Mutual Information Relevance Networks: Functional Genomic Networks Built From Pair-wise Entropy Measurements

by

Atul Janardhan Butte
M. D., Brown University School of Medicine, 1995

Submitted to the Division of Health Science Technology
in Partial Fulfillment of the Requirements for the Degree of
Master Of Science in Medical Informatics
at the Massachusetts Institute Of Technology

February 2002

Signature of Author: _____
Division of Health Science Technology
January 18, 2002

Certified by: _____
Isaac S. Kohane
Director, Children's Hospital Informatics Program
Thesis Co-supervisor

Certified by: _____
Peter Szolovits
Professor of Computer Science and Electrical Engineering
Thesis Co-supervisor

Accepted by: _____
Martha L. Gray
J W Kieckhefer Professor of Electrical Engineering
Co-Director, Harvard-MIT Division of Health Sciences Technology

Mutual Information Relevance Networks: Functional Genomic Networks Built From Pair-wise Entropy Measurements

by

Atul Janardhan Butte

Submitted to the Division Of Health Science Technology
on January 18, 2002 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in
Medical Informatics

ABSTRACT

Increasing numbers of methodologies are available to find functional genomic clusters in RNA expression data. We describe a technique that computes comprehensive pair-wise mutual information for all genes in such a data set. An association with a high mutual information means that one gene is non-randomly associated with another; we hypothesize this means the two are related biologically. By picking a threshold mutual information and using only associations at or above the threshold, we show how this technique was used on a public data set of 79 RNA expression measurements of 2,467 genes to construct 22 clusters, or relevance networks. The biological significance of each Relevance Network is explained.

Thesis Co-supervisor: Isaac S. Kohane
Title: Associate Professor of Pediatrics, Harvard Medical School

Thesis Co-supervisor: Peter Szolovits
Title: Professor of Computer Science and Engineering

Table of Contents

1 Introduction

1.1 Increasing number of methodologies available to functionally cluster genes

With the human genome sequencing nearing completion in one year and with the increasing use of microarrays to determine expression levels across the known genome, the problem of predicting the function of newly discovered genes has taken center stage. Newly developed techniques in bioinformatics use sequence, organism, and expression information to create clusters of genes with related functions. Current methodologies in functional genomics that use RNA expression data for clustering can be roughly divided into three categories: simple criteria matching, those that use Euclidean distance, and comprehensive pair-wise comparisons.

The first category contains the simplest use of RNA expression data sets. Levels are measured before and after an intervention. Fold-differences are calculated for each gene and the genes are sorted accordingly. Genes that demonstrate a fold-change greater than a given threshold are then considered "clustered" with the intervention. There have been several studies using this technique. [1, 2]

Self-organizing maps (SOM) are in the second category. This methodology uses multi-dimensional points corresponding to genes. Coordinates for these points represent expression levels at various time points. A grid of centroids is imposed in the multi-dimensional space, then allowed to drift towards collections of points. When completed, centroids reflect clusters of genes demonstrating similar time-course behavior. In this way, related genes have a smaller Euclidean distance in the multi-dimensional space. Tamayo, et al., used this technique to functionally cluster genes into various patterned time-courses in HL-60 cell macrophage differentiation. [3] Törönen, et al., used a hierarchical SOM to cluster yeast genes responsible for diauxic shift. [4]

The third category reflects those methodologies that comprehensively compare all genes against each other using a dissimilarity measure. Eisen, et al., took expression levels at various time points and created a vector for each gene. He then compared all genes against each other and recorded the correlation coefficient between vectors, then constructed a phylogenetic-type tree with branch lengths proportional to the correlation coefficients. [5, 6]

One methodology in both the second and third categories involves the construction of phylogenetic-type trees with branch length proportional to the Euclidean distance between genes, with coordinates again representing expression levels at various time points. Wen, et al., used this technique to find five waves of expression during embryonic neural development. [7, 8]

1.2 Relevance Networks

We have previously developed a methodology, termed *relevance networks*, that takes large data sets of clinical laboratory results and ascertains facts of human physiology by performing pair-wise correlation coefficients. [9]

Relevance networks are a technique where one evaluates the similarity of features by comprehensively comparing all features with each other in a pair-wise manner over the same set of cases. Strictly speaking, relevance networks are defined and implemented as a graph

$$G = \left\{ g_1, g_2, ..., g_n, \left\{ e_{1_1}, e_{1_2}, ..., e_{1_m} \right\}, \left\{ e_{2_1}, e_{2_2}, ..., e_{2_m} \right\}, ..., \left\{ e_{1_1}, e_{p_2}, ..., e_{p_m} \right\} \right\}$$

where n nodes $(g_1, g_2, ..., g_{n_m})$ are connected by p sets of m edges $(\{e_{1_1}, e_{1_2}, ..., e_{1_m}\}, \{e_{2_1}, e_{2_2}, ..., e_{2_m}\}, ..., \{e_{1_1}, e_{p_2}, ..., e_{p_m}\})$, where $m = \frac{n^2 - n}{2}$, and where each edge has a single value. In other words, each set of p edges completely connects the n nodes, where each pair of nodes is connected by a single edge. In practice, each set of p

