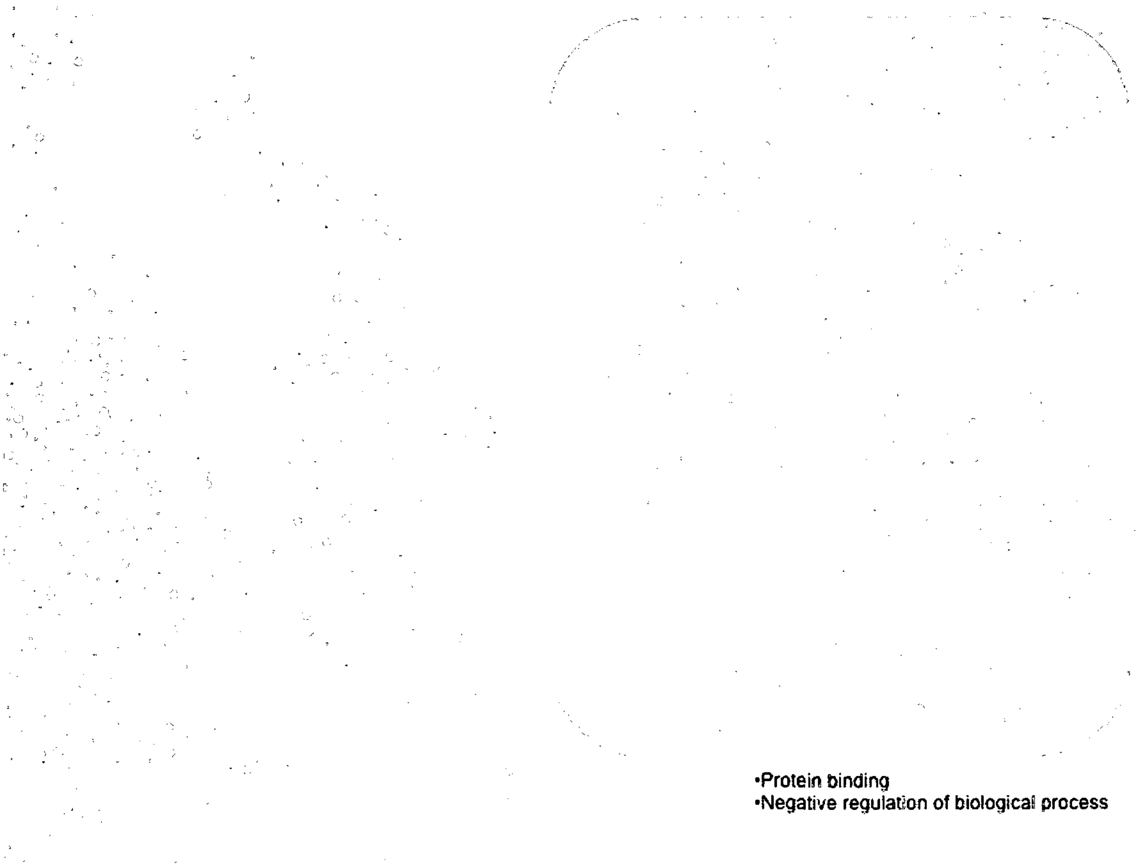


Figure 16c: Genes in the highlighted module have functions such as ribosomal proteins, voltage gated channel, kinase proteins, and G protein coupled receptors.

