Mutual Information Based Extrinsic Similarity for Microarray Analysis

Duygu Ucar[1], Fatih Altiparmak[2], Hakan Ferhatosmanoglu[1], and Srinivasan Parthasarathy[1]

[1] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH
[2] ASELSAN A.S. Radar, EW, and Intelligence Systems Division, Turkey

Abstract. Genes responding similarly to changing conditions are believed to be functionally related. Identification of such functional relations is crucial for annotation of unknown genes as well as the exploration of the underlying regulatory program. Gene expression profiling experiments provide noisy datasets about how cells respond to different experimental conditions. One way of analyzing these datasets is the identification of gene groups with similar expression patterns. A prevailing technique to find gene pairs with correlated expression profiles is to use linear measures like Pearson's correlation coefficient or Euclidean distance. Similar genes are later compiled into a co-expression network to explore the system-level functionality of genes. However, the noise inherent in microarray datasets reduces the sensitivity of these measures and produces many spurious pairs with no real biological relevance. In this paper, we explore an extrinsic way of calculating similarity of two genes based on their relations with other genes. We show that 'similar' pairs identified by extrinsic measures overlap better with known biological annotations available in the Gene Ontology database. Our results also indicate that extrinsic measures are useful in enhancing the quality of co-expression networks and their functional subnetworks.

1 Introduction and Related Work

Microarray experiments are now being used to profile expression levels of genes under changing experimental conditions. To analyze these profiles in an attempt to answer diverse biological questions, various techniques and ideas have been proposed. Of particular interest to many scientists is the identification of genes whose expression profiles are similar, since genes with similar cellular functions have been theorized to respond similarly to changing conditions [9]. As a result, an efficient similarity measure for microarray analysis is fundamental for understanding the cellular processes [24] and annotating unknown genes.

There has been a growing interest in linking genes whose expression profiles are similar to construct co-expression networks. These networks and their highly modular subnetworks are invaluable sources of information for system-level gene processes [29,4]. Similarity of two genes can be deduced from expression levels

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 424–436, 2009.

of these genes across all samples [12,29,7]. However, the noise inherent in microarray datasets limits the sensitivity of such analysis. Since any microarray measurement is likely to fluctuate due to many possible sources of error, a similarity based solely on expression measurements of two genes is more error-prone than a similarity based on expression measurements of many genes. In addition, inferring the similarity of two genes based on their relations with a set of other genes will be in accordance with the biological hypothesis about gene products acting as complexes to accomplish certain cellular level tasks [23]. Thus, here we investigate use of extrinsic similarity measures to analyze microarray studies.

The use of extrinsic measures and their advantages have been previously studied for various data mining problems [5,6]. Das et al. [5] proposed using extrinsic measures on market basket data in order to derive similarity between two products from the buying patterns of customers. Palmer et al. [19], defined an extrinsic similarity measure (REP) with an analogy to electric circuits. Both groups concluded that extrinsic measures can give additional insight into the data. Recently, Ravasz et al. [20], took a step towards using extrinsic properties along with the intrinsic similarity. Their measure, the Topological Overlap Measure (TOM), infers similarity of two nodes in a biochemical network in terms of their pairwise similarity as well as the number of their common neighbors. In a previous work we discussed using mutual independence notion to derive an effective extrinsic dissimilarity measures [25].

We introduce application of extrinsic similarity measures for identification of co-expressed genes. We propose extrinsic measures motivated by *Mutual Information* notions from Information Theory. The proposed similarity measures are evaluated on a well-studied cancer microarray dataset [1] obtained with Affymetrix oligonucleotide arrays, as well as a yeast microarray data generated with custom complementary DNA (cDNA) arrays [10]. For both datasets and platforms, we showed that gene pairs obtained by extrinsic similarity measures better overlap with known biological annotations from the Gene Ontology (GO) database when compared to the Pearson's correlation coefficient and the TOM. To further analyze efficacy of extrinsic measures for gene function inference, we constructed co-expression networks by using different measures. We observe that co-expression networks constructed based on extrinsic measures contain less spurious and more biologically verified edges compared to their counterparts generated with other measures. We also studied modular structure of these networks by decomposing them into co-expressed modules. We found that gene modules extracted from Extrinsic Gene Networks are also functionally more homogeneous in comparison.