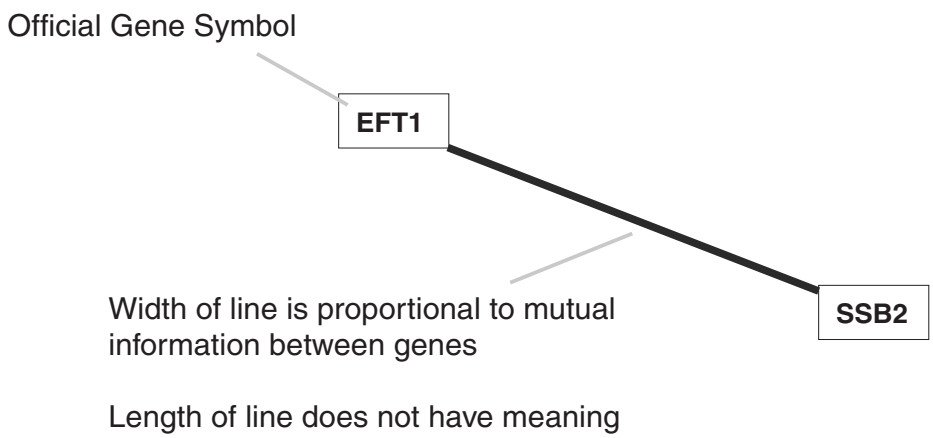

Figure 1: Graphical representation of relevance networks.

edges represents a different dissimilarity measure (such as Euclidean distance, correlation coefficient, or mutual information), and several of these may be simultaneously applied to the same set of *n* nodes. When used with microarray data, genes are represented as nodes, and edges are labeled with a real-valued score, which represents the strength of association between two genes.

Relevance networks are viewed or displayed by creating a subset $G_s(f_1, f_2, \ldots, f_p, t_1, t_2, \ldots, t_p)$, where $t_1 \ldots t_p$ are values (considered thresholds), and $f_1 \ldots f_p$ are functions applying the threshold to each set of edges. For each i of the p sets of edges in G, only those edges where $f_i\left(t_i, \left\{e_{i_1}, e_{i_2}, \ldots, e_{i_m}\right\}\right)$ is true are kept in the subset G_s. Typically, if the edges $\left\{e_{i_1}, e_{i_2}, \ldots, e_{i_m}\right\}$ contain values between –1.0 and 1.0, then t_i is set to a value between 0 and 1 and f_i returns true for e_j if $\left|e_j\right| >= t_i$, for all j between 1 and m, where $\left|e_j\right|$ means the absolute value of e_j. When applied to microarray data, this translates into a biological hypothesis that those edges assigned a higher positive value or lower negative value are more likely to represent hypotheses of a biological relationship. Using a threshold serves to break apart the completely connected network graph into a set of smaller graphs. The resultant relevance networks are displayed in a graphical manner similar to figure 1.

The choice of dissimilarity measures used to calculate the scores is arbitrary, and a previous implementation of relevance networks used correlation coefficient to score the

similarity of patterns of features. [10] Relevance networks are interpreted by translating the scores, relations, networks and the dissimilarity measure used into a set of hypotheses. The hypothesis behind any interpretation of relevance networks is that those pairs of features with the highest (or lowest) scores correspond to a hypothesis of interaction that can be tested.

Relevance networks have eight advantages over other clustering methods used in functional genomics. First, clustering algorithms based on Euclidean distance, like self-organizing maps, cannot handle missing data. For example, it is not obvious where to place a gene in a multi-dimensional space when even a single gene expression measurement is inaccurately measured or missing. Relevance networks handle missing data without difficulty by ignoring any case in the calculation of a pair-wise dissimilarity measure that is missing either of the two measurements.

Second, many clustering algorithms cannot handle negative interactions. As an example of a negative interaction in biology, *p53* is a tumor suppressor gene, in that increased levels of *p53* are known to be associated with decreased expression of other genes. The concept of negative interaction is clearly different than the concept of no interaction. Since Euclidean distance ranges from zero through the positive numbers, where zero is the strongest score, there is no representation for a negative interaction. In a multi-dimensional space, *p53* and the genes under its control would not be close together, and clustering techniques using Euclidean distance would not cluster these genes together. Dendrograms, as commonly used, also miss negatively correlated interactions. Without taking into account negative interactions, the behavior of tumor suppressor genes and other negative transcriptional factors will be ignored. Because relevance networks use correlation coefficients and mutual information as dissimilarity measures, the methodology take into account negative interactions between genes.

Third, adding additional experiments worsens the dimensionality problem in

Euclidean distance. If a multi-dimensional space is constructed with each experiment as a separate coordinate axis, then adding additional experiments drastically increases the amount of "empty space". This makes it increasingly difficult to find clusters that are biologically meaningful. Again, relevance networks can use dissimilarity measures where the confidence interval improves as additional experiments are added. [11]

Fourth, constructing a dendrogram always attempts to connect all leaves. Phylogenetic-type trees are always constructed by trying to connect all leaves, and there is no rapid method for determining the stronger links (*i.e.* the more believable ones) compared to the weaker ones. In essence, relevance networks provide a dial for "believability." One can quickly construct relevance networks at a high threshold, and then, if more novel hypotheses are needed, the dial can be lowered gradually to introduce slightly weaker links.