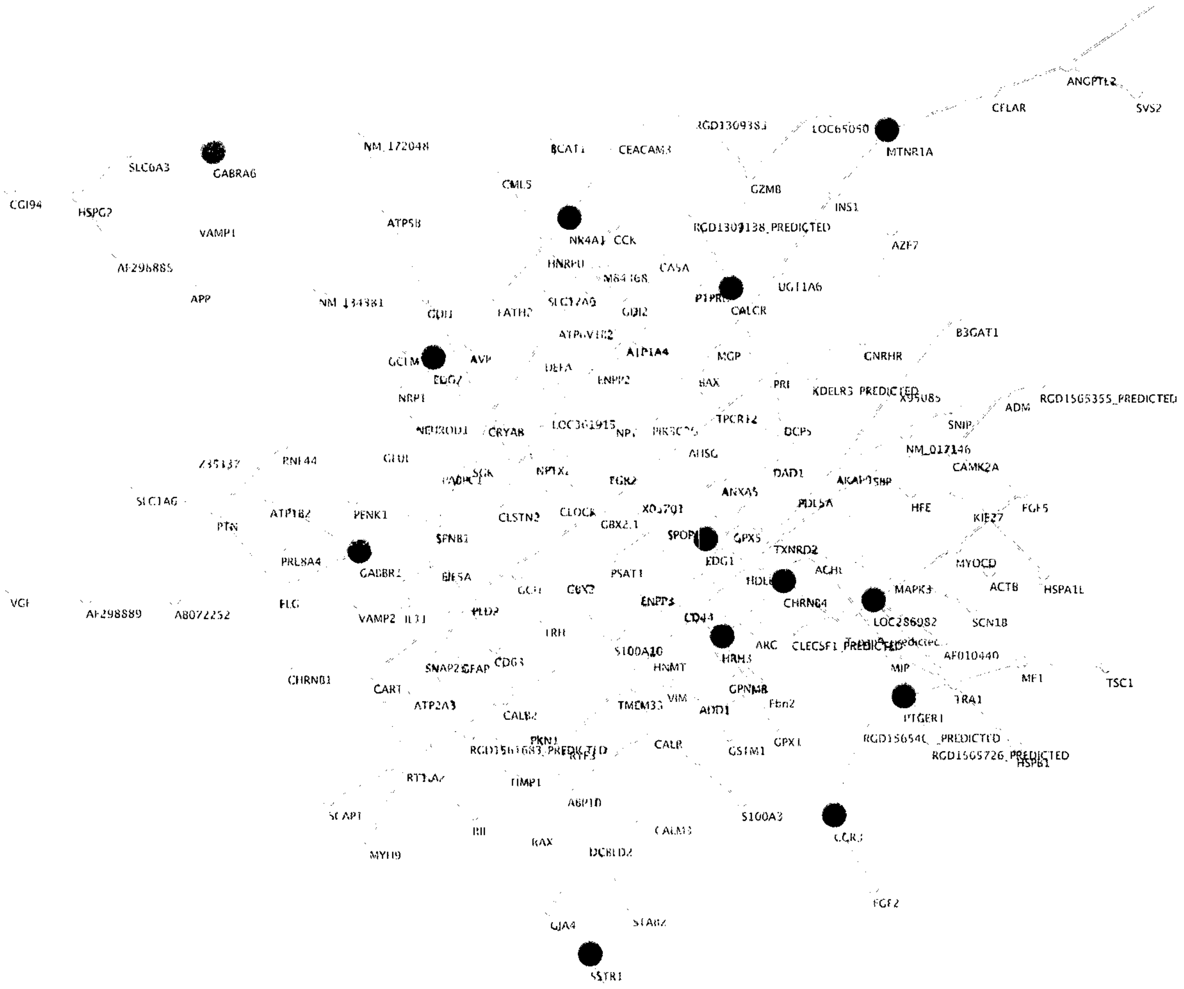


Figure 16 (d) Crystallin, neurotransmitter inhibitors, and glutamate receptors playing central role in the network. GO term surrounding the receptors is the signal transducer activity and the regulation of cyclase activity.