To summarize, our main contributions in this study are:

- The study of Information Theory concepts, Conditional Mutual Information and Specific Mutual Information, for genes derived from their associations with other genes
- The introduction of extrinsic measures for microarray datasets based on Conditional Mutual Information and Specific Mutual Information

– The demonstration of the efficacy of using extrinsic measures in inferring pairwise gene similarities, constructing co-expression networks, and identifying co-expressed modules.

2 Similarity Measures

To quantify the resemblance of two points, one needs a measure of similarity. Similarity measures can be categorized into two: *extrinsic* and *intrinsic* similarity. An *intrinsic* similarity of two points i and j is purely defined in terms of the values of i and j. On the other hand, an *extrinsic* similarity measure takes into account other points to infer similarity of i and j. Previous studies have shown the usability of extrinsic similarity measures in other domains [5,6]. The standard method to infer similarity of two genes from their expression patterns is to use a linear *intrinsic* similarity such as the Pearson's correlation coefficient. To our knowledge, we are the first to study extrinsic measures for microarray datasets [25].

2.1 Intrinsic Similarity

Intrinsic similarity is purely defined on the points in question. In the context of microarray analysis, the *intrinsic* similarity of two genes is defined on the measured expression levels of these two genes over all samples. In a typical microarray experiment, each gene is expressed at some certain level at each condition which is defined as the expression profile of the gene. More formally, a gene (say, x) is associated with a profile vector (V_x) composed of its expression values over all samples, such that $V_x = [x_1, x_2, ..., x_n]$, where n denotes the number of samples in the dataset. Thus, *intrinsic* similarity between genes x and y, is a measure defined on their profile vectors, V_x and V_y. A prevailing measure used for inferring similarity of two genes based on their gene profiles is Pearson's correlation coefficient [17]. Throughout our analysis, we employ absolute value of Pearson's correlation scores since both positive and negative correlations can play an important role in gene association. Recently, Ravasz et al [20], proposed the Topological Overlap Measure (TOM) which takes into a step in incorporating external information to infer similarity of two nodes in a biological network. This measure is considered as an improvement over the *intrinsic* similarity which amalgamates an additional external knowledge derived from the network topology (i.e., number of common neighbors).

2.2 Extrinsic Similarity

Extrinsic similarity of two attributes (i.e., genes) is defined over other attributes in the dataset [5]. In general, an extrinsic similarity between two attributes, i and j, can be defined as $ES_P(i,j) = \sum_{k \in P} f(i,j,k)$. Here, $f(i,j,k)$ denotes a function that signifies the association between attributes i and j, with respect to a third attribute k. P refers to the set of attributes that will contribute to

the *extrinsic* similarity of attributes i and j. As noted by Das et al [5], proper choice of the attribute set P and function f is crucial for the usefulness of the resulting *extrinsic* measure. Different choices of P and f will result in different similarity notions. Das et. al. [5] preferred to define an extrinsic dissimilarity measure based on the confidence of association rules.

In this work, we propose using Mutual Information of Information Theory to derive efficient extrinsic gene similarity measures. Our final goal is to surmise the similarity of two genes by the similarity of their relation with other genes. We believe that an extrinsic measure for microarray analysis has a twofold advantage over the use of intrinsic measures. First, it may reduce the impact of noise inherent in the dataset on the similarity analysis. It is well known that expression level of each gene is likely to fluctuate due to many sources of variability in a typical microarray analysis. Thus, the similarity deduced from expression levels of two genes is likely to be more error-prone than a similarity deduced from relative positions of these two genes with respect to many other genes. Second, it suits well with the biological hypothesis about genes and gene products acting in the form of complexes (i.e., groups) to accomplish certain tasks in the cell. As hypothesized, two gene products that belong to the same complex behave similarly with the members of this complex. Thus a similarity notion that is defined based on the relation of two genes with other genes can potentially capture the modular structure of the genomic interactions. Moreover, known modular structure of a biological system can be incorporated into the similarity analysis, by defining the P set by using this known structure.

