Cellular Pathway Networking Using Gene Expression Profiles and Transcription Factor Binding Sites

Melanie J. Smith[1] and Brad Windle[2]

[1] Biological Sciences Department, Cedar Crest College, 100 College Drive, Allentown PA 18104
[2] Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond VA 23298

Abstract

In the midst of the post-genomic era, the wish to engineer pharmaceuticals, vaccines, and organisms is hindered by the need to define the cell, tissue, and the organism. The way that life organizes itself is a necessity for altering the network of life-events. Now that numerous genomes from a variety of organisms have been sequenced, there is a need to define the function of all the genes and genetic elements. One approach that this paper will illustrate is the use of microarray data from different cell types to define the function of a gene, based on gene expression profile-comparisons to previously defined genes. In order to accomplish this, Random Forest and Neural Network classification models were used to predict the functions of genes. Data from these models were used to relate cell functions and construct a network of pathways. In order to understand how these pathways interact and are regulated by the cell, the upstream transcription factor binding sites (TFBS) were studied. Using Random Forest, genes were classified into defined Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways based on the TFBS. Using the model output, the important TFBS that define a pathway were found.

Introduction

The two major questions that were asked during the course of this research project were addressing the function of a gene as well as its regulation. These questions were addressed with respect to biochemical pathways and the network of how these pathways are related to one another.

Initially, the goal of gene identification was for the purpose of understanding what genes were targeted by pharmaceutical drugs and why. Other reasons for gene identification are for characterizing disease and developing drug therapy. A great deal of research is conducted with the goal of creating gene prediction models, but the reliability and reproducibility of the models are still at the beginning stages of a success (Friedberg 2006). Even still, models to identify genes within the DNA sequences are still in development (Mathé *et al* 2002).

Gene expression data has been used to predict characteristics of biological phenomena in many fields of research (Hughes *et al* 2000). The NCI 60 cell line gene expression profiles have been used to identify types of cancer (Ross *et al* 2000). Previous work using the model in this paper has been published that assesses the performance of the model as a predictor (Ko *et al* 2005). Applications of similar projects are also studied by others for purposes of personalized cancer treatment (Nevins 2003).

The gene prediction model that was created uses a set of previously defined genes to compare unknown genes in order to identify their function. The training genes were defined by the KEGG into their respective pathways. The model classified unknown genes into their most probable pathways and also indicated possible secondary, tertiary, and quaternary pathways that the gene could be affiliated with. Using this data, a network of pathways was constructed.

It is assumed that these pathways were related by the model based on the fact that genes from different pathways are co-expressed. It is well accepted among the scientific community that co-expressed genes are a part of a well-organized network. Based on this assumption, the transcriptional regulatory properties of the genes were then studied to find if a correlation between the regulation and expression of genes can be predicted by computer modeling.

Methods

Gene Expression Profile Model (GEP)

Neural Network and Random Forest models were used to discover what cell functions are related based on gene expression profiles of NCI's 60 cell lines. The tumor cell lines are from lung, colon, breast, prostate, kidney, melanoma, leukemia, central nervous system, and ovary cancers (Developmental Therapeutics Program NCI / NIH). The models were created to classify genes based on their function. In order for the models to correctly classify a gene, a large number of references was needed. For the 6,165 genes analyzed, 367 of the genes were used as a reference. These genes are well known and their functions were determined by the KEGG. Each gene's expression profile was compared to the reference set, and classified into the pathway that best fit. The 367 training genes only represented 21 pathways of the 300^+ pathways from KEGG, so the models could only classify genes into these 21 pathways. These 21 pathways are listed in Table 1. The 21 functions of the cell were chosen from metabolic, genetic information, environmental information, and cellular processes that are well known. Also, a human disease, Huntington's Chorea, was included among the cell functions.

Table 1. The 21 Pathways from the KEGG that were used to classify genes in the gene expression profile model and the TFBS model. Pathways fall into five categories including metabolism, genetic information processing, cellular processes, environmental information processing, and human diseases.