Fifth, although biological functional clusters likely have variable numbers of genes in them, phylogenetic-type trees connect all clusters into a single structure. A visual inspection is often needed to determine where to cut the tree apart. Relevance networks create multiple networks containing varying numbers of genes.

Sixth, phylogenetic-type trees can only cluster data of a single data-type. Mixing phenotypic measurements with expression measurements, for example, will produce trees with the leaves of phenotypic measurements scattered throughout the tree, which is not useful. Relevance networks can easily mix phenotypic measurements with expression measurements, and direct hypotheses and links are provided between different type data. [10]

Seventh, features can only be positioned in a single place within a dendrogram. Each gene is directly connected to the tree with only one stem. In reality, a transcription factor may be responsible for regulating the expression of multiple other genes, but a phylogenetic-type tree methodology will link that transcription factor only with the one

gene it most closely resembles in expression pattern. Relevance networks clearly show that if a gene is closely linked to few or many other genes, then each link is shown separately. This is also important when a gene is similar to two different groups of genes, or when a pharmaceutical agent displays activity similar to two different classes of compounds. [10]

Eighth, trees are constructed with only a single dissimilarity measure. A dendrogram can only cluster genes based on correlation coefficient or Euclidean distance, but not both. As mentioned above, relevance networks can mix multiple types of dissimilarity measures. For example gene A and gene B may be linked because of a high correlation coefficient, but gene B and gene C may be linked because of a high mutual information. Relevance networks can be constructed to include both types of links simultaneously.

1.3 Using entropy and mutual information to evaluate gene-gene associations

Our goal was to use this method to take large data sets of RNA expression measured under varying conditions and generate networks of hypotheses of gene-gene interactions. We compute the *entropy* of gene expression patterns and the *mutual information* between RNA expression patterns for each pair of genes. The entropy of an RNA expression pattern is a measure of the information content in that pattern, and is calculated using equation 1

$$H(A) = -\sum_{i=1}^{n} p(x_i)\ log2(\ p(x_i)\ ) \qquad \textit{(Equation 1)}$$

where *log2* is base 2 logarithm. [12] Higher entropy for a gene means that its expression levels are more randomly distributed.

Gene expression is measured on a continuous scale, yet equation 1 shows how entropy is computed using discrete probabilities. To calculate entropy, we use a histogram or binning technique. We first calculate the range of values for each gene

separately, and then divide that range into n sub-ranges. In equation 1, $p(x_i)$ equals the proportion of measurements in sub-range x_i. As n approaches infinity, the histogram will more accurately model the probability density function (PDF) for the gene.

Ideally, n would be set to a specific appropriate value for each gene. For example, if a particular gene's range of expression level is known to have only two functional states, say "on" and "off," n for that gene could be set to 2. Previous work in applying mutual information to gene expression measurements have operated in this way, by assigning only two allowable states for all genes, and quantizing continuous expression measurements into these two states. [7] By assuming only binary values for each gene, one can represent a state of expression of all genes as a single Boolean vector. State transitions can then be modeled as a Boolean network, and logical rules can be intuited from these networks. [13]

However, many genes are known to interact with other genes and proteins in a dose-response type of manner, where the specific amount of a gene affects downstream processes on a continuous scale. Most important, the specific number of functional states is currently unknown for the majority of genes. For our computations, we arbitrarily set $n = 10$ for all genes, though we acknowledge that changing this piece of *a priori* information can have a large impact on the ordering of gene-gene interactions by mutual information. [14]

The mutual information is a measure of the additional information known about one expression pattern when given another, as shown in equation 2.

$$MI(A, B) = H(A) - H(A \mid B) \qquad \textit{(Equation 2)}$$

Equation 2 can be restated as equation 3. Mutual information can be calculated by subtracting the entropy of the joint RNA expression patterns from the individual gene entropies.

$$MI(A, B) = H(A) + H(B) - H(A,B) \qquad \textit{(Equation 3)}$$

A mutual information at zero means that the joint distribution of expression values holds no more information than the genes considered separately. A higher mutual information between two genes means that one gene is non-randomly associated with the other. In this way, mutual information can be used as a dissimilarity measure between two genes related to their degree of independence. We hypothesize that the higher mutual information is between two genes, the more likely it is they have a biological relationship.

1.4 Construction of Relevance Networks

We used a publicly available RNA expression data set from Stanford, containing 79 separate measurements of 2,467 genes in *Saccharomyces cerevisiae*. [5] The specific methodology of how RNA expression was measured has been previously described. [15] Genes were measured under a variety of conditions, including diauxic shift, mitotic cell division cycle, sporulation, and temperature and reducing shocks, and at various time points for each condition. Measurements of all genes were compared against each other, resulting in 3,041,811 total pair-wise calculations of mutual information, ranging from 0.2 to 2.8. Each gene was thus completely connected to every other gene with a calculated mutual information.

We then chose a threshold mutual information (TMI) and displayed only those genes that were linked to others with a mutual information higher than the threshold. Out of the completely connected network of genes, we were left with clusters of genes, or relevance networks, that were more strongly connected to each other than the TMI.

We displayed the relevance networks graphically with nodes representing genes and lines between nodes representing hypothetical associations of genes. Relationships with higher mutual information were drawn with a thicker line. Nodes were positioned and edge crossings minimized using the Graph Editor Toolkit (Tom Sawyer Software, Berkeley, California).