To define proper extrinsic measures, we first need to determine the gene set, P, and the association function, f, that will constitute our measures. For the P set, we make use of the close proximity of each gene determined by an *intrinsic* similarity notion. We propose to use Conditional Mutual Information and Specific Mutual Information as the association functions.

Choice of Attribute Set (P): To derive an efficient *extrinsic* measure, we need an effective gene set that will be used to infer the *extrinsic* similarity of two genes. To compile such a set, we initially identified for each gene a set of genes that are intrinsically similar to that gene. We refer this as the neighborhood list of gene i and define it as $N_i = \{j | j \in G, |r_{ij}| > \kappa\}$, where G denotes the set of all genes in our dataset and $|r_{ij}|$ refers to the absolute value of the Pearson's correlation coefficient between genes i and j. We investigated the effect of the threshold parameter κ in our previous work and observed that size of the neighborhood lists can help us set this parameter [25]. Next, the attribute set P that will be used to infer similarity of two genes is designated as the intersection of their neighborhood lists (i.e., $P = N_i \cap N_j$). Using the common elements in two neighborhood lists, has two important implications. First, it significantly reduces the required number of calculations. Hence, instead of using the whole gene set (G), a smaller size set is taken into consideration for each similarity calculation. Secondly, it filters out irrelevant information which enhances the power of the *extrinsic* measure. Moreover, by using the *intrinsic* similarity to determine elements in set P, we take advantage of both *extrinsic* and *intrinsic*

properties. We believe this will be helpful in reducing the noisy inference that can be introduced into the similarity inference by using each technique separately.

Choice of Association Function (f): Das et al [5], proposed using *confidence* of association rules in an application on market basket dataset. We previously discussed Das's external dissimilarity measure and its applicability on gene expression datasets [25]. Our analysis showed that it is possible to improve their measure for the task of similar gene identification by using Mutual Independence of genes. We here propose using Conditional Mutual Information and Specific Mutual Information to derive effective extrinsic microarray measures.

To leverage Mutual Information of genes we used probability of occurrence and co-occurrence for genes in the neighborhood lists. Formally we define these probabilities as follows:

Definition 1: Probability of occurrence for a gene i, $P(i)$, is defined as the frequency of encountering that gene in all neighborhood lists. Note that genes with indistinct expression profiles will have higher frequency of occurrence values.

Definition 2: Probability of co-occurrence for two genes, i and j, $P(i,j)$, is defined as the frequency of encountering these two genes together in the neighborhood lists.

Conditional Mutual Information based Gene Similarity: Conditional Mutual Information between variables X and Y, $I(X,Y|C)$, signifies the quantity of information shared between X and Y when C is known. Formally, it is defined as, $I(X,Y|C) = H(X|C) - H(X|Y,C)$ where $H(X)$ signifies the Shannon entropy of the discrete random variable, X. For our calculations, H is defined for the occurrence of a gene in the neighborhood lists. Mutual information calculates the quantity of information shared between X and Y when C is given. $I(X,Y|C)$ is equal to zero iff X and Y are conditionally independent given C. Probabilities of occurrence and co-occurrence are used to calculate Conditional Mutual Information of two genes given neighborhood list of a third gene. A high Conditional Mutual Information between two genes implies that these two genes prefer to co-occur with the same set of genes when a third gene is known to be occurring in the neighborhood lists. If they are not co-occurring with the same set of genes, they will have a smaller Conditional Mutual Information. If two genes bring the same information to the Neighborhood Lists of many third parties, we expect these two genes to be regulated by the same mechanism. Based on this heuristic we define Conditional Mutual Information based Extrinsic Gene Similarity as follows:

$$CMI_P(i,j) = \sum_{k \in P} I(i,j|k=1) \tag{1}$$

This measure calculates the quantity of information shared by i and j, given that a third gene k is occurring in the neighborhood lists. As can be seen above, the final score is the sum of Conditional Mutual Information between i and j, with respect to all elements in set P. If i and j tend to share the same information, they will have a high CMI similarity value.