- **Metabolism**
  - Arginine and Proline
  - N-glycans
  - Glycolysis
  - Oxidative Phosphorylation
  - TCA Cycle
  - Pyrimidine
  - Glutathione
  - Valine, Leucine, Isoleucine
  - Porphyrin
  - Purine
  - Glycerolipids
- **Genetic Information Processing**
  - Ribosome
  - Polymerases
  - tRNA
  - Proteasome
- **Cellular Processes**
  - Cell Cycle
  - Apoptosis
  - Cell Adhesion
- **Environmental Information Processing**
  - MAP Kinase
  - PI Kinase
- **Human Diseases**
  - Huntington's Disease

The neural network model determines how probable it is that a gene is in a pathway and assigns a relative probability for all 21 pathways. For the Random Forest Model, a gene is classified into one pathway, but each time the model is run, it may classify the gene differently. The pathway the gene is classified into most of the time is most likely to be the pathway it belongs to. So, the model was run 500 times, the number of times the gene was classified into each pathway was recorded, and the "votes" were converted to relative probabilities (votes/500). This process was repeated 100 times to get a more representative answer. All of the probabilities from both models were analyzed using Microsoft Excel and can be obtained from supplementary material.

In order to decide what pathways were linked, programs were written in Microsoft Visual Basic. The first program counted how many genes were in a primary pathway and how many of those genes went to the same secondary, tertiary, or quaternary pathway. The second program calculated the binomial distribution to determine how probable it is that the two pathways are related. The number of successes was how many genes from the primary pathway were in the secondary pathway. The number of trials was how many genes in total were classified to the secondary pathway. The probability of success was determined by dividing the number of genes in the primary pathway by the total number of genes in the model. These values were used as p-values and were converted to –log (p-values) to be more easily read. The p-values for both models can be found in supplementary materials.

Using the pathway association scores, a network of how the different pathways are connected was constructed for both models and for primary versus secondary, tertiary, and quaternary functions. These diagrams were created in Microsoft Excel and can be found in supplementary materials.

Transcription Factor Binding Site Model (TFBS)

The second gene prediction method involved transcription factor binding site data (TFBS) for the training genes. 288 genes defined using the same 21 KEGG pathways as the GEP model were used to train a Random Forest Model. The Telis Database website was used to find TFBS that were located 300bp or less upstream of the training genes (Cole 2004). A spreadsheet file was created indicating the TFBS and number of copies of the site that each gene possessed (Supplementary Material). Using Microsoft Visual Basic, code was written to operate the Random Forest Model using the R plug-in for Microsoft Excel. Random Forest was created by Leo Breiman, Liaw and Wiener developed the program to run in R, and Thomas Baier and Erich Neuwirth developed the R-Excel add-in that controls R (Breiman 2001). Twenty variables (TFBS) were used at each branching point of the decision tree and 5,000 trees were compiled. To verify the model, just the training genes are used so that all but one gene are used as a reference, and the one gene is classified. Three output files are given (Supplementary Materials). A confusion matrix shows how many genes were classified to each pathway, and what pathway they should have been classified in. A second file shows for each gene, the fraction of 5,000 trees that classified the gene to each pathway. The highest value was used as the classification. The percent error, and purity was found for each pathway. The third file shows importance values for each element with respect to each pathway. The importance value is based on how well the pathway performed without the element present. The higher the value, the more important the element was for classifying the genes correctly for that pathway.

To find what element was most important for each pathway, two approaches were taken. First, the pathway that received the highest score for an element was best correlated with that element. Second, elements with the highest importance value for a pathway were considered most relevant to that pathway. Threshold values for the importance value were taken for each individual pathway, based on how many training genes of that pathway contained that TFBS. After the TFBS were identified for each pathway, the training genes that were correctly classified and contained the important TFBS were identified. The TFBS that contained the correctly identified genes were analyzed through literature searches (Supplementary Materials).

