

# Investigating the use of Extrinsic Similarity Measures for Microarray Analysis \*

D. Ucar, F. Altıparmak, H. Ferhatosmanoglu and S. Parthasarathy<sup>†</sup>  
Department of Computer Science and Engineering  
The Ohio State University  
Columbus, Ohio  
Contact : srini@cse.ohio-state.edu

## ABSTRACT

Genes behaving similarly over changing conditions are believed to be part of the same functional module. Identifying functional modules of genes plays an important role in understanding gene regulatory behavior as well as in facilitating function prediction of unknown genes. Subsequently, determining ‘similar’ gene pairs or groups based on their gene expression profiles is an important task towards extracting modules from microarray datasets. A prevailing technique is to use a linear similarity measure like Pearson’s correlation coefficient or Euclidean distance, to find similar gene pairs. 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 gene similarity 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 to enhance the quality of gene networks constructed from similar gene pairs by reducing spurious edges and introducing missing edges between network nodes.

## Categories and Subject Descriptors

H.2.8 [Database Applications]: Data Mining

## Keywords

Bioinformatics, Microarray analysis, Extrinsic similarity

## 1. INTRODUCTION AND RELATED WORK

\*This work is supported in part by DOE Early Career Principal Investigator Award No. DE-FG02-04ER25611 and NSF CAREER Grant IIS-0347662.

<sup>†</sup>To whom correspondence should be addressed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BIOKDD’07 August 12, 2007, San Jose, California, USA.

Copyright 2007 ACM 978-1-59593-839-8/07/0008 ...\$5.00.

Due to advances in technology (e.g., oligonucleotide microarray chips), scientists are now able to accumulate a wealth of information on the expression of genes during the life cycle of an organism. Such datasets provide vital information that can be used to gain insight into diverse biological questions. To analyze and mine these datasets for potential useful information, various techniques and ideas have been proposed. Of particular interest to many scientists is the problem of identifying gene groups that have similar expression patterns over various samples, known as co-expressed genes. Genes with similar cellular functions have been theorized to behave similarly over different conditions [10]. Thus, obtaining groups of similar genes is fundamental to understanding the molecular and biochemical processes that sustain the physiological state of the cell [23].

There has been a growing interest in representing co-expressed genes as an association network to explore the system-level functionality of genes [25, 6]. Here, nodes represent genes and two nodes are linked if the corresponding genes are significantly co-expressed (correlated) across the samples. Earlier approaches have used expression levels of two genes over all samples to surmise their correlation. However, this similarity notion does not necessarily imply that genes are functionally related. Given the noise inherent in microarray datasets, it is our hypothesis that intrinsic similarity measures are not adequate to distinguish accidentally regulated genes from those that are biologically motivated. We argue that since any given gene is likely to fluctuate in its measured expression level due to many possible sources of error, a similarity based on two genes’ measurements is more error-prone than using relative positions of many genes as a reference to deduce the same information. In addition, gene products act as complexes to accomplish certain cellular level tasks [22], which is potentially suitable to infer two gene’s similarity via their relations with other genes. Thus, we propose and investigate the use of extrinsic similarity measures to induce gene similarity.

The use of extrinsic measures and their advantages have been previously studied for various data mining problems [8, 9]. Das et al [8], 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 [18], 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 [19], proposed the Topological Overlap Measure (TOM), which is one of the few to use extrinsic

properties along with the intrinsic ones. Their measure infers similarity of two nodes in a biochemical network in terms of their pairwise similarity as well as the number of common neighbors they share.