Specific Mutual Information based Gene Dissimilarity: The Specific Mutual Information is a measure of association commonly used in the Information Theory to infer mutual dependency. Specific Mutual Information of two variables, X and Y, given their joint distribution, $P(X,Y)$, and individual distributions, $P(X)$ and $P(Y)$, is defined as $\frac{P(X,Y)}{P(X)P(Y)}$, where $P(X,Y)$ is the observed value (O) for joint probability of events X and Y, whereas $P(X)P(Y)$ is its expected value (E). This test can be used to deduce the co-occurrence relation between two genes when their neighbors are considered. If Specific Mutual Information of two genes is 1, it can be concluded that these two genes are independent. In this context, being independent means genes i and j are randomly appearing together in the neighborhood lists. However, if two genes are not independent, occurrence of a gene in a neighborhood list makes it either less probable or more probable for the other gene to occur in that list. Based on this analysis we propose the following *extrinsic* measure to quantify dissimilarity of two genes (i and j).

$$SMI_P(i,j) = \sum_{k \in P} |\frac{P(i,k)}{P(i)P(k)} - \frac{P(j,k)}{P(j)P(k)}| \tag{2}$$

This definition ensures that two genes having the same co-occurrence relations with their common neighbors are closely related to each other (SMI value close to 0). Whereas two genes that have different independency relations with their common neighbors are dissimilar and associated with higher values of SMI.

We compare the proposed Mutual Information based extrinsic measures with the existing measures in the literature.

3 Domain Based Evaluation

'Similar' pairs identified according to different similarity/dissimilarity measures are evaluated based on Pairwise Semantic Similarity measure of Resnik [18]. This measure makes use of known annotations in the Gene Ontology (GO) database. GO is a controlled vocabulary designed to accumulate the result of all investigations in the area of genomic and biomedicine by providing a large database of known associations. Biological relevance of two genes can be quantified with respect to the significance of their shared GO annotations using the Semantic Similarity (SS) measure defined by Resnik [18]. Resnik's measure is preferred among other semantic similarity measures [11,13], since it has been shown to outperform the others and suit better to be used for GO analysis [21]. We calculated pairwise semantic similarity for the pairs labeled as similar according to different similarity/dissimilarity measures. We did not take into consideration relations among unannotated genes since there is not enough information to speculate about the biological concordance of these genes.

We then constructed association gene networks by linking the most similar gene pairs identified with respect to alternative similarity definitions. We obtained clusters of densely linked genes from these networks to study their efficacy

in understanding the molecular and biological processes. The obtained clusters are evaluated with an enrichment score that shows the statistical significance of the GO term homogeneity in a cluster. Details of this enrichment score can be found elsewhere [26].

4 Datasets and Pre-processing

For this study, we employ a well-studied cancer dataset and the Rosetta compendium yeast data (i.e., *Saccharomyces cerevisiae*) [10]. Our first dataset is composed of gene expression values of 62 colon tissue samples where the Affymetrix Hum6000 array with 6819 probes is used [1]. 42 of these are collected from colon adenocarcinoma patients and 20 of them are collected from normal colon tissue of the patients. Among all probes, 2000 were selected from 6817 by Alon et al according to the highest minimum intensity [1]. Our second dataset, Rosetta yeast data is obtained using a two-color cDNA microarray hybridization assay [10]. It is composed of 300 compendium experiments on the *Saccharomyces cerevisiae* organism. As suggested by the authors, we used the scale factor for our further analysis, which is defined as the standard deviation of log10(ratio)/[error of log10(ratio)] over all experiments. We perform thresholding, log transformation and normalization (quantile normalization) on these two datasets as suggested by our analysis. In addition to these, we further standardize datasets using a robust standardization method, median absolute deviation (MAD). Genes with zero MAD values implying that they are co-expressed at very similar levels across all of the samples are excluded from further analysis.