Results

Gene Expression Profile Model

The gene expression profile model produced a complex set of cellular networks based on the pathway associations. All of these networks can be found in supplementary material; however, figure 1 shows an example of the type of networks that were created. Three major pathways, cell cycle, ribosome synthesis, and PI kinase signaling, possessed the greatest number of connections. Table 2 shows the different cellular functions that were often associated with the three major pathways.

Figure 1. Cellular function networks were created from pairwise comparisons of the primary, secondary, tertiary, and quaternary pathways. The figure shows a comparison of the primary and secondary pathways from the Neural Network model.

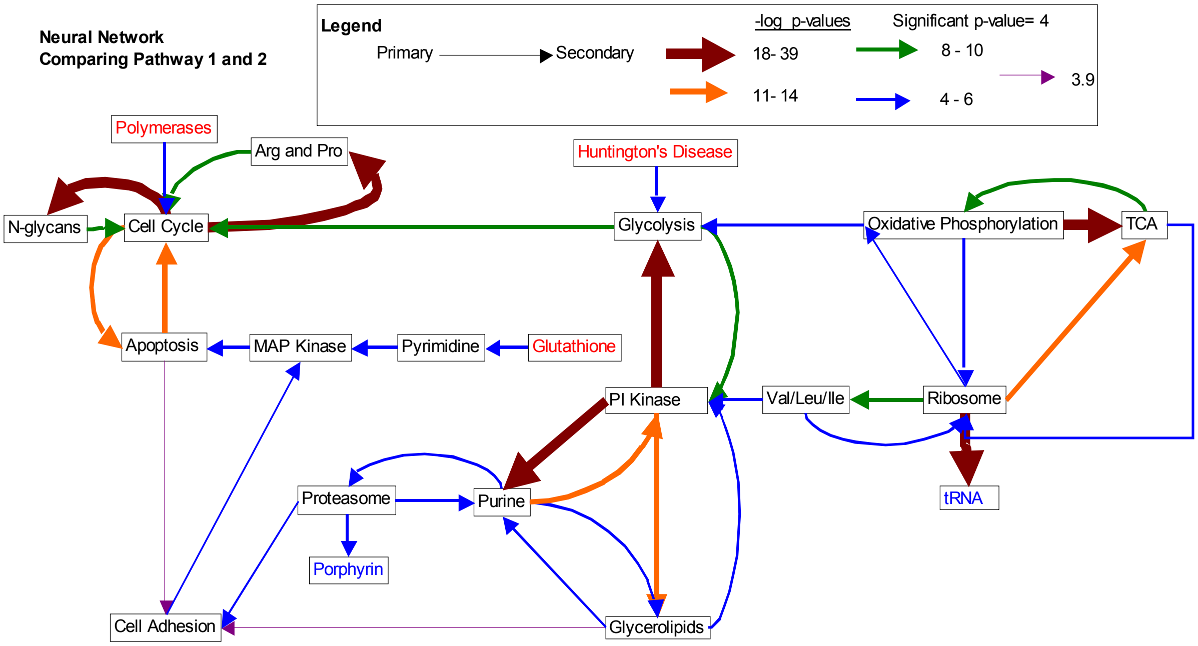


Table 2. The gene expression profile model associated many functions to three major pathways: cell cycle, ribosome synthesis, and PI kinase signaling pathways. The associated pathways covered a broad set of functions including metabolism, cellular processes, genetic information processes, environmental processing, and human diseases.