In this paper, we introduce a methodology for the application of extrinsic similarity measures on microarray datasets. We propose two different extrinsic measures motivated by the notion of *mutual independence analysis*. The proposed similarity measures are evaluated on two well-studied cancer microarray datasets [1, 4]. In order to quantify the biological concordance of different similarity notions, we employ domain based validation metrics. We find that extrinsically similar gene pairs 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 their usability for gene function inference, we construct association networks from ‘similar’ gene pairs identified by different measures. Our analyses show that association networks constructed based on our extrinsic measures contain less spurious and more biologically verified edges compared to their counterparts generated using other measures. We obtain densely connected clusters of genes from these networks to study their usability in understanding the molecular and biological processes that sustain health or cause cancer. We find that clusters extracted from the extrinsically similar gene networks show evidence of cancer related pathways and functional modules such as signal transduction pathway, apoptosis etc.

To summarize, our main contributions in this study are:

- Introducing the notion of *mutual independence* of two genes based on their associations with other genes
- Proposing two extrinsic similarity measures suitable for microarray analysis motivated by the *mutual independence* analysis
- Investigating and demonstrating the efficacy of using extrinsic measures in inferring pairwise gene similarities, constructing gene networks and clustering genes

## 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 measures. 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  $i$  and  $j$ ’s similarity.

Previous studies have shown the usability of external similarity measures in other domains [8, 9]. To our knowledge, usability of *extrinsic* similarity measures have not been investigated for identifying ‘similar’ genes. A prevailing method to infer similarity of two genes from their expression patterns is to use a linear *intrinsic* similarity (e.g. Euclidean distance, Pearson’s correlation coefficient) measure. We discuss *intrinsic* similarity measures next.

### 2.1 Intrinsic Measure

*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 these genes’ expression levels over all samples.

In a typical microarray experiment, each gene is expressed at some certain level at each condition which is defined as the gene’s expression profile. More formally, a gene (say,  $x$ ) is associated with a profile vector ( $V_x$ ) composed of its expression values over all samples, such that  $V_x = [x_1, x_2, \dots, x_n]$ , 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  $V_y$ .

The most commonly used and accepted measure in the literature for the task at hand is the Pearson’s correlation coefficient. This is defined as [16]:

$$r_{xy} = \frac{\sum_{i=1}^n (V_x^i - \bar{V}_x)(V_y^i - \bar{V}_y)}{\sqrt{\sum_{i=1}^n (V_x^i - \bar{V}_x)^2 \sum_{i=1}^n (V_y^i - \bar{V}_y)^2}} \quad (1)$$

where  $\bar{V}_x$  and  $\bar{V}_y$  are the profile averages. Here,  $V_x^i$  represents the  $i^{th}$  entry of the vector  $V_x$ . According to this definition, genes which are positively (or negatively) correlated have a value close to 1 (or -1) whereas dissimilar gene pairs have values close to 0. Absolute value of Pearson’s correlation scores is used in this study since both positive and negative correlations can play an important role in gene association.

### 2.2 Extrinsic Measures

*Extrinsic* similarity of two attributes (i.e., genes) is defined over other attributes in the dataset. Before defining its specifics, a general definition of an *extrinsic* measure is as follows [8]:

$$ES_P(i, j) = \sum_{k \in P} |f(i, k) - f(j, k)| \quad (2)$$

Here,  $f(i, k)$  denotes a function that signifies association between  $i$  and  $k$ .  $P$  refers to the set of attributes that will contribute to the *extrinsic* similarity calculation of attributes  $i$  and  $j$ .

As noted by Das et al [8], proper choice of the attribute set  $P$  and function  $f$  is crucial for the usefulness of the resulting *extrinsic* measure. Different choices will result in different similarity notions. In the following section we will discuss a methodology to derive effectual *extrinsic* similarity measures to be used in inferring gene similarity.

### 2.3 Proposed Methodology

Our goal in developing an *extrinsic* similarity for microarray analysis is to surmise the similarity of two genes by the similarity of their relation with other genes. We believe that use of an extrinsic measure for microarray analysis has a twofold advantage over the use of intrinsic measures. First, it reduces the impact of noise inherent in the dataset on the similarity inference since more evidence are taken into consideration per inference. Second, it suits well with the biological hypothesis that genes act as complexes to accomplish certain tasks in the cell. As hypothesized, two genes behaving similarly with the elements of a gene complex, presumably belongs to that complex and share their functionality. Thus defining two genes’ similarity by taking into consideration their relation with other genes can potentially benefit from the modular structure of the genomic interactions.