5 Experiments

Throughout this section, we discuss usability of extrinsic measures for microarray analysis. First, we present biological relevance of 'similar' gene pairs with different measures. We then linked these 'similar' genes to construct gene co-expression networks. Each of these networks are partitioned into its functional modules to study the effect of *extrinsic* similarity on the quality of information extracted from these networks.

5.1 Effect on Top 'Similar' Pairs

To choose a suitable κ threshold, there are two things that we should take into consideration. First, we want the neighborhood list of a gene to be composed only of genes that are within close proximity of that gene. Second, we do not prefer a set composed of a few genes since this would limit the power of inference based on common neighbors and increase the impact of noise on the final scores. Our previous study showed that average size of the neighborhood lists can guide us while setting the κ parameter [25]. Consequently, we set the κ threshold to 0.5 for the colon cancer dataset and 0.9 for the yeast data, which generates neighborhood lists of size 40 in average.

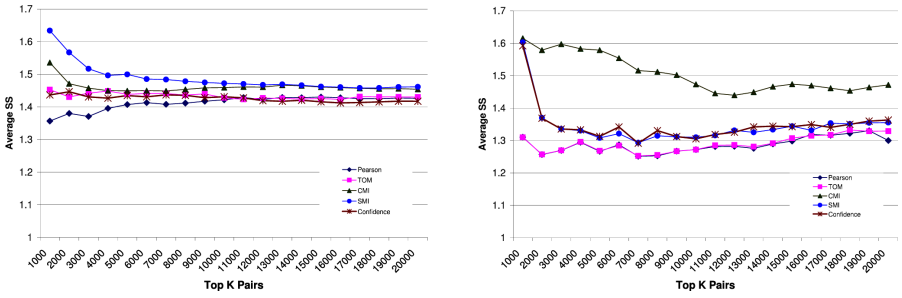

Fig. 1. Average semantic similarity (SS) is calculated for the top 'similar' pairs identified via alternativee measures from (a) Colon cancer and (b) Yeast microarray datasets. 1K represents the top 1000 pairs identified with each measure.

In our first experiment, we compare gene pairs that are labeled as 'similar' according to discussed measures. For each measure, gene pairs are sorted starting from the most 'similar' (or least 'dissimilar') one. We calculated semantic similarity of all annotated pairs and calculate the average semantic similarity for the whole set of gene pairs. Different number of top scoring pairs (varying between 1000 and 20000) are compared based on their average semantic similarity values. When we analyze the distribution of average semantic similarities, we observe that extrinsic measures outperform existing measures. For both datasets, a significant improvement in semantic similarity is observed.

For the colon cancer dataset, we observe that extrinsic measures significantly overlap with the biological relevance of genes. As can be seen in Figure 1a, the pairs identified with the SMI measure show greater biological relevance when compared to the pairs identified by other measures. For the top 1000 pairs, the improvement in the average semantic similarity score is up to 15%, when an extrinsic measure is used instead of an intrinsic one. Since semantic similarity calculations are based on the information content of each GO term which is in the logarithmic scale, this improvement is significant in real world, as our further analysis indicate. Although TOM measure is also able to improve the Pearson's correlation, this improvement is not as significant as our Mutual Information based extrinsic measures.

When we analyze the yeast dataset, we again observe that extrinsic measures identify biologically more relevant gene pairs. As can be seen in Figure 1b, the improvement is more significant (up to 22%) when top pairs obtained by CMI measure are compared to top pairs identified by the standard measure. Note that in contrast to colon cancer dataset, yeast data is obtained using cDNA assays. Our analysis show that extrinsic measures are effective for analysis of both cDNA and oligonucleotide arrays. As can be observed in this figure, TOM contributes even less to standard measure in this case, since mean r values are higher for this dataset.