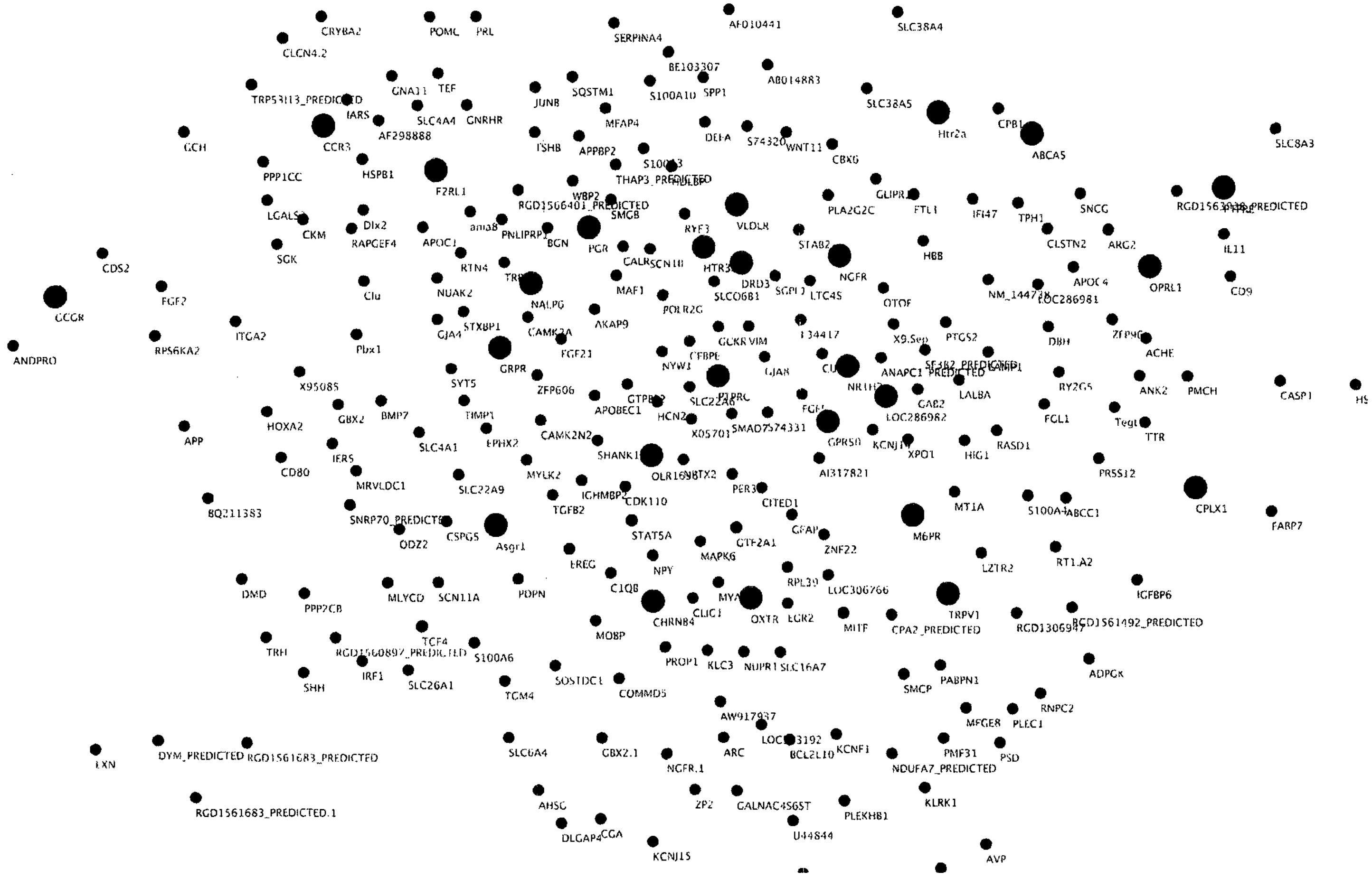


Figure 17a: Receptors (red circles) and transcription factors (blue circles) are playing a central role in the day 2 network.