To define a proper measure, we first need to determine over which set of genes,  $P$ , and using which association function,  $f$ , *extrinsic* similarity of two genes should be defined.

Here, we investigate the use of close proximity of genes according to *intrinsic* notions when choosing a proper set  $P$ . In addition, two functions based on *mutual independence analysis* from the Information Theory are evaluated. We compare the proposed similarity measures with the currently available techniques described in Section 3, as well as the most popular *intrinsic* measure (i.e., Pearson’s correlation coefficient).

### 2.3.1 Choice of Attribute Set ( $P$ )

To derive an efficient *extrinsic* measure for microarray analysis, we first need to identify a gene set,  $P$ , that will be used to infer the *extrinsic* similarity of two genes. For this purpose, we use the group of genes that are similar to both of the genes under question. Thus, initially for each gene we identify a set of genes that are intrinsically similar to that gene (i.e., the gene’s close neighbors). We refer this as a gene’s neighborhood list ( $N_i$ ) and define it as follows:

$$N_i = \{j | j \in G, |r_{ij}| > \kappa\} \quad (3)$$

Here,  $G$  denotes the set of all genes in our dataset and  $|r_{ij}|$  refers to the absolute value of the Pearson’s correlation coefficient of genes  $i$  and  $j$ . Effect of the threshold parameter  $\kappa$ , on the *extrinsic* measures and guidance of the size of neighborhood lists to set this parameter is discussed in Section 6<sup>1</sup>. Next, the attribute set  $P$  that will be used to infer two genes’ similarity is designated as the intersection of their neighborhood lists (i.e.,  $P = N_i \cap N_j$ ). Using common neighbors of two genes as the set of attributes ( $P$ ) has two important implications. First, it significantly reduces the required number of calculations. Thus, instead of using the whole gene set ( $G$ ), a smaller size set is taken into consideration. Secondly, it filters out irrelevant information which improves the success of the *extrinsic* measure. By using the *intrinsic* similarity to determine elements in set  $P$ , we take advantage of both *extrinsic* and *intrinsic* properties. Our hypothesis is that this helps to reduce the noisy inference that can be introduced into the similarity inference by using these measures separately. It is noteworthy that an *extrinsic* measure can be easily expandable to other groups of related genes. For instance, one can prefer using an attribute set containing genes mapped to close chromosomal locations with two genes whose similarity is under investigation.

### 2.3.2 Choice of Association Function ( $f$ )

After establishing the notion of an *extrinsic* similarity, and defining the set  $P$ , the next step is to determine which association function ( $f$ ) to use for our calculations. Das et al [8], proposed using the *confidence* of association rules in an application on market basket dataset. Their approach and its applicability on gene expression datasets will be discussed in details in Section 3. We propose using two appropriate functions that are motivated by the *mutual independence analysis*. We leverage mutual independence of two genes by analyzing their frequency of occurrence and co-occurrence in the neighborhood lists.

Before defining mutual dependency of two genes, first, we explore three possible type of relations between any two genes motivated by Das et al [8]. Accordingly, two genes can either be, *complementary*, *independent* or *correlated*. If two genes are *complementary*, then they do not to co-occur

<sup>1</sup>Our analysis indicated that relatively loose values produce more useful *extrinsic* measures.

in the neighborhood lists. If they are *independent*, neighbors of gene  $i$  are neighbors of gene  $j$  with the same probability as the genes that are not neighbors of gene  $i$ . And if they are *correlated*, neighbors of gene  $i$  are also neighbors of gene  $j$ . These concepts are formally defined using neighborhood lists as follows:

**Definition 1:** *Frequency of occurrence* for a gene  $i$ ,  $P(i)$ , is defined as the frequency of encountering that gene in all neighborhood lists. Since Pearson’s correlation coefficient is a symmetric measure a gene has as many neighbors as the number of times it occurs in all neighborhood lists. Thus, frequency of a gene’s occurrence can be simplified to the following:

$$P(i) = \frac{|N_i|}{|G|} \quad (4)$$

where ‘ $|u|$ ’ denotes the number of elements (cardinality) in its argument. Note that *frequency of occurrence* is an indication of the discriminatory nature of a gene’s expression profile. Genes with indistinct expression profiles such as the housekeeping genes will have higher values of *frequency of occurrence*.

**Definition 2:** *Frequency of co-occurrence* for genes  $i$  and  $j$ ,  $P(i, j)$ , is defined as the frequency of encountering these two genes together in the neighborhood lists. More formally, based on the symmetric Pearson’s measure,  $P(i, j)$  can be defined as follows:

$$P(i, j) = \frac{|\{a | a \in G, i \in N_a, j \in N_a\}|}{|G|} \quad (5)$$

By itself high *frequency of co-occurrence* does not imply that two genes are *correlated*. In order to conclude that two genes are not randomly co-occurring (*independent*) but there is a biological trigger behind their co-occurrence (*correlated*), we need to test if one gene’s *frequency of occurrence* is helpful in predicting that of the other gene which is a notion known as mutual independence. Note that, in this context, independence of two genes implies that occurrence of a gene in a neighborhood list makes it neither more nor less probable for the other gene to occur in that list. Thus, mutual independence of two genes only holds when  $P(i, j) = P(i)P(j)$ . We propose using two different independence tests to leverage *mutual dependency* of two genes.

#### Specific Mutual Information Measure:

The Specific Mutual Information (*smi*) is a measure of association commonly used in the Information Theory to infer mutual dependency. *Smi* of two variables,  $X$  and  $Y$ , given their joint distribution,  $P(X, Y)$ , and individual distributions,  $P(X)$  and  $P(Y)$ , is defined as follows:

$$I(X, Y) = \frac{O}{E} = \frac{P(X, Y)}{P(X)P(Y)} \quad (6)$$

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 type of relation between two genes. If their *smi* value is 1, it can be concluded that these two genes are *independent*. On the other hand, a value greater than 1 implies being *correlated* and a value smaller than 1 implies being *complementary*.

If two genes have similar relations with their common neighbors, it is reasonable to conclude that they are similar. Based on this analysis and the notion of specific mutual information, we propose the following *extrinsic* measure to quantify dissimilarity of two genes ( $i$  and  $j$ ).

$$smi_P(i, j) = \frac{\sum_{k \in P} \left| \frac{P(i,k)}{P(i)P(k)} - \frac{P(j,k)}{P(j)P(k)} \right|}{|P|} \quad (7)$$

This definition ensures that two genes having similar relations (i.e., *complementary*, *correlated* or *independent*) with their common neighbors are closely related to each other (*smi* value close to 0). Whereas two genes that have different relations with their common neighbors are dissimilar and associated with higher values of *smi*. Note that, the *smi* measure is normalized by dividing by the size of the attribute set  $P$ .

### Chi-Square Based Measure:

Pearson’s chi-square test is another method to assess *mutual dependency* of two events. Formally, it is defined as follows:

$$chi(X, Y) = \frac{(O - E)^2}{E} = \frac{(P(X, Y) - P(X)P(Y))^2}{P(X)P(Y)} \quad (8)$$

This test tells us how far the observed value deviates from the expected value under the assumption of independence.