2 Results

2.1 Distribution of Mutual Information Calculations

The distribution of the 3,041,811 pair-wise calculations of mutual information is shown in figure 2. The mode of the distribution of mutual information was 0.7.

To determine the significance of this distribution, a permutation analysis was performed. [16-18] The expression measurements for each gene were independently randomly permuted, so that the average expression level for each gene remained constant. The pair-wise calculations of mutual information were then repeated, and a distribution of the new pair-wise mutual information was recalculated. This process was repeated independently 30 times, such that 30 independently-computed random permutations of the data were used, and 30 permuted distributions of pair-wise mutual information were calculated.

The 30 permuted distributions were summed, and an *average distribution* was computed by dividing by 30. The average of the 30 permuted distributions is included in figure 2. Permutation was unable to create any associations with mutual information over 1.3. Thus, associations found in the original data set with mutual information over 1.3 could be viewed as significant.

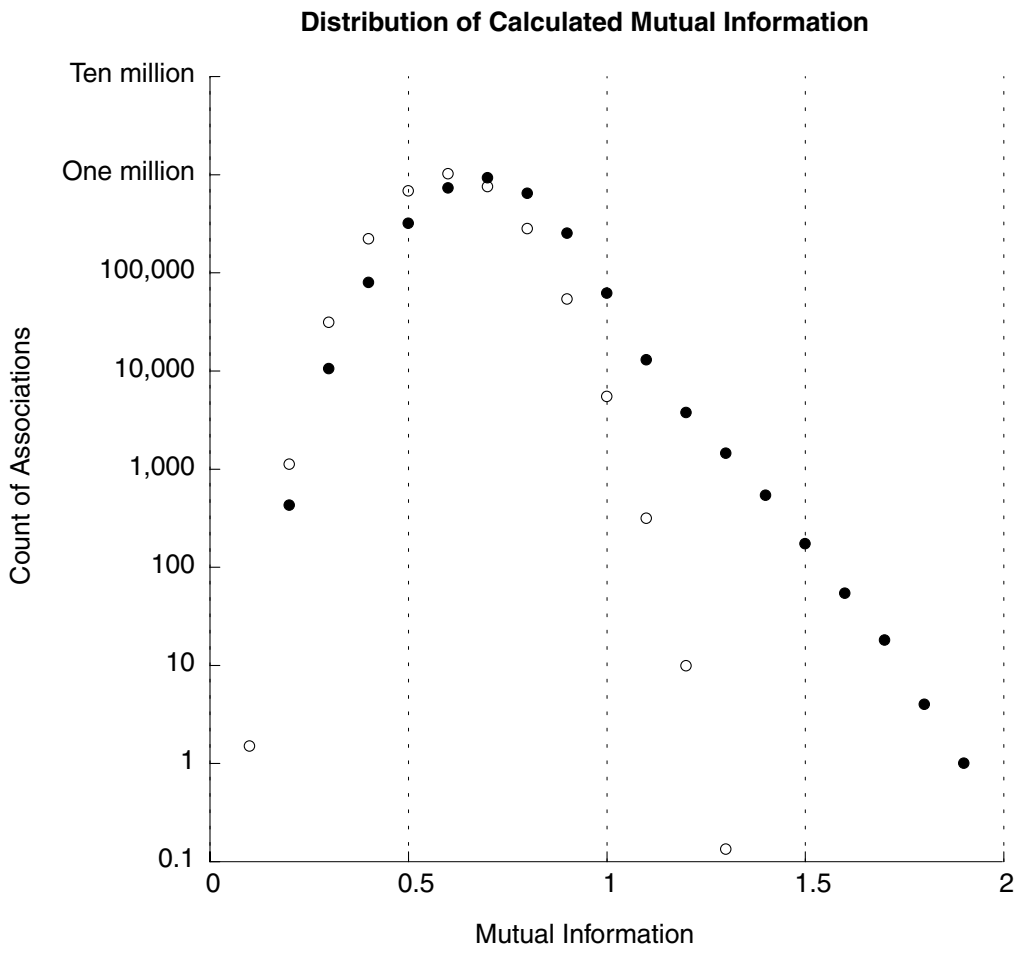

Figure 2: Pair-wise mutual information was calculated between 79 measurements of RNA expression of 2,467 genes in *Saccharomyces cerevisiae* and the distribution of these is shown with filled circles. The same was calculated using permuted RNA expression measurements; the average distribution from 30 permuted repetitions is shown with open circles.

2.2 Changing Threshold Affects Size and Number of Relevance Networks

As the TMI is dropped from 2.0 to 1.2, the number and size of the relevance networks increases, as shown in figure 3. More nodes are introduced, and these nodes form large numbers of small networks. With an increasing number of nodes, the number of potential links between them increases; yet the *connectivity*, defined as the number of actual links relative to the potential number of links, drops from 100% to 1%. This indicates that most nodes are connected to only a few other nodes. When the TMI is decreased from 1.2 to 0.8, the number of networks drops as the newly included nodes serve to merge existing networks with each other.

At a TMI of 0.8, all the genes belong to a single relevance network. The connectivity of the networks then quickly increases until the TMI reaches 0.2, when the connectivity reaches 100%.