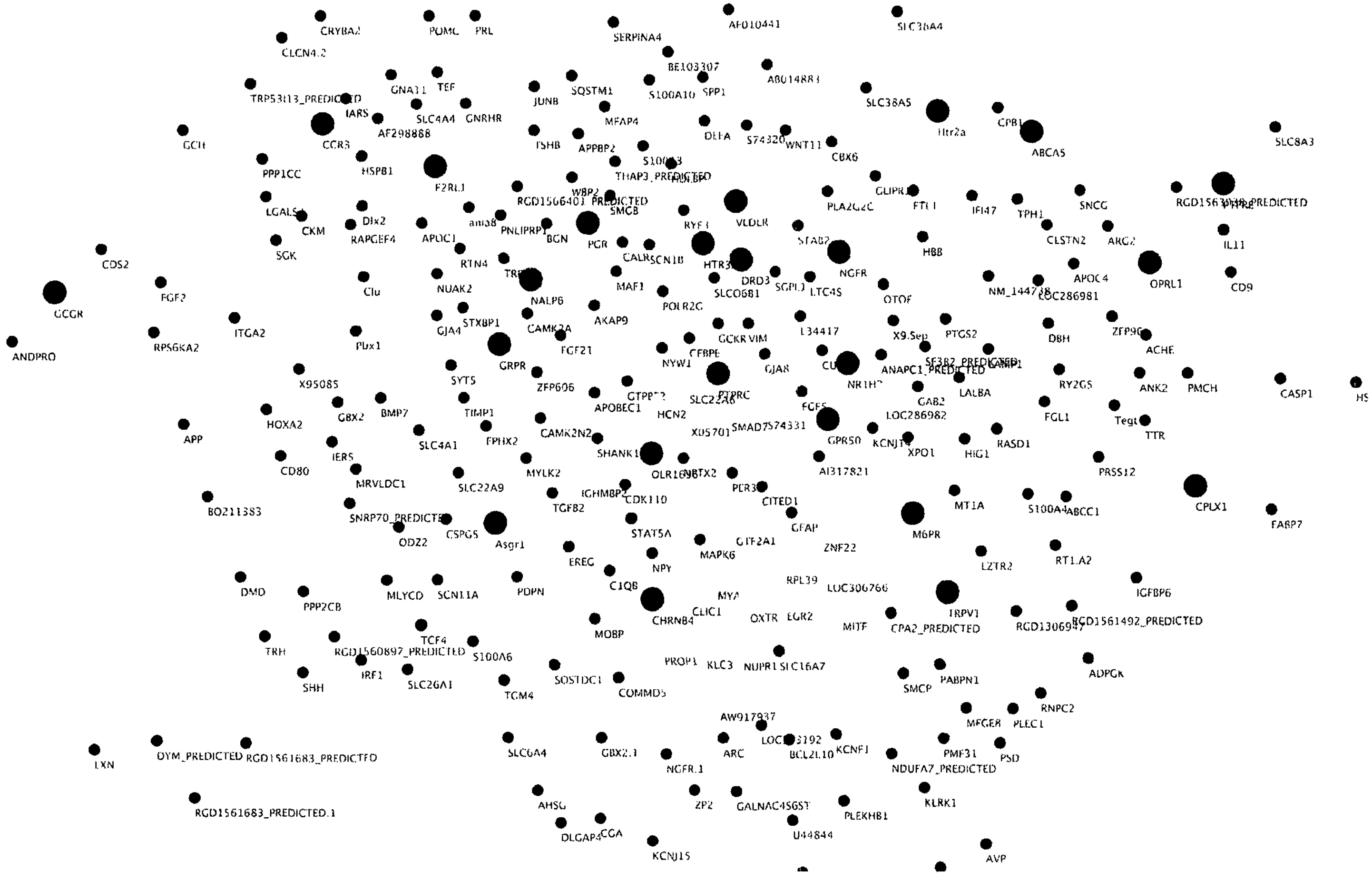


Figure 17b: GO terms such as Transcription regulator activity, protein binding and response to stress was overrepresented in the day 2 network

Many receptors, including glutamate receptors and neurotransmitter inhibitor receptor, were playing a central role in the early time point network model. We found some literature evidence for the interactions involving the two neurotransmitters (Gabra6, Gabbr1) and vesicle-associated membrane proteins (vamp2, Vamp2). Another highly interacting node was a receptor LOC286982, which may play a role in neuroendocrine responses and behavior. Gene in day 2 network had calcium binding proteins, voltage gated channel proteins, serotonin receptor, acetylcholinesterase, and transcription factors playing a central role, but no glutamate or neurotransmitters were active (Figure 17a-b). Genes surrounding the transcription factors and receptors were mainly involved in molecular activities such as transcription regulator activity, protein binding, and response to stress. Literature evidence was found for the Interactions such as SMAD7 and STAT5, Dopamine and Apoptosis regulators, GnRH receptor gene, and thyroid stimulating hormones.