According to this definition, two genes will have zero *chi* value if they are *independent*. They will have higher *chi* values otherwise. We employ a signed version of this test to surmise the type of relation between two genes. Given this, external dissimilarity of two genes based on the chi-square analysis,  $chi_P(i, j)$ , is defined as follows:

$$\frac{\sum_{k \in P} \left| \frac{s_{ik}(P(i,k) - P(i)P(k))^2}{P(i)P(k)} - \frac{s_{jk}(P(j,k) - P(j)P(k))^2}{P(j)P(k)} \right|}{|P|} \quad (9)$$

where  $s_{ab}$  denotes the sign of the term  $P(a, b) - P(a)P(b)$ . Note that signs are included into the measure to differentiate a *correlated* pair from a *complementary* one. Similar to the *smi* measure, two genes that have similar relations with their common neighbors will have smaller *chi* values whereas two genes that have dissimilar relations with their common neighbors will have higher values<sup>2</sup>. *Chi* measure is also normalized by dividing by the size of the attribute set.

## 3. PREVIOUS WORK

### 3.1 Topological Overlap Measure

Recently, Ravasz et al [19], proposed the Topological Overlap Measure (TOM) which takes into a step in using *extrinsic* measures to infer similarity between two nodes of 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). According to their definition, two nodes have high topological overlap if they are connected to roughly the same group of nodes. More formally, TOM of two genes  $i$  and  $j$  can be expressed as follows:

$$TOM(i, j) = \frac{|N_i \cap N_j| + r_{ij}}{\min\{|N_i|, |N_j|\} + 1 - r_{ij}} \quad (10)$$

<sup>2</sup>Only the positive information is considered for the chi square test.

where  $r_{ij}$  is the pairwise similarity between these two genes. The inclusion of the *intrinsic* similarity ( $r_{ij}$ ), into this definition, makes TOM measure explicitly dependent on the *intrinsic* similarity of two nodes in question. Drawbacks of this dependency will be discussed in Section 6.

### 3.2 Confidence of Association Rules

Das et al [8, 9], previously studied the *extrinsic* similarity of attributes in a market basket dataset where *confidence* of association rules are used as the association function,  $f$ . In a market-basket problem, each customer fills their market basket with a subset of large number of items (e.g., bread, milk). Such datasets are mined for association rules of the form  $(X_1, \dots, X_n \Rightarrow Y)$  to identify the relation between items. The *confidence* of an association rule is defined as the frequency of encountering the head of the rule  $(X_1, \dots, X_n)$  among all the groups containing the body  $(Y)$ . Das et al [8], proposed using the *confidence* of association rules as the association function  $f$ . Thus, their proposed *extrinsic* similarity measure reduces to the following.

$$ES_P(A, B) = \sum_{D \in P} |conf(A \Rightarrow D) - conf(B \Rightarrow D)| \quad (11)$$

where  $conf(A \Rightarrow D)$  is defined as  $\frac{P(A, D)}{P(A)}$ .

For the task at hand, an analogy to a market basket is a neighborhood list. Accordingly, we use the *frequency of occurrence* ( $P(i)$ ) and the *frequency of co-occurrence* ( $P(i, j)$ ) to derive a corresponding *confidence* based *extrinsic* measure suitable for microarray analysis. We again normalize this measure by dividing it by the size of the set  $P$ .

We compare the newly proposed *extrinsic* similarity measures (*smi* and *chi*) with the existing ideas in the literature (i.e., TOM and *confidence*) as well as the most commonly used and accepted intrinsic measure for microarray analysis, namely the Pearson’s correlation coefficient.

## 4. DOMAIN BASED EVALUATION

‘Similar’ pairs identified according to different similarity measures are evaluated based on the Pairwise Semantic Similarity measure of Resnik [17]. 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 [17]. Resnik’s measure is preferred among other semantic similarity measures [11, 12], since it has been shown to outperform the others and suit better for use in GO [20].

