Pacific Symposium on Biocomputing 5:415-426 (2000)

MUTUAL INFORMATION RELEVANCE NETWORKS: FUNCTIONAL GENOMIC CLUSTERING USING PAIRWISE ENTROPY MEASUREMENTS

A. J. BUTTE, I. S. KOHANE
Children's Hospital Informatics Program and
Division of Endocrinology,
300 Longwood Avenue,
Boston, MA 02115, USA

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.

1 Introduction

1.1 Increasing number of methodologies available to functionally cluster genes

With the human genome sequencing nearing completion in four years 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 pairwise 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 metric. 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 Using entropy and mutual information to evaluate gene-gene associations

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]

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. Instead of calculating correlation coefficients, 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. Higher entropy for a gene means that its expression levels are more randomly distributed.

Mutual information calculated from binary measurements of gene expression has previously been proposed as a method of determining cell state transition rules.[10] However, gene expressions are measured on a continuous scale, not binary. Yet entropy is computed using discrete probabilities. Thus, to calculate entropy, we use a histogram technique. We first calculate the range of values for each gene, then

divide that range into n sub-ranges. In equation 1, $p(x_i)$ equals the proportion of measurements in sub-range x_i (or frequency). As n approaches infinity, the histogram will more accurately model the probability density function for the gene. For our computations, we set $n = 10$.

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 metric 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.

1.3 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.[11] 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 pairwise 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

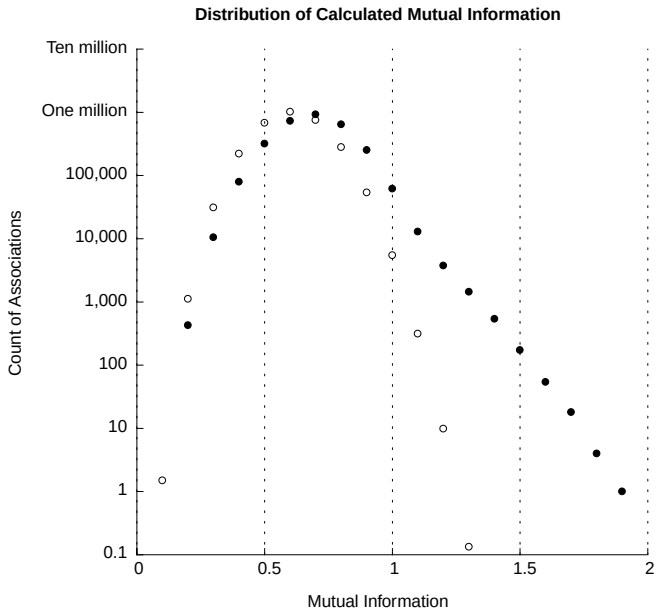

Figure 1: Pair-wise mutual information was calculated between 79 measurements of RNA expression of 2,467 genes in *Saccharomyces cerevisiae* and the distribution of these is shown with filled circles. The same was calculated using permuted RNA expression measurements; the average distribution from 30 permuted repetitions is shown with open circles.

were positioned and line 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 1. The mode of calculated mutual informations was around 0.7.

To determine the significance of this distribution, the RNA expression measurements were permuted 30 times and a distribution of the new pair-wise mutual informations was recalculated for each permutation. The average of the 30 permuted distributions is also shown in figure 1. 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.

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 2. 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%.