CHAPTER IV

CONCLUSION

Analysis of differential expression may provide new information about the biological pathways involved in a process. This is often done by looking for over-representation of functional classes in gene clusters derived from expression data [21]. Even simple pair-wise comparisons can indicate novel interactions [22]. Using the technique of linking all genes by calculating comprehensive pair-wise mutual information and then isolating clusters of genes, or Relevance Networks, by removing links under a threshold, we were able to find biologically relevant clusters. Although Relevance Networks can be made at any threshold mutual information (TMI), we successfully clustered 192 genes into 14 Relevance Networks at the TMI of 0.90. Decreasing the TMI will introduce more genes and hypothetical associations. Even though some of these associations are noise because some high mutual information may be calculated by chance, the associations at lower TMI may represent novel hypotheses. Increasing the TMI will restrict the Relevance Networks to include only the strongest hypothetical associations.

We have used Mutual Information (MI), Bayesian Network (BN), and GGM models for gene networks (GNs) and tested each model with both artificial and real biological networks. We analyzed the toxicogenomic and demonstrated the usefulness of GNs as a computational approach for the analysis of transcriptional regulation. In summary, a GN can be used, among other things, to i) define transcriptional factors (activators and inhibitors) for a target gene and ii) find co-regulated genes. The intention of the efforts for developing both theories and software for network analysis is that these networks could provide useful clues about biological systems, thus helping with the design and refinement of wet experiments.

The MI model is suitable for a large network, whereas BN and GGM models are suitable for small networks. A learning scheme that scales up to a large number of variables should be investigated, and is a future goal. Nowadays, the finding of an efficient reconstruction method with no constraints in the number of nodes using BN is a cutting-edge problem (Bar-Joseph, 2004).

Comparative toxicogenomics has the potential to identify conserved responses between humans and animal research models that are associated with toxicity, which can be used to develop predictive toxicity tools. In addition, these approaches are likely to provide empirical evidence supporting the transfer of functional annotation from known human and mouse genes to unknown genes or ESTs in the rat or ecologically relevant species like fish, based on sequence similarity and comparable expression patterns. However, platform differences, inaccurate annotation across species and microarrays, the lack of tools to facilitate comparative analysis, one-to-many relationships between genes and probes (e.g., one gene in rat has two or more orthologs in humans), incomplete or poorly annotated genomes, discrepancies between databases which define orthologous relationships (National Center for Biotechnology Information (NCBI) vs. European Bioinformatics Institute (EBI)), and the limited availability of functional annotation complicate effective cross-species comparisons and confound comparative analyses. Current gene ontologies are also imprecise, incomplete, and inconsistent across species, which compromises the accurate interpretation of toxicogenomic data relative to a phenotypic endpoint. Therefore, consistent approaches to annotation curation are required to ensure the accurate interpretation of the data. In addition, despite more complete and accurate annotation for the human and mouse genomes, the rat continues to be the traditional rodent model of choice for toxicology studies.

The interpretation of toxicogenomics data will continue to be a difficult task, and more effective tools to facilitate their integration and interpretation are required. Typically, toxicity is a persistent and easily identified endpoint; however, toxicogenomic responses are dynamic and subject to reversible temporal changes that can be displaced in time relative to toxicity. The added challenge is to accurately determine whether acute or short term toxicogenomic responses are predictive of subchronic or chronic toxicity outcomes. In addition, dose-response studies are required to differentiate adaptive versus toxic responses and to establish toxicogenomic thresholds that need to be exceeded prior to the initiation of the cascade of molecular responses leading to an adverse effect.

Our studies show that individual responses are not independent but form a network of interacting networks. The challenge that remains is to comprehensively integrate the disparate chemical, biological, toxicological, and toxicogenomic data in order to elucidate the mechanisms and networks involved in toxicity and to develop quantitative models capable of accurately predicting thresholds. Complex network theory has been used to investigate technological and social networks, and similar principles have also been shown to govern complex biological networks. Therefore, the most significant challenge will be the application of comparable network approaches that integrate disparate toxicity data in order to reduce uncertainties and to support mechanistically based quantitative risk assessment.