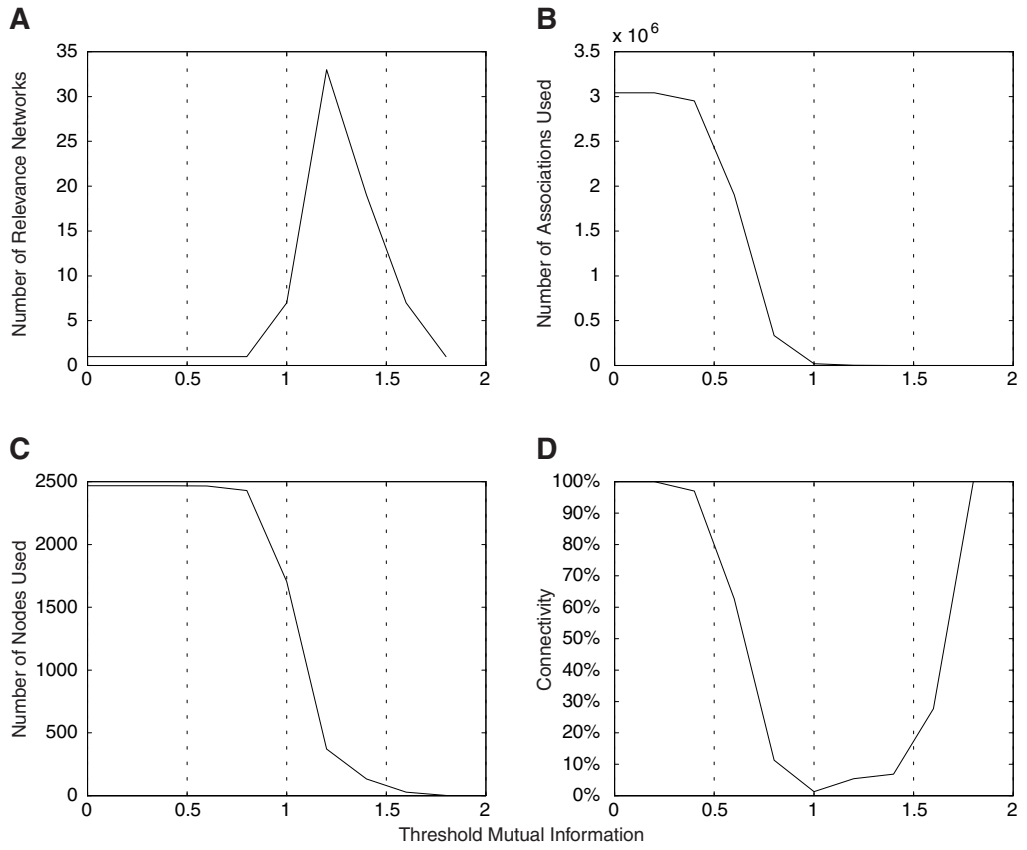

Figure 3: (A) Number of relevance networks as a function of the chosen threshold mutual information (TMI). (B) Number of generated hypothetical gene-gene associations versus TMI. (C) Number of genes used versus TMI. (D) Connectivity of relevance networks versus TMI.

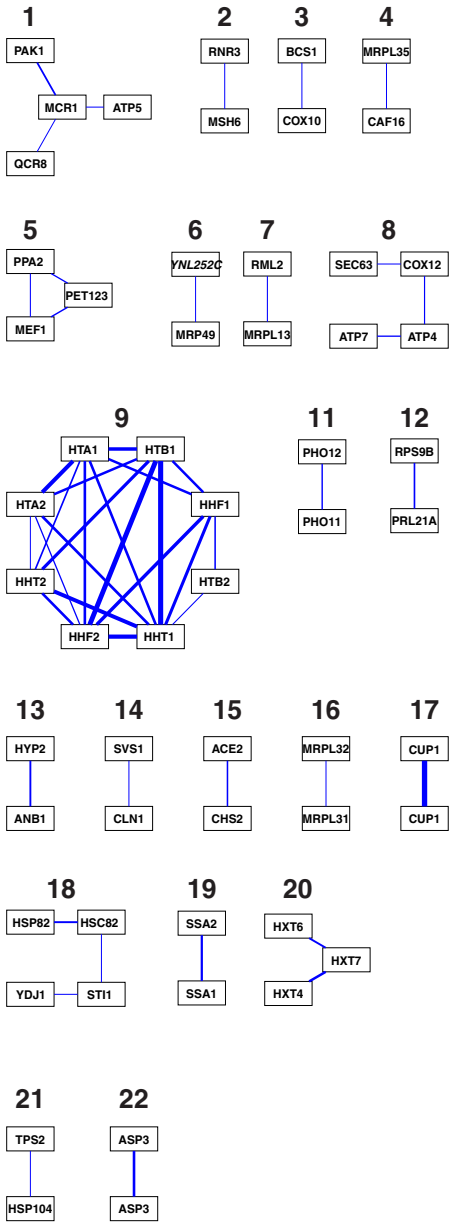

Figure 4: Twenty-one of the 22 relevance networks created with TMI set to 1.3. Node labels represent gene abbreviations; names can be found using http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html and are explained in the text.