	Primary Pathway		
Functional Category	**Cell Cycle**	**Ribosome**	**PI Kinase**
Metabolism	Arginine and Proline N-glycans	Glycolysis Oxidative Phosphorylation TCA Cycle	Glycerolipids Glycolysis Purine Valine, Leucine, and Isoleucine
Cellular Processes	Apoptosis		Cell Cycle
Genetic Information Processes	Polymerases Proteasome	tRNA	
Environmental Processing	PI Kinase		
Human Diseases	Huntington's Disease		Huntington's Disease

Transcription Factor Binding Sites Model

Of the 21 cellular pathways, six of these pathways contained correctly classified genes from verification analysis of training genes. The percent error and purity for these pathways are shown in Table 3. The number of correctly and incorrectly classified genes for each pathway were found as well as the total number of genes that were classified to a pathway (Table 3). TFBS that had high importance values for the six pathways are listed in Table 4, and Table 5 and 6 describes the basic functions of the transcription factors, as defined by the Gene Ontology. Literature searches were performed to investigate the shared transcription factors among different pathways. Figure 2 shows a network of functions compiled based on previous research that explains associations made in the GEP and TFBS models.

Table 3. The TFBS model correctly classified six pathways. Percent error and purity were calculated for the pathway. Performance of the individual pathways was also analyzed by the number of genes that were correctly and incorrectly classified for each defined pathway, as well as the total number of genes that the TFBS classified to that pathway.

	Cell Cycle	**MAP Kinase Signaling**	**Oxidative Phosphorylation**	**PI Kinase Signaling**	**Purine**	**Ribosome Synthesis**
Error (%)	74	89	95	54	82	42
Purity (%)	14	50	6	22	14	18
Correctly Classified Genes	9	1	1	12	3	21
Incorrectly Classified Genes	25	8	21	14	14	14
Total Classified in Pathway	65	2	17	55	21	117

Table 4. Using the importance values, the important factors for the cell cycle, MAP kinase signaling, oxidative phosphorylation, PI kinase signaling, purine synthesis, and ribosome synthesis pathways were determined. Transcription factors in bold are shared among other pathways.

	Cell Cycle	MAP Kinase Signaling	Oxidative Phosphorylation	PI Kinase Signaling	Purine Synthesis	Ribosome Synthesis
Transcription Factors	Ahr	AP2	**AP2alpha**	**AP1**	Egr3	**AP1**
	ARNT	**AP2alpha**	deltaEF1	**AP2alpha**	MyoD	**Dof**
	Elk	c_ETS	GATA2	**Dof**	Nkx2.5	**Elk**
	c-Myc			c-Myb		FREAC3
	Max			NF1		GATA3
	NFY					MZF
	PBX1					SAP
						SP1
						TBP

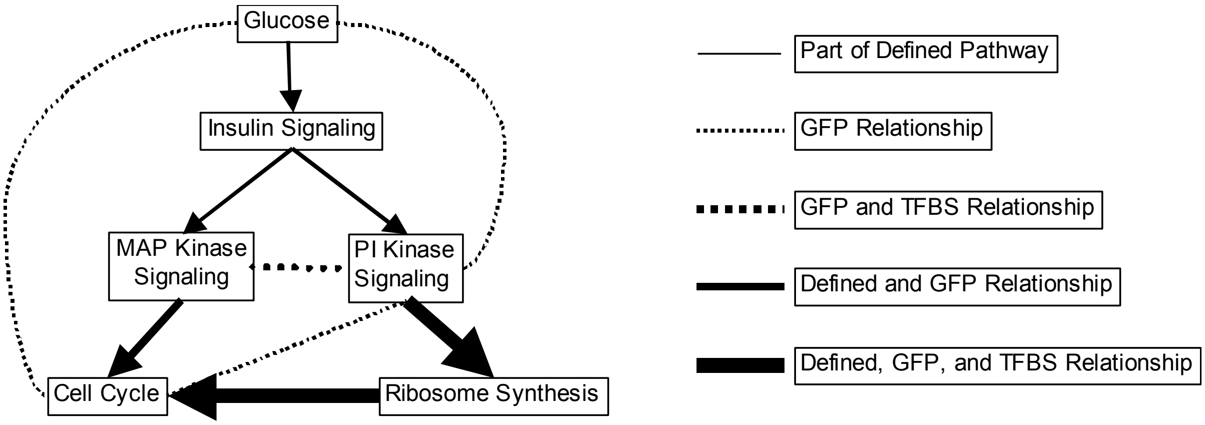
Table 5. Using the Gene Ontology definitions of the transcription factors associated with each pathway, cellular functions were correlated with the cell cycle, ribosome synthesis, and PI kinase signaling pathways.