REFERENCES

1. Eckel-Passow JE, Hoering A, Therneau TM, Ghobrial I. Experimental design and analysis of antibody microarrays: applying methods from cDNA arrays. Cancer Res. (2005); 65(8):2985-9.

2. Emmert-Streib, Frank and Dehmer, Matthias. Analysis of Microarray Data: A Network-Based Approach Wiley-VCH (2008), ISBN-10-3-527-31822-4

3. Vinciotti V, Khanin R, D'Alimonte D, Liu X, Cattini N, Hotchkiss G, Bucca G, de Jesus O, Rasaiyaah J, Smith CP, Kellam P, Wit E. Experimental Evaluation of loop versus a reference design for two-channel microarrays. Bioinformatics (2005), 21(4): 492-501.

4. Lee M, Kuo F, Whitmore G, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci (2000), 97:9834-9839.

5. Yang YH, Speed TP. Design issues for cDNA microarray experiments. Nature Reviews Genetics (2002), 3:579-588.

6. Black MA, Doerge RW. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold changes in microarray experiments. Bioinformatics (2002), 18(12):1609-16.

7. Ziv Bar-Joseph. Analyzing time series gene expression data. Bioinformatics (2004), 20(16):2493-2503.

8. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol

Cell. (1998), 9(12):3273-97.

9. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. Gene expression during the life cycle of Drosophila melanogaster. Science(2002), 297(5590):2270-5.

10. Stolovitzky, G, Monroe, D, and Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann N Y Acad Sci. (200, 1115:1-22.

11. Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. BMC Bioinformatics (2007), 8(Suppl 6):S9

12. Bay, S.D., Chrisman,L., Pohorille,A. and Shrager,J. Temporal aggregation bias and inference of causal regulatory networks. In Proceedings of the IJCAI Workshop on Learning Graphical Models for Computational Genomics, Morgan Kaufmann. (2003),

13. Peter M. Haverty, Ulla Hansen, Zhiping Weng. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acid Research (2004), 32(1):179-188.

14. Friedman, N, Linial M, Nachman, I & Pe'er, D. Using Bayesian networks to analyze expression data. J. Comput. Biol. (2000), 7:601-620.

15. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics (2005), 21(1):71-9.

16. Schwarz, G Estimating dimension of a model. Ann Stat (1978), 6:461-464.

17. Chickering, DM. Learning from Data: Artificial Intelligence and Statistics. Springer-

Verlag. New York (1996).

18. Margolin, AA and Califano A. Theory and Limitations of Genetic Network Inference from Microarray Data. Ann N.Y. Acad. Sci. (2007), 1115:51-72

19. Butte, A. J. and Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput (2000), 418-29.

20. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics (2006), 7(Suppl 1):S7.

21. A. de la Fuente, N. Bing, I. Hoeschele and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics (2004), 20(18):3565-74.

22. Whittaker, J. Graphical Models in Applied Multivariate Statistics. Wiley (1990), New York.

23. http://strimmerlab.org/notes/ggm.html

24. Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F. Gene networks inference using dynamic Bayesian networks. Bioinformatics (2003), Suppl 2:ii138-48.

25. Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics (2003), 19(17):2271-2282.

26. Breen MS, Villeneuve DL, Breen M, Ankley GT, Conolly RB. Mechanistic

computational model of ovarian steroidogenesis to predict biochemical responses to endocrine active compounds. Ann. Biomed. Eng. (2007), 35, 970-81.