Our analysis confirm that extrinsic measures better capture the biological relevance of two genes when compared to the standard intrinsic measure. We

believe their power can be attributed to two reasons: the noisy nature of microarray datasets and the functional modularity of genes. *Intrinsic* measures directly possess and reflect the noise inherent in the data since they are purely defined on the expression levels of genes under study. We also believe that since TOM measure is also dependent on the intrinsic measure in its definition, it would also be effected by the noise inherent in these datasets. The poor performance of TOM measure with respect to our extrinsic measures can be attributed to the fact that erroneous measurements will have a more drastic impact on any *intrinsic* or intrinsic based measure. On the other hand, *extrinsic* measures are dependent on more evidence since similarity of two genes are inferred from their relative positions with respect to a set of other genes. Hence, we expect the impact of erroneous measurements to be less severe on the *extrinsic* similarity measures. Our experimental results are also in accordance with this expectation where extrinsic measures produce biologically more relevant pairs. In addition, inferring similarity of two genes from a set of other genes can benefit from the group level interactions known to take place between genes and gene products when accomplishing certain cellular tasks [23].

5.2 Effect on Gene Networks

In this experiment, we constructed gene association networks by linking top similar pairs identified with each measure. Here, nodes represent genes, and two nodes are linked if the corresponding genes are 'similar' to each other. To keep the same size for all networks, we only used the top 0.01% of all gene pairs sorted with respect to a similarity/dissimilarity measure. Accordingly, colon cancer networks are composed of 12,438 edges and yeast networks are composed of 74,267 edges. Tightly connected subnetworks of a co-expression network can provide insight into the vital molecular and biochemical processes. Moreover, groups of genes that are densely linked in gene networks have been theorized to have similar cellular functions with great implications for gene annotation at a global scale [9,22,3]. Thus, we extracted and studied densely linked sub-networks of these networks.

To identify densely interacting subnetworks of these networks, we employ a graph partitioning algorithm, Graclus [8], that is shown to be effective in analyzing gene association networks [27]. This algorithm is effective in obtaining balanced-size clusters while minimizing the normalized cuts criterion. To our knowledge, no entirely reliable method exists for identifying correct number of partitions (i.e., k) in a network. That is why, we partitioned colon cancer networks into 100 clusters, and yeast networks into 200 clusters, to make sure reasonable size clusters will be generated at the end. In average 20 genes are located into each partition. Each partitioning is validated using the enrichment score p-value that signifies the homogeneity of each cluster in terms of its known GO annotations. Smaller p-values imply that the grouping is not random and is functionally more homogeneous. A cut-off parameter is used to differentiate significant groups from the insignificant ones. If a cluster is associated with a p-value greater than the cut-off, it is considered insignificant. We used the recommended cut-off of 0.05 for all our validations. The

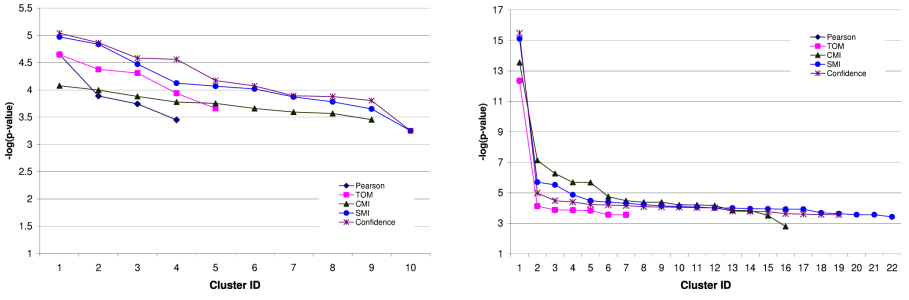

Fig. 2. P-value distribution of significant clusters extracted from (a)Colon Cancer and (b)Yeast gene networks. The y axis represents the $-log$ of the enrichment score of each corresponding cluster.