2.3 Relevance Networks Seen in Saccharomyces cerevisiae

Using the analyses above, we determined that the largest number of relevance networks was at a TMI of 1.2, and the highest mutual information reached in permuted data was 1.3. Thus, we set the TMI to 1.3, which produced 22 relevance networks using a total of 199 genes. Twenty-one relevance networks are shown in figure 4. Enlarged versions of these networks are available at http://www.chip.org/genomics/. We saw four main classes of networks: those that linked identical genes, those linking genes with similar functions, those that linked genes in the same biological pathway, and combinations of these. The majority of the hypothetical associations could be validated using the biological literature.

Two networks were found to link identical genes. Network 17 linked two repeated open reading frames encoding *cup1*, a copper metallothionein, and network 22 connected two copies of L-aspariginase II found on chromosome 12.

Nine networks clustered genes that have similar functions. Network 9 tightly linked eight genes coding for histones. Network 11 linked *pho10* and *pho11*, two secreted acid phosphatases and network 12 linked *s9b* and *l21a*, two ribosomal proteins. Network 13 connected *hyp2* and *anb1*, both of which are involved in translation initiation. Network 19 connected *ssa1* and *ssa2*, both 70 kilodalton heat shock proteins. Network 20 clustered the three hexose transporters *hxt4*, *hxt6* and *hxt7*, which are known to have increased transcription when extracellular glucose increases. Networks 6, 7 and 16 linked mitochondrial ribosomal proteins.

Five networks linked genes known to be involved in the same biological pathway. Network 2 linked *msh6*, which repairs base pair mismatches and *rnr3*, induced as a response to DNA damage. Network 3 connected *bcs1* and *cox10*, both known to be involved in assembly of the cytochrome complex. Network 21 linked *tps2*, trehalose-6-phosphate phosphatase, and *hsp104*, a chaparone. This exact interaction has been

described in the literature; *hsp104* contributes to the heat shock accumulation and degradation of trehalose.

Network 15 linked *ace2*, a known regulator of chitinase expression, and *chs2*, chitin synthase II. Network 18 connected the two isoforms of the chaperone *hsp90*, *hsp82* and *hsc82*. *Sti1*, which is also connected in this network, is known to regulate *hsp90* ATPase activity and is involved in regulating activity of the glucocorticoid receptor. *Ydj1* works earlier in the maturation of the glucocorticoid receptor and was linked to *sti1* in network 18.

The remaining six networks contained various types of links, including a few associations not presently explained in the biological literature. Network 1 linked cytochrome B5 to F1F0-ATPase 5p, ubiquinol:cytochrome-C reductase subunit VIII, and the *pak1* protein kinase. Ubiquinol:cytochrome C reductase is known to regulate cytochrome B5. F1F0-ATPase is known to regulate cytochrome C. The link to *pak1* is unexplained in the biological literature.

Network 4 connected *mrpl35*, a mitochondrial ribosomal protein, and *caf16*, possibly involved in essential mitochondrial function. Network 14 linked *cln1*, a G1 cyclin and *svs1*, a gene required for vanadate resistance, but with no known role in cell cycle regulation.

Network 5 linked *pet123*, a protein involved in mitochondrial translation and *mef1*, a mitochondrial translation factor. These both were linked to *ppa2*, a mitochondrial inorganic pyrophosphatase essential for mitochondrial function, but which has not been implicated in mitochondrial translation.