27. Watanabe, K.H., Jensen, K.M., Orlando, E.F., Ankley, G.T. 2007. What is normal? A characterization of the values and variability in reproductive endpoints of the fathead minnow, *Pimephales promelas*. Comp. Biochem. Physiol. C Toxicol. Pharmacol. (2007), 146, 348-56.
28. Ankley, GT & Villeneuve, DL. The fathead minnow in aquatic toxicology: Past, present and future. Aquatic Toxicology (2006), 78(1): 91-102.
29. Miller, W. Molecular biology of steroid hormone synthesis. Endocr Rev. (1998), 9:295-318.
30. Ankley GT, Johnson RD. Small fish models for identifying and assessing the effects of endocrine-disrupting chemicals. *ILAR J.* (2004), 45:469-83.
31. Stuart, JM, Segal, E, Koller, D, Kim, SK. A Gene Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (2003), 302(5643):249-55.
32. Werhli, AV, Grzegorczyk, M, and Husmeier, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics (2006), 22: 2523-2531.
33. Wille, A. Zimmermann, P., Vranova, E., Furholz, A., LAule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L. *Sparse Graphical Gaussian modelling of the isoprenoid gene network in Arabidopsis thaliana.* Genome Biology (2004), 5:R92.
34. Magwene, P.M. and Kim, J. Estimating genomic co-expression networks using first-order conditional independence. Genome Biology (2004), 5:R100.

35. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. (2003), 4(5):P3.

36. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques (2003), 34(2):374-8.

37. Schäfer J. and Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. (2005), 4: 32.

38. Friedman N, Murphy K, Russell S. Learning the Structure of Dynamic Probabilistic Networks. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (1998), 139-147.

39. Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics (2003), 4:2.

40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Idekar T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research (2003), 13(11):2498-504.

41. Maere S, Heymans K, and Kuiper M. BinGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks.

Bioinformatics (2005), 21(16):3448-3449.

42. E. M. Colin, A. G. Uitterlinden, J. B. J. Meurs, A. P. Bergink, M. Van De Klift, Y. Fang, P. P. Arp, A. Hofman, J. P. T. M. van Leeuwen and H. A. P. Pols. Interaction between Vitamin D Receptor Genotype and Estrogen Receptor {alpha} Genotype Influences Vertebral Fracture Risk. The Journal of Clinical Endocrinology & Metabolism (2003), 88(8):3777-3784.

43. Filby AL, Thorpe KL, Maack G, Tyler CR. Gene expression profiles revealing the mechanisms of anti-androgen- and estrogen-induced feminization in fish. Aquat Toxicol. (2007), 28;81(2):219-31.

44. Christopher J. Martyniuk, Emily R. Gerrie, Jason T. Popesku, Marc Ekker, Vance L. Trudeau. Microarray analysis in the zebrafish (Danio rerio) liver and telencephalon after exposure to low concentration of 17alpha-ethinylestradiol. Aquatic Toxicology (2007), 84:38–49.

45. Venkatadri Kolla, Noreen M. Robertson and Gerald Litwack. Identification of a Mineralocorticoid/Glucocorticoid Response Element in the Human Na/K ATPase α1 Gene Promoter. Biochemical and Biophysical Research Communications (1999), 266(1):5-14.

46. Breen MS, Villeneuve DL, Breen M, Ankley GT, Conolly RB. Mechanistic computational model of ovarian steroidogenesis to predict biochemical responses to endocrine active compounds. Ann. Biomed. Eng. (2007), 35, 970-81.

47. Bak B, Carpio L, Kipp JL, Lamba P, Wang Y, Ge RS, Hardy MP, Mayo KE, Bernard DJ. Activins regulate 17beta-hydroxysteroid dehydrogenase type I transcription in murine gonadotrope cells. J Endocrinol (2009), 201(1):89-104.

48. W. Yang, C. Qiu, N. Biswas, J. Jin, S. C. Watkins, R. C. Montelaro, C. B. Coyne and T. Wang. Correlation of the Tight Junction-like Distribution of Claudin-1 to the Cellular Tropism of Hepatitis C Virus. J Biol Chem. (2008), 283(13): 8643-53.

49. N. Ben-Jonathan, S. Chen, J. A. Dunckley, C. LaPensee and S. Kansra. Estrogen receptor-alpha mediates the epidermal growth factor-stimulated prolactin expression and release in lactotrophs. Endocrinology (2009), 150(2):795-802.

50. Y. Oshima, K. Noguchi and M. Nakamura. Expression of Lhx9 isoforms in the developing gonads of *Rana rugosa.* Zoolog Sci (2007), 24(8): 798-802.

51. Jone MR, Wilson SG, Mullin BH, Watts GF, and Stuckey BGA. Polymorphism of the follistatin gene in polyscystic ovary syndrome. Molecular Human Reproduction. (2007), 13(4):237-241.