p-value distributions for the significant clusters extracted from various gene association networks are shown in Figure 2[1]. As can be observed from the figure, extrinsic similarity measures produce more number of clusters that are significantly enriched with Biological Process GO term annotations. For the colon cancer data, we are able to identify only 4 clusters that are functionally homogeneous when Pearson correlation is used. However, with the use of extrinsic measures this number increases to 10 for SMI and 9 for CMI. Similarly, for the yeast data, number of significant clusters and their significance scores drastically improve when extrinsic measures are used instead of the intrinsic measure. By using SMI measure instead of Pearson's correlation, number of significant clusters that can be deduced from the same data increased more than threefold. These results suggest that using extrinsic measure has a twofold enhancement for co-expression network analysis. First, these measures enhance functional homogeneity of clusters that can be identified with a standard measure as smaller p-values obtained for extrinsic based networks suggest. Also it enables identification of clusters that cannot be detected by standard measures, as evident from the increase in number of significant clusters.

6 Discussion

In this section, we investigate the usability of clusters extracted from different gene similarity networks by running a dataset specific analysis. For this part of our analysis, we analyze the colon cancer dataset which is composed of tumorous and non-tumorous tissues of the human colon and rectum. A more detailed analysis of the significant clusters obtained from the colon cancer data revealed that they can be very useful in understanding and treating the colorectal cancer. We discuss several of these clusters and their relation with colon cancer in the rest of this section.

By using the CMI measure, we obtained a cluster that is annotated with 'aldehyde dehydrogenase (NAD) activity'. Previous studies showed that activity of aldehyde dehydrogenase was measured in primary and metastatic human colonic

[1] Biological Process GO terms are used for this analysis.

adenocarcinomas [14]. We also identified clusters annotated with 'phospholipase activity' by employing the CMI measure. It has been shown that Phospholipase D (PLD) has a possible impact on carcinogenesis and its progression [16]. Another cluster obtained with CMI measure is annotated with 'NF-kappaB binding'. NF-kappaB pathway is shown to be taking part in the regulation of Inhibitors of apoptosis (IAP) family in human colon cancers [28]. Identification of clusters that are known to be related to colon cancer is vital for developing new therapeutic targets and identifying potential tumor markers for colorectal cancer. However, we cannot identify such clusters via standard analysis of the same dataset.

From the *SMI* network, we extracted a cluster that is composed of genes associated with the GO term 'cytoskeleton-dependent intracellular transport'. Recent evidence indicates that the interaction of a tumor suppressor gene (APC) with the cytoskeleton might contribute to colorectal tumor initiation and progression [15]. That is why, we believe that locating these genes together in a cluster is triggered by the role they play in colon cancer tumorigenesis. Unfortunately, it is still unknown that how APC interacts with the cytoskeleton and how their interaction plays a role in the formation of colorectal tumors [15]. We believe that once functionally coherent clusters are identified, relations between these clusters can be used to reveal function level interactions vital for understanding the cause of some diseases.

7 Conclusion

In this paper, we have introduced the notion of Mutual Information of genes based on their relations with other genes. We have presented two *extrinsic* measures for microarray analysis based on Conditional Mutual Information and Specific Mutual Information. We also discussed a method to employ a previously suggested extrinsic measure for market basket datasets in microarray analysis. We have investigated the efficacy of the proposed measures and run thorough analysis to compare them with standard similarity measures. Our experimental results prove that by using the *extrinsic* measures, it is possible to identify gene pairs that are biologically more relevant. In addition, association networks generated based with these measures are shown to be more informative and useful for further analysis. These results suggest that different similarity notions can reveal different aspects of the same dataset. Previously, we have studied different ensemble techniques to improve clustering results on a scale-free protein interaction network [2]. In the future, we plan to investigate an ensemble approach for integrating different aspects of a dataset captured by different similarity measures.