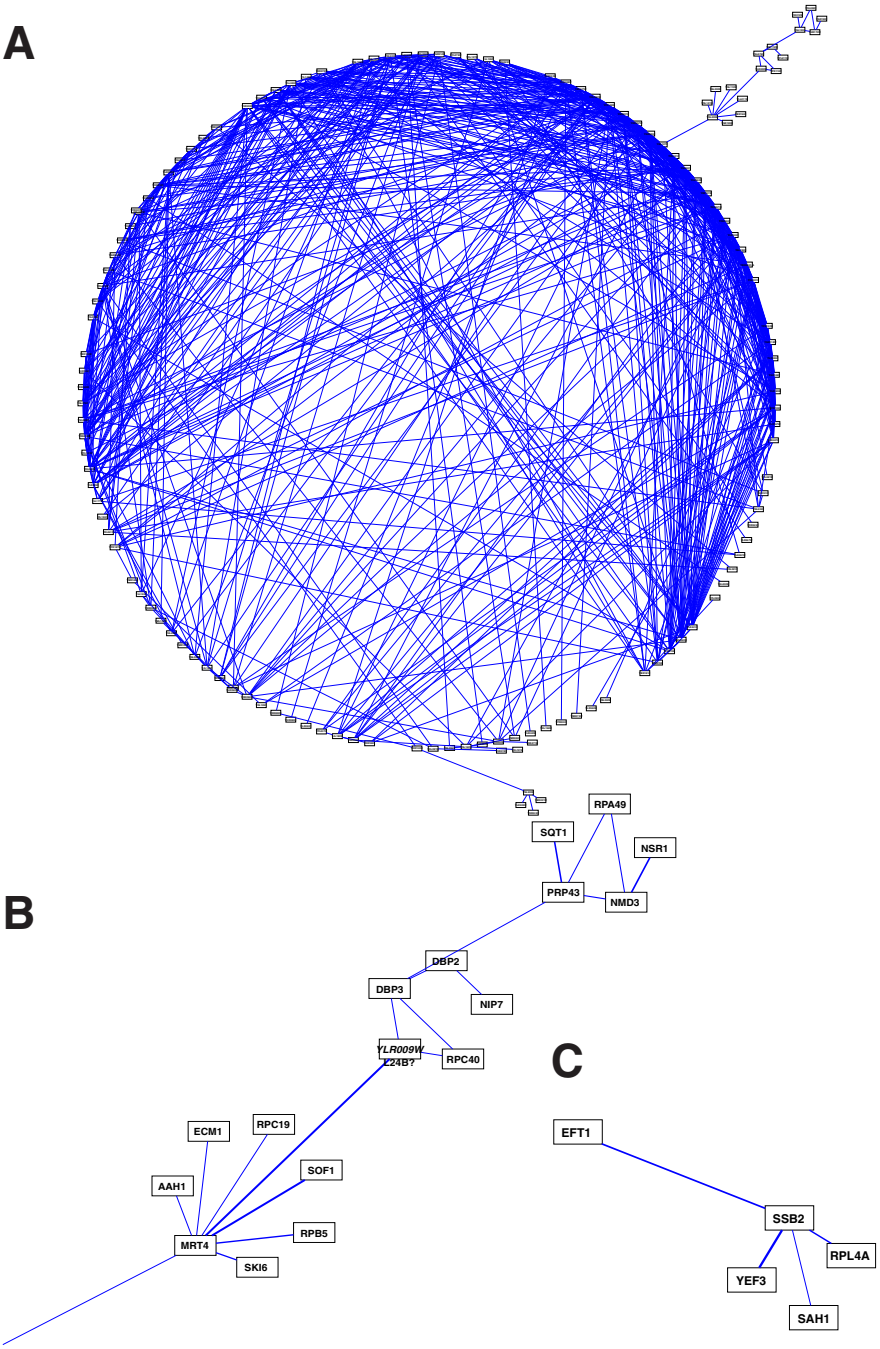

Figure 5: (A) Largest of the Relevance Networks created with TMI was set to 1.3. (B) and (C) Two branches enlarged from (A) and explained in the text.

The largest network, network 10, clustered 143 genes and is shown in figure 5a. Of these, 102 were various components of the large and small ribosomal subunits and 8 were translation initiation factors. One branch from the larger network is shown in figure 5b. Here, *mrt4*, presumed to be involved in mRNA turnover, was linked to *aah1*, involved in purine salvage; *ski6*, which represses double-stranded RNA replication; *sof1*, a protein involved in nucleolar rRNA processing; *rpb5*, a subunit of RNA polymerases I, II and III; and open reading frame (ORF) YLR009W, whose function is unknown. This ORF was linked to *rpc40*, a shared subunit of RNA polymerase I and III and *dbp3*, an RNA helicase, which in turn was linked to *dbp2*, another RNA helicase and *prp43*, an RNA helicase-like factor, and other ribosomal and RNA processing proteins.

Another branch is shown in figure 5c, where *eft1*, an elongation factor, was linked to *ssb2*, a 70 kilodalton heat shock protein associated with translating ribosomes, which was linked to *yef3*, another elongation factor; *sah1*, s-adenosyl-l-homocysteine hydrolase, a cytoplasmic adenosine-binding protein; and *rpl4a*, one of two genes encoding ribosomal protein L4.

3 Discussion

3.1 Summary of Findings

Using this technique of linking all genes by calculating comprehensive pair-wise mutual information, 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 199 genes into 22 relevance networks at the TMI of 1.3. Decreasing the TMI will introduce more genes and hypothetical associations. Even though some of these new associations may be noise because high mutual information may still be calculated by chance, the associations at lower TMI may also represent novel hypotheses. Increasing the TMI will restrict the relevance networks to include only the strongest hypothetical associations.

3.2 Strengths of Relevance Networks

We found three specific advantages of the relevance network methodology. First, using mutual information is more general than using correlation coefficients to model the relationship between genes. The correlation coefficient is more easily distorted when points are not uniformly distributed across the axes. For example, two genes with a single high expression level measured in the same cellular condition will have a higher correlation coefficient regardless of the expression levels seen in other cellular conditions. In this way, outlying points bias correlation coefficients. Mutual information uses each expression level measurement equally regardless of the actual value, and thus is not biased by outliers.

Because mutual information is a more general model, complex relationships

between genes can be modeled. For example, if one gene acts as a transcription factor only when it is expressed at a midrange level, then the scatterplot between this transcription factor and other genes might more closely fit a bell curve rather than a linear model, and might be scored with a low correlation coefficient. Mutual information does not require an *a priori* choosing of any particular model.

A second advantage of relevance networks is that relationships are displayed in a graph instead of a phylogenetic-type tree. The advantage is that complex interactions are more easily visualized. Although biological functional clusters likely have variable numbers of genes in them, phylogenetic-type trees connect all clusters into one structure; relevance networks have variable size.

In a phylogenetic-type tree, each gene is directly connected to only one other gene, the one it is most closely related to. Relevance networks connect nodes directly and indirectly with many or few links. There is valuable information in the number of links within a Relevance network. Nodes that are connected directly and indirectly with more links represent genes that are not only related directly to each other, but also as an aggregate. Relevance networks with higher degrees of cross-connection are thus more trusted, because they suggest that not only are two genes related, but that other genes exist that are related to both similarly.