Pairwise *SS* scores are used to infer functional relevance of probe pairs. For this purpose, we plot *SS* values for all annotated pairs of the arrays under study and observe that for both arrays *SS* values roughly follow normal distributions. We believe that to reduce the impact of missing information in GO database, it is desirable to limit ourselves to upper and lower tail of the distribution for inference. Accordingly, we label each pair as a ‘TP’ if their *SS* score is greater than the 95<sup>th</sup> percentile of all pairwise *SS* values. Similarly, a pair is accepted as a ‘FP’ when their *SS* value is smaller than the 5<sup>th</sup> percentile of the distribution. We run an analysis to test the effect of using greater percentile cut-offs on the overall

results which is presented in the Experiments section. We want to note that, not every gene pair will be classified as a ‘TP’ or a ‘FP’ using this labeling methodology. A pair that is composed of at least one unannotated gene is not labeled since there is not enough information to conclude about the biological concordance of these two genes. In addition, a gene pair with an  $SS$  score between the percentile cut-offs is not labeled since considering it as a ‘TP’ or a ‘FP’ pair is a matter of specifying the granularity of biological similarity.

Pairs extracted by using different similarity notions are accumulated into association networks. We define the Cluster-wise Positive Predictive Value measure ( $CPPV$ ) to evaluate the biological quality of the dense regions extracted from these clusters.  $CPPV$  of a cluster, (say,  $C_i$ ), is defined as  $CPPV_i = \frac{|TP_i|}{|TP_i| + |FP_i|}$ . Here,  $TP_i$  and  $FP_i$  denote the set of ‘TP’ and ‘FP’ pairs in that cluster. Our calculations are based on every possible gene pair in a cluster. Higher values of  $CPPV$  imply that the cluster is enriched in ‘TP’ pairs. On the contrary, lower values indicate that the cluster is composed of biologically dissimilar genes.

## 5. DATASETS AND PRE-PROCESSING

For this study, we employ two well-studied cancer datasets. 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]. Second dataset is composed of 86 lung adenocarcinoma and 10 normal samples which is analyzed by the Affymetrix HuGene FL array [4]. Beer et al [4] trimmed the dataset of genes expressed at extremely low levels resulting in 4966 probes for investigation.

Initially, we consider 2000 and 4966 probes for colon and lung adenocarcinoma datasets respectively. 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. After pre-processing 1578 genes for colon cancer and 4228 genes for lung cancer datasets are examined.

## 6. EXPERIMENTS

We discuss the usability of external similarity measures as a way of identifying similar genes throughout this section. First, we give results for biological relevance of gene pairs that are identified as ‘similar’ with different measures. Then, co-expression networks generated from these ‘similar’ pairs are analyzed for biological soundness. Finally, genes in each of these networks are clustered to study the effect of *extrinsic* similarity on the quality of gene clustering.

### 6.1 Setting the $\kappa$ parameter

Before comparing newly proposed measures with the existing ones, we first investigate the effect of  $\kappa$  parameter on the neighborhood lists. To choose a suitable  $\kappa$  threshold, there are two things that we should take into consideration. First, we want a gene’s neighborhood list to be composed

only of genes that are within close proximity of that gene. Second, it is not desirable to have a set that is only composed of a few genes since this would limit the power of inference based on common neighbors. Accordingly, we vary  $\kappa$  parameter between 0.3 and 0.9 and observe the average size of neighborhood lists for each of these values. As expected, for both datasets, smaller values of  $\kappa$  resulted in lists bigger in size with many dissimilar genes. On the other hand, higher  $\kappa$  values resulted in very small size lists which are very restrictive to draw any conclusions. Given that observation, we believe that average size of the neighborhood lists can guide us for setting the  $\kappa$  parameter. Consequently, a reasonable  $\kappa$  threshold value, 0.5, is determined for both datasets where neighborhood lists contain around 40 genes. We test the effect of  $\kappa$  parameter on the efficacy of extrinsic similarity measures in the next section.