References

1. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. 96, 6745–6750 (1999)
2. Asur, S., Ucar, D., Parthasarathy, S.: An ensemble framework for clustering protein-protein interaction networks. In: Proc. 15th Annual Int'l Conference on Intelligent Systems for Molecular Biology (ISMB) (2007)

3. Bader, G., Hogue, C.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4(2) (2003)
4. Carter, S., Brechbhler, C., Griffin, M., Bond, A.T.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20(14), 2242–2250 (2004)
5. Das, G., Mannila, H., Ronkainen, P.: Similarity of attributes by external probes. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD 1998), pp. 23–29 (1998)
6. Das, G., Mannila, H., Ronkainen, P.: Similarity of attributes by external probes. Report C-1997-66, University of Helsinki, Department of Computer Science (October 1997)
7. Datta, S., Datta, S.: Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics 7(397) (2006)
8. Dhillon, I., Guan, Y., Kulis, B.: Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1944–1957 (2007)
9. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95(25), 14863–14868 (1998)
10. Hughes, T., et al.: Functional discovery via a compendium of expression profiles. Cell, 102 (2000)
11. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. Int'l Conf. Research in Computational Linguistics, ROCKLING X (1997)
12. Lee, H., Hsu, A., Sajdak, J., Qin, J., Pavlidis, P.: Coexpression analysis of human genes across many microarray data sets. Genome Research 14, 1085–1094 (2004)
13. Lin, D.: An information-theoretic definition of similarity. In: Proc. 15th Int'l Conf. Machine Learning (1998)
14. Marselos, M., Michalopoulos, G.: Changes in the pattern of aldehyde dehydrogenase activity in primary and metastatic adenocarcinomas of the human colon. Cancer letters 34(1), 27–37 (1987)
15. Näthke, I.: Cytoskeleton out of the cupboard: colon cancer and cytoskeletal changes induced by loss of apc. Nature Reviews Cancer 6, 967–974 (2006)
16. Oshimoto, H., Okamura, S., Yoshida, M., Mori, M.: Increased Activity and Expression of Phospholipase D2 in Human Colorectal Cancer
17. Ostel, B.: Statistics in research basic concepts and techniques for research workers. Iowa State University Press, Ames (1963)
18. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 448–453 (1995)
19. Palmer, C., Faloutsos, C.: Electricity based external similarity of categorical attributes. In: Whang, K.-Y., Jeon, J., Shim, K., Srivastava, J. (eds.) PAKDD 2003. LNCS, vol. 2637. Springer, Heidelberg (2003)
20. Ravasz, E., et al.: Hierarchical organization of modularity in metabolic networks. Science 297(5586), 1551–1555 (2002)
21. Sevilla, J.L., et al.: Correlation between gene expression and go semantic similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2(4) (2005)
22. Snel, B., Bork, P., Huynen, M.: The identification of functional modules from the genomic association of genes. Proc. Natl. Acad. Sci. 99, 5890–5895 (2002)

23. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. PNAS 100(21) (2003)
24. Stuart, J., Segal, E., Koller, D., Kim, S.: A gene coexpression network for global discovery of conserved genetic modules. Science 302(5643), 249–255 (2003)
25. Ucar, D., Altiparmak, F., Ferhatosmanoglu, H., Parthasarathy, S.: Investigating the use of extrinsic similarity measures for microarray analysis. In: Proceedings of the BIOKDD workshop at the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD) (2007)
26. Ucar, D., Asur, S., Catalyurek, U., Parthasarathy, S.: Improving Functional Modularity in Protein-Protein Interactions Graphs Using Hub-Induced Subgraphs. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS, vol. 4213, pp. 371–382. Springer, Heidelberg (2006)
27. Ucar, D., Neuhaus, I., Ross-MacDonald, P., Tilford, C., Parthasarathy, S., Siemers, N., Ji, R.: Construction of a reference gene association network from multiple profiling data: application to data analysis. Bioinformatics 23(20), 2716 (2007)
28. Wang, Q., Wang, X., Evers, B.: Induction of cIAP-2 in human colon cancer cells through PKC/NF-B. J. Biol. Chem. 278, 51091–51099 (2003)
29. Zhang, B., Horvath, S.: A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4(1) (2005)