52. Fitzpatrick SL, Richards JS. Identification of a cyclic adenosine 3',5'-monophosphate-response element in the rat aromatase promoter that is required for transcriptional activation in rat granulosa cells and R2C Leydig cells. Mol Endocrinol (1994), 8:1309–1319.

53. Minnie Hsieh, Sabine M. Mulders, Robert R. Friis, Arun Dharmarajan and JoAnne S. Richards . Expression and Localization of Secreted Frizzled-Related Protein-4 in the Rodent Ovary: Evidence for Selective Up-Regulation in Luteinized Granulosa Cells. Endocrinology (2003), 144(10):4597-4606.

54. Craig TA, Sommer S, Sussman CR, Grande JP, Kumar R. Expression and regulation of the vitamin D receptor in the zebrafish, Danio rerio. J Bone Miner Res. (2008), 23(9):1486-96.

55. Ankley G, Kahl MD, Jensen KM, Hornung MW, Korte JJ, Makynen EA, Leino RL.

Evaluation of the Aromatase Inhibitor Fadrozole in a short-term Reproduction Assay with the Fathead Minnow (Pimephales promelas). Toxicological Sciences (2002), 67:121-130.

56. Thomas F. Jenkins a,*, Alan D. Hewitt a, Clarence L. Grant c, Sonia Thiboutot b, Guy Ampleman b, Marianne E. Walsh a, Thomas A. Ranney d, Charles A. Ramsey e, Antonio J. Palazzo a, Judith C. Pennington. Identity and distribution of residues of energetic compounds at army live-fire training ranges. Chemosphere 63 (2006), 1280–1290.

57. Beller HR, Tiemeier K. Use of liquid chromatography/tandem mass spectrometry to detect distinctive indicators of in situ RDX transformation in contaminated groundwater. Environ Sci Technol (2002), 36:2060–2066.

58. Burdette, L.J., Cook, LL, and Dyer, RS. Convulsant properties of cyclotrimethylenetrinitramine (RDX): Spontaneous audiogenic, and amygdaloid kidnled seizure activity. Toxicol. Appl. Pharmacol. (1998), 92. 436-444.

59. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature (1999), 402:83-86.

60. Michael Hecker. Gene Regulatory Network Reconstruction: Best Practice Guide. Jena GmbH.

61. Momiao Xionga, Jun Lia, and Xiangzhong Fang. Identification of Genetic Networks. Genetics (2004), 166:1037-1052.

62. Edward J Perkins, Wenjun Bao,Xin Guan, Choo-Yaw Ang, Russell D Wolfinger, Tzu-Ming Chu, Sharon A Meyer, and Laura S Inouye. Comparison of

transcriptional responses in liver tissue and primary hepatocyte cell cultures after exposure to hexahydro-1,3,5-trinitro-1,3,5-triazine. BMC Bioinformatics (2006), (suppl 4): 522.

63. Trond Hellem B, Bjarte Dysvik, and Inge Jonassen. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research (2004), 32:3.

64. Jeffrey P Townsend, and Daniel L Hartl. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biology (2002), 3.

65. Bartek Wilczynski and Norbert Dojer. BNFinder: exact and efficient method for learning Bayesian networks. Bioinformatics (2009), 25(2):286-287.

66. Dojer N. Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinformatics (2006), 7:249.

67. Zhenjun Hu, Joseph Mellor, Jie Wu, and Charles DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics (2004), 5:17.

68. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res (2003), 31:248-250.

69. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res (2002), 30:42-46.

70. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C. Predictome: a database of putative functional links between proteins. Nucleic Acids Res (2002), 30:306-309.

71. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a

database of predicted functional associations between proteins. Nucleic Acids Res (2003), 31:258-261.

72. Ashburner, M., C. A. Ball. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet (2000), 25(1): 25-9.

73. Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. Bioinformatics (2007), 23(22):3024-3031.

74. Johnson, WE, Rabinovic, A, and Li, C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics (2007), 8(1):118-127.

75. Meyer, P E, Lafitte , F, and and Bontempi, G. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics (2008), 9: 461.