### 6.2 Effect on Top ‘Similar’ Pairs

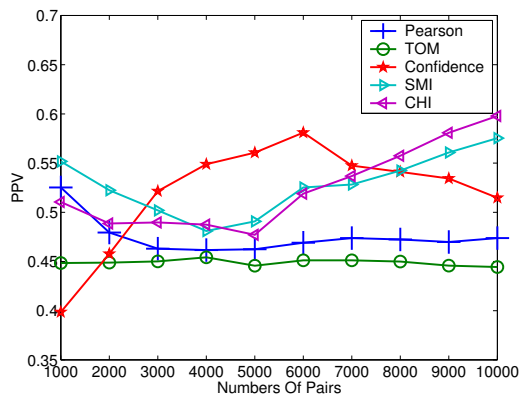
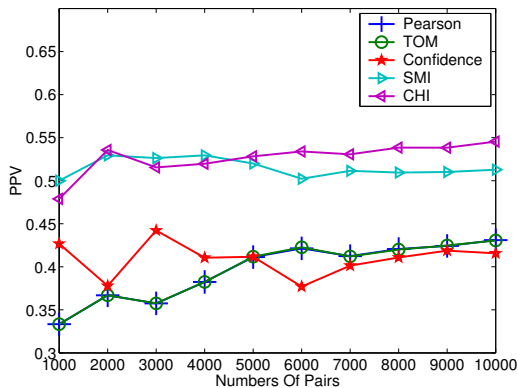
In the first experiment, we compare gene pairs that are labeled as ‘similar’ according to the discussed measures. For each measure, gene pairs are sorted starting from the most ‘similar’ one. These pairs are labeled as ‘TP’ or ‘FP’s based on their semantic similarity scores<sup>3</sup>. Different number of top scoring pairs (varying between 1000 and 10000) are compared based on the number of ‘FP’ and ‘TP’s among them (depicted in the below table)<sup>4</sup>.

	Pearson		TOM		Confidence		Smi		Chi	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1000	24	48	24	48	35	47	34	34	47	25
2000	51	88	50	87	65	107	72	64	75	65
3000	74	133	75	134	111	140	111	99	100	94
4000	109	176	109	177	140	201	153	136	132	122
5000	153	219	154	220	170	243	195	180	168	150
6000	193	265	194	265	187	309	224	222	204	178
7000	226	322	225	321	236	352	268	256	242	214
8000	265	365	265	366	265	380	296	285	294	252
9000	297	403	299	405	304	422	328	315	330	283
10000	337	445	338	447	330	464	361	343	366	305

In each case, *smi* and *chi* measures produce more ‘TP’ pairs compared to the TOM and the Pearson measures. In addition, *smi* and *chi* measures also generate significantly less ‘FP’ pairs in comparison to other measures. These results confirm that *smi* and *chi* measures better capture the biological relevance of two genes than the available measures in the literature. This improvement 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. As high values of ‘FP’ counts for the Pearson measure imply, erroneous measurements have a drastic impact on this *intrinsic* measure. It is notable that despite taking into consideration an *extrinsic* feature, TOM is similarly affected by the noise inherent in the dataset. This result shows that TOM is mainly dominated by the *intrinsic* factor in its definition. On the other hand, *extrinsic* measures are dependent on more evidence where mutual independence is inferred from all neighborhood lists. As a result, impact of erroneous measurements expected to be less severe on the *extrinsic* similarity measures. Our experimental results

<sup>3</sup>Not every gene pair can be labeled as a ‘TP’ or a ‘FP’.

<sup>4</sup>Colon cancer dataset follows similar trends.



**Figure 1: PPV of the top ‘similar’ pairs identified from our experimental datasets ( $\kappa = 0.5$ ): (a) Colon Cancer (b) Lung Cancer.**

are also in accordance with this expectation where extrinsic measures generate less ‘FP’ pairs. In addition, inferring two genes’ similarity from a set of other genes can benefit from the group level interactions known to take place between gene products when accomplishing certain cellular tasks [22]. High ‘TP’ counts associated with *extrinsic* measures are also in accordance with this biological premise. Poor results of the *confidence* measure indicate that choosing a proper association function  $f$  is also vital when defining an *extrinsic* similarity measure.

We also evaluate the Positive Predictive Value