The third strength is that relevance networks need not be restricted to genomic clustering. Histological or clinical features can be quantitated and added to the array; pair-wise calculation of mutual information can easily include them and can thus potentially cluster expression of particular genes with specific phenotypes.

3.3 Future Directions

One important next step will be to construct relevance networks while modeling measurement noise. Like many measurement systems, RNA expression levels as measured by microarrays are not perfectly reproducible when experiments are repeated.

This noise in RNA expression level measurement can come from many sources: intrachip defects, variation within a single lot of chips, variation within an experiment, and biological variation for a particular gene. In our current methodology to calculate the mutual information between two genes, we model each sample/case as a point with a pair of quantized expression values as coordinates, and we use a two-dimensional histogram to approximate the joint PDF.

If this were expanded into a continuous scale, and if one assumed a normal distribution for the measurement noise, one could instead represent each sample/case as a two-dimensional normal PDF, with the mean of the distribution centered at the actual coordinate pair of expression measurements, and the variance of the distribution calculated as a function proportional to the noise of the two measurements. Then, instead of using a histogram to calculate the mutual information from discrete points, one could instead use a Parzan density function to model the overall joint PDF as a sum of all the individual PDF. [14] From this, a continuous mutual information may be estimated or calculated. [12]

This is important because as more is learned about the noise and reproducibility of expression level measurements, this methodology could be used to represent gene expression levels as a distribution instead of just a single point and can still find functional patterns.

A gene-gene association with a high mutual information means the expression of one gene is predictable given the other. However, we acknowledge that there may be many gene-gene interactions that have high mutual information, yet contain a few points that do not fit the overall interaction. These samples/cases may indicate significant deviations from, or exceptions to, a gene-gene interaction model and should be studied separately.

Finally, this technique will be used to analyze human gene expression patterns,

not only to find the functional clusters in normal physiology, but also to hopefully find targets susceptible to therapy in disease physiology.

Acknowledgments

Portions of this paper were originally published in [19]. I thank World Scientific Publishing Co. for granting me permission to derive from that work.

This research was supported in part by the National Library of Medicine grant "Research Training in Health Informatics", 5T15 LM07092-07 and R01 LM06587-01, the National Heart Lung and Blood Institute, 1U01HL066582-01, the National Institute of Diabetes and Digestive and Kidney Diseases, 1U24DK058739-01, National Institute of Neurological Disorders and Stroke, 1P01NS040828-01A1, National Institute of Allergy and Infectious Diseases, 1R01AI050987-01, the Merck / Massachusetts Institute of Technology Graduate Research Fellowship, the Lawson Wilkins Pediatric Endocrine Society Award, the Endocrine Fellows Foundation, and the Genentech Center for Clinical Research and Education.

I would like to thank my parents, Janardhan and Mangala Butte, for their support and encouragement and wisdom in exposing me to computers at an early age. I would like to recognize my brother, Dr. Manish Butte, for being a role model to me with his academic pursuits. My wife, Dr. Tarangini Deshpande, was not only a continuous source of encouragement, but was also an expert source for discussions in molecular biology.

I would like to thank my division chief Dr. Joseph Majzoub for his support of my clinical and research environment, allowing me to pursue this and future degrees. I would like to thank Professor Peter Szolovits for his guidance and advice, as well as his support for my upcoming pursuit of a PhD degree in Medical Engineering / Medical Physics. Finally, I would like to thank my mentor and friend, Dr. Isaac S. Kohane, for his support and development of my career in Medical and Bio-informatics. I sincerely owe my being

in this field of research to Dr. Kohane's strong collaborative connections to biologists and his continuing support of my exploration in functional genomics.

References

1. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14(4):457-460.

2. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci U S A 1997;94(6):2150-2155.

3. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 1999;96(6):2907-2912.

4. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999;451(2):142-6.

5. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863-14868.

6. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, et al. The transcriptional program in the response of human fibroblasts to serum. Science 1999;283(5398):83-87.

7. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. Pac Symp Biocomput 1998:42-53.

8. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, et al. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci U S A 1998;95(1):334-339.

9. Butte A, Kohane I. Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks. In: Lorenzi N, editor. Fall Symposium, American Medical Informatics Association; 1999; Washington, DC: Hanley and Belfus; 1999. p. 711-715.

10. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A 2000;97(22):12182-6.

11. Kleinbaum DG, Kupper LL, Muller KE. The Correlation Coefficient and Straight-Line Regression Analysis. In: Applied Regression Analysis and Other Multivariable Methods. Belmont, California: Duxbury Press; 1988. p. 90-91.
12. Shannon CE, Weaver W. The Mathematical Theory of Communication. Chicago: University of Illinois Press; 1949.
13. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 1998:18-29.
14. Bishop CM. Neural Networks for Pattern Recognition. Oxford: Clarendon Press; 1995.
15. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 1996;93(20):10614-9.
16. Park PJ, Pagano M, Bonetti M. A nonparametric scoring algorithm for identifying informative genes from microarray data. Pac Symp Biocomput 2001:52-63.
17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98(9):5116-21.
18. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531-7.
19. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000:418-29.