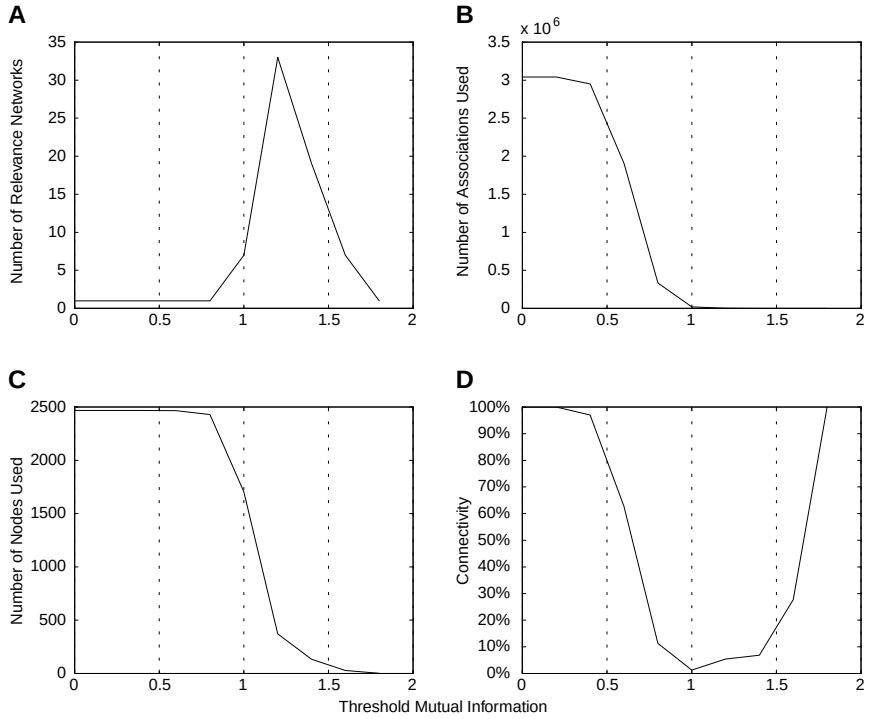

Figure 2: (A) Number of Relevance Networks as a function of the chosen threshold mutual information (TMI). (B) Number of generated hypothetical gene-gene associations versus TMI. (C) Number of genes used versus TMI. (D) Connectivity of Relevance Networks versus TMI.

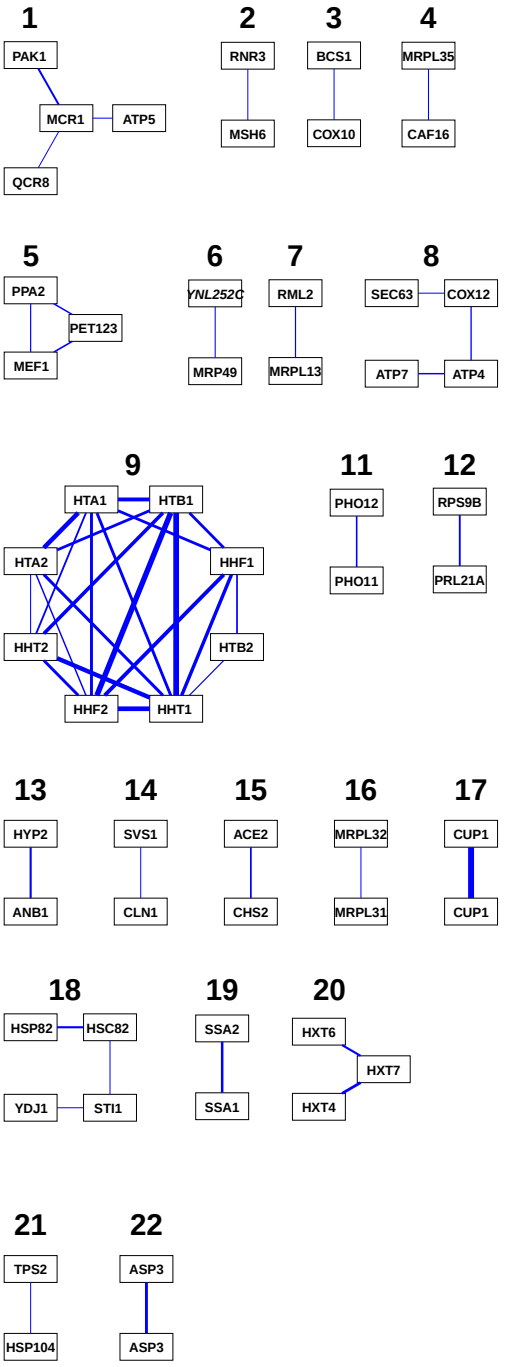

Figure 3: Twenty-one of the 22 Relevance Networks created with TMI was set to 1.3. Node labels represent gene abbreviations; names can be found using http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html and are explained in the text.

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 3. 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.