	Cellular Pathways		
Functional Categories	**Cell Cycle**	**Ribosome Synthesis**	**PI Kinase Signaling**
Cellular Processes	Cell Cycle Apoptosis	Cell Cycle Apoptosis	Cell Cycle
Developmental Processes	Embryonic Development Hindbrain Development Sex Differentiation	Nervous System Development Plant Growth Development Embryonic Development Circulatory and Heart development Multiple Organ Development	Ectoderm Development Odontogenesis
Differentiation/ Proliferation	Differentiation Proliferation	Leading Edge Cell Differentiation Proliferation	Cell Differentiation
Environmental Processing	Response to Stress Response to Radiation	Inflammatory Response	Signal Transduction

Table 6. Using the Gene Ontology definitions of the transcription factors associated with each pathway, cellular functions were correlated with the purine synthesis, MAP kinase signaling, and oxidative phosphorylation pathways.

	Cellular Pathways		
Functional Categories	**Purine Synthesis**	**Map Kinase Signaling**	**Oxidative Phosphorylation**
Developmental Processes	Peripheral Nervous system development Muscle Development	Ectoderm Development	CNS development Pituitary gland development Ectoderm Development
Differentiation/ Proliferation	Cell Differentiation	Positive regulation of erythrocyte differentiation Negative regulation of cell proliferation	Cell Proliferation Cell Fate determination Embryonic morpheogenesis Cell Maturation Neuron differentiation
Environmental Processing	Protein amino acid phosphorylation Neuromuscular synaptic transmission Circadian Rhythm	Immune Response Signal Transduction	Immune Response Phagocytosis Signal Transduction

Figure 2. Based on literature searches, the insulin signaling network including glucose, insulin, MAP kinase, PI kinase, ribosome synthesis, and cell cycle was constructed. Solid lines were defined by literature, and several were correlated by the models. Dotted lines were related by the models, but are not a defined relationship in the insulin pathway.



Discussion

Gene Expression Profile Model

The networks of pathways created are based on the assumption that co-expressed genes are related in function. That is, there must be some cooperative function between pathway A and B if their expression is occurring at the same time. The validity of the networks created are based on the number of occurrences of the relationship and literature searching that was performed. The number of occurrences was considered via p-values as well as the number of correlations between primary, secondary, tertiary, and quaternary pairwise comparisons. Obvious correlations that are well established were found, such as relationships found between the ribosome synthesis and tRNA synthesis as well as glycolysis, TCA, and oxidative phosphorylation (Figure 1). Table 2 shows major connections that were found in the two models.

Transcription Factor Binding Site Model

The TFBS model has a very high error, and does not seem to be able to classify the genes with any sort of accuracy. There are enough reasons as to why the model may not be performing as to make it difficult to say for certain what the problem is. It is already well-known that transcription factors enhance transcription for a large assortment of genes. But, it is also proposed that modules or clusters of TBFS cooperate together to elicit transcription of a particular array of genes. Initially, it was hoped that the model would find unique factors for a pathway or that the number of copies of a TFBS would be unique to a pathway. However, neither of these cases were found. In fact, replacing all non-zero numbers with one so that genes either contained or did not contain a TFBS gave nearly identical results as the original data set. In order to improve this model, defining clusters of TFBS for genes may be a solution. Also, the TFBS for genes were only computationally found, and are not biologically proven to be used *in vivo*. This may have introduced error into the system, which may be considerably significant. For instance, TFBS that were found significant for the ribosome synthesis and PI kinase signaling pathway were four factors from the Dof family of transcription factors that are only found in plants. This may have been biologically interesting if it weren't for the fact that the genes are human. Further investigation of this anomaly was not pursued, although the Dof TFBS may be an artifact of evolution, and may very well have significance in the plant regulatory system.

Another important consideration for the model is DNA and factor modifications that occur in the cellular environment. Methylation, phosphorylation, acetylation, etc. all play an important role in the activity of molecules. For instance, phosphorylation of the cAMP response element binding protein (CREB) largely determines what other factors CREB will bind to, and what genes will be transcribed.

Also, another type of modeling tool could be used to interpret the data. For instance, Neural Network modeling could be used.

Despite the shortcomings of the model, some significant data does seem to have come from the results. The functional identity of the TFBS for the six pathways was found using gene ontology (Table 5 and 6). Then, the pathways that had TFBS in common were searched in the literature for regulatory connections. Interestingly, many of the factors dealt with differentiation, proliferation, and developmental processes and were also often found to be tissue-specific. What was found was that all pathways play a significant role in the control of the cell cycle. All pathways were found to be included in the insulin pathway (Figure 2). Two basic principles that govern the progression of the cell cycle are the environmental state and the health of the cell. The environment of the cell can be determined by the specific tissue, such as muscle, that the cell is a part of. Depending on the location of the cell, the cell will divide, differentiate, or stay the same. Once the cell decides what to do, it requires energy and supplies to perform the task. So, once MAP kinase gives the signal to start the cell cycle, the requires ribosomal synthesis to occur. And, the insulin pathway controls ribosomal synthesis to ensure that enough energy is present to make all the proteins necessary for DNA replication and cellular division.

Both the gene expression profile and the transcription factor binding site model explain the insulin pathway. The GEP model shows glycolysis to be co-expressed with the cell cycle and PI kinase signaling. This connection makes sense since G1 and G2 phase of the cell cycle require a great deal of energy for the high rate of protein synthesis. Also, PI kinase plays a role in the regulation of cell cycle progression (Kenney *et al* 2004). In fact, both kinase-signaling pathways were found to be linked to the cell cycle in the GEP model. Although all of the pathways were found to be co-expressed, they are not all co-regulated. If glucose and the other pathways were regulated in the same way, they would always be expressed at once, and that would defeat the purpose of glucose levels regulating the function of the other pathways. The same argument explains the lack of regulatory connection between MAP kinase and the cell cycle. However, a regulatory connection is found between the ribosome synthesis and PI kinase signaling and the ribosome synthesis and the cell cycle. The co-regulation of the ribosome synthesis pathway and the cell cycle is important to ensure efficient production of proteins during G1 and 2 (Thomas 2000). Since PI kinase signaling is needed to activate synthesis of the ribosome, the co-regulation of these pathways is also required (Roquest and Vidal 1999). It is worth mentioning that PI kinase and the cell cycle did not have any regulatory connections. This is likely important for the same reason as MAP kinase signaling.

Although these findings are interesting they are largely incomplete, since it is only 6 of the 21 pathways that were studied. Although it was easy to find descriptions of transcription factors involved in cell cycle progress, factors important in metabolic and other processes were not found. Although some research has been done on metabolic regulation, much more work is necessary for a complete picture (Desvergne *et al* 2006). In contrast, more detailed work has been done to depict the embryonic development, tissue-specific regulation, and organ-specific regulation (Davidson *et* al 2002, Smith *et al* 2006, Olson 2006). Recently, studies have begun to elucidate body plan changes and speciation based on regulatory evolution (Prud'homme *et al* 2007). An ideal model of gene networks would be the integration of all of these levels of control, from the individual cell to the species level, and perhaps beyond.