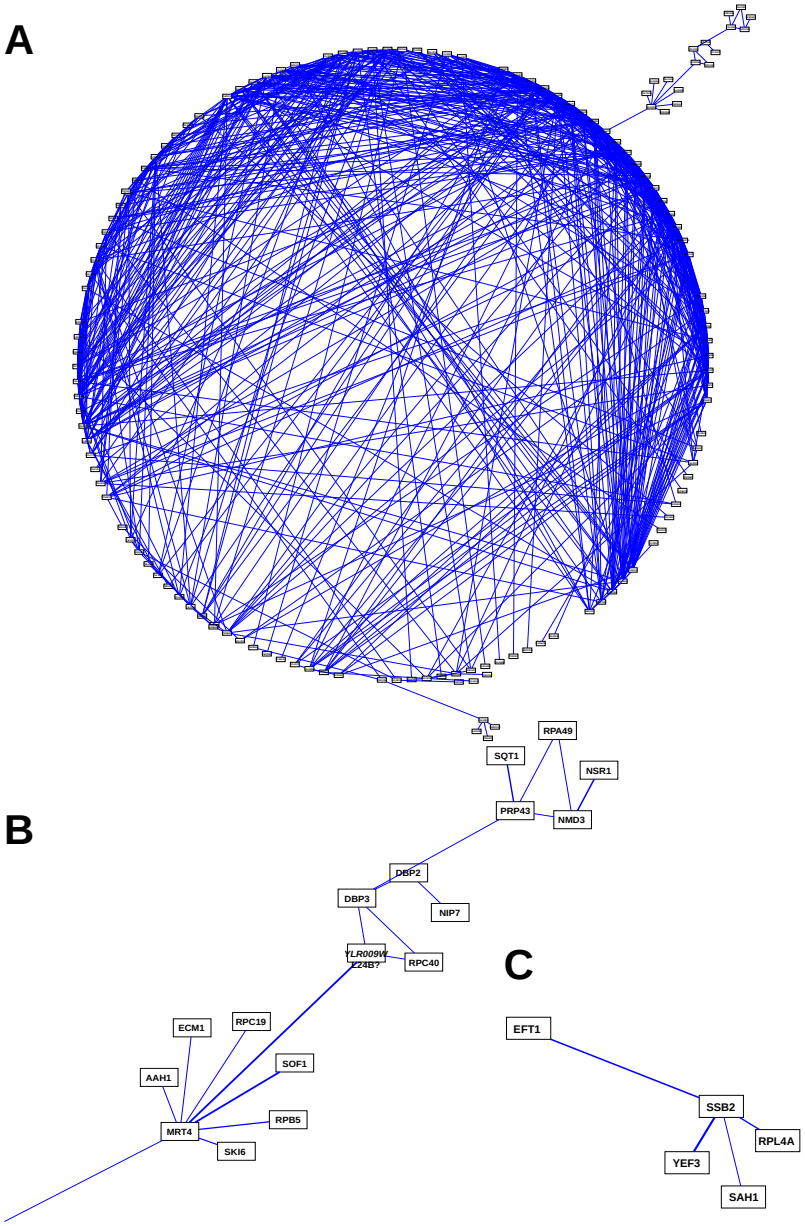

Figure 4: (A) Largest of the Relevance Networks created with TMI was set to 1.3. (B) and (C) Two branches enlarged from (A) and explained in the text. Node label in italics represents accession number of open reading frames with no abbreviation.

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.

Network 8 connected one subunit of the F1F0-ATPase complex to F0-ATP synthase, then links that to a subunit of cytochrome-c oxidase. These three are known to be involved in ATP synthesis and oxidative phosphorylation. Cytochrome-c oxidase was linked to *sec63*, a gene that assists transit of secretory proteins across the endoplasmic reticulum. This link has not previously been described.

The largest network, network 10, clustered 143 genes and is shown in figure 4a. 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 4b. 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 4c, 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 associations are noise because some high mutual informations 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.

3.2 Strengths of Relevance Networks

We found four 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 resemble a normal distribution 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.

A third advantage of constructing Relevance Networks using mutual information is that expression levels can be modeled to include measurement noise. Like any measurement, RNA expression levels are not always replicated 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, we use a two-dimensional histogram to approximate the joint probability density function between two genes to calculate the mutual information. However, instead of using a histogram, this methodology can be expanded to use a Parzan density function, where the joint probability distribution function is represented as the sum of multiple normal distributions. This is important because as more is learned about the noise and reproducibility of expression level measurements, this methodology can be modified to represent RNA expression levels as a distribution instead of just a single point and can still find functional patterns.

The fourth 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

In addition to expanding the technique to model measurement noise and clinical measurements, we intend to introduce temporality into Relevance Networks. Expression of a particular RNA at a given time may be associated with the expression of another RNA some time in the future. By modeling this, we can start to approach assigning causality. In addition to this, Relevance Networks may themselves change over time.

A gene-gene association with a high mutual information means the expression of one RNA is predictable given the other. However, there are many exceptions where the expression of an RNA was not what was predicted, even in a strong association. These exceptions may indicate significant deviations of the model and should be studied.

Finally, this technique will be used to analyze human RNA 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

This research was supported in part by the grant “Research Training in Health Informatics” funded by the National Library of Medicine, 5T15 LM07092-07 and R01 LM06587-01.

References

1. DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]. *Nat Genet* 1996;**14**(4)**:**457-60.
2. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci U S A* 1997;**94**(6)**:**2150-5.
3. Tamayo P, Slonim D, Mesirov J, 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 [In Process Citation]. *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)**:**14863-8.
6. Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum [see comments]. *Science* 1999;**283**(5398)**:**83-7.
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, et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci U S A* 1998;**95**(1)**:**334-9.
9. Butte A, Kohane I. Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks. *Proc Amia Symp* 1999**:**In Press.
10. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput* 1998**:**18-29.
11. 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.