Supplementary Material

Smith, M. 2007. Bioinformatics and Bioengineering Summer Institute. Virginia Commonwealth University. http://ramsites.net/~msmith37/.

- Gene Expression Profile Model Data
    - Pathway Networks
    - Pathway Probablities
    - Pathway Association Scores
- Transcription Factor Binding Site Model Data
    - Gene IDs and TFBS Data
    - Pathway Probabilities
    - Confusion Matrix
    - TFBS Importance Values
    - TFBS Gene Ontology

References

Breiman, L. 2001. Random Forests. Machine Learning. 45(1): 5-32.

Breiman, L. and A. Cutler. Random Forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (8 June 2006).

Cole, S. and the UC Regents. 2004. The transcription Element Listening System. http://www.telis.ucla.edu/index.htm.

Collins, J. Developmental Therapeutics Program NCI / NIH. (2006). http://dtp.nci.nih.gov./index.html. (4 Aug 2006).

Davidson, E., J. Rast, P. Oliveri, A. Ransick, C. Calestani, C. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. Brown, C. Livi, P. Lee, R. Revilla, A. Rust, Z. Pan, M. Schilstra, P. Clarke, M. Arnone, L. Rowen, R. Cameron, D. McClay, L. Hood, and H. Bolouri. 2002. A genomic regulatory netowork for development. Science. 295: 1669-1678.

Desvergne, B., L. Michalik, and W. Wahli. 2006. Transcriptional regulation of metabolism. American Physiological Society. 86: 465-514.

Friedberg, I. 2006. Automated protein function prediction—the genomic challenge. Briefings in bioinformatics. 7(3): 225-242.

Gene Ontology. (2006). http://www.geneontology.org/index.shtml (8 June 2006).

Hughes, T., M. Marton, A. Jones, C. Roberts, R. Stoughton, C. Armour, H. Bennett, E. Coffey, H. Dai, Y. He, M. Kidd, A. King, M. Meyer, D. Slade, P. Lum, S. Stepaniants, D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. Friend. 2000. Functional discovery via a compendium of expression profiles. Cell. 102: 109-126.

Kenney A., H. Widlund, and D. Rowitch. 2004. Hedgehog and PI-3 kinase signaling converge on Nmyc1 to promote cell cycle progression in cerebellar neuronal precursors. Development. 131: 217-228.

Ko, D., W. Xu, and B. Windle. 2005. Gene function classification using NCI-60 cell line gene expression profiles. Computational biology and chemistry. 29:412-419.

Kyoto encyclopedia of genes and genomes. (2006). http://www.genome.ad.jp/kegg/. (8 June 2006).

Mathé, C., M. Sagot, T. Scheix, and P. Rouze. 2002. Survey and Summary: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 30(19): 4103-4117.

Nevins, J. 2003. Reviews: Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Human Molecular Genetics. 12: 153-157.

Olson, E. 2006. Review: Gene regulatory networks in the evoluton and development of the heart. Science. 313: 1922-1927.

Prud'homme, B., N. Gompel, and S. Carroll. 2007. Emerging principles of regulatory evolution. The national academy of sciences. 104: 8605-8612.

Roquest, M. and H. Vidal. 1999. A phosphatidylinositol 3-kinase/p70 ribosomal S6 protein kinase pathway is required for the regulation by insulin of the p85α regulatory subunit of phosphatidylinositol 3-kinase gene expression in human muscle cells. The journal of biological chemistry. 274(48): 34005-34010.

Ross, D., U. Scherf, M. Eisen, C. Perou, C. Rees, P. Spellman, V. Iyer, S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Bolstein, and P. Brown. 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics. 2000 March, 24 (3): 225-234.

Smith, A., P. Sumazin, Z. Xuan, and M. Zhang. 2006. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. The nathional academy of sciences. 103(16): 6275-6280.

Thomas, G. 2000. Commentary: An encore for ribosome biogenesis in the control of cell proliferation. Nature cell biology. 2(5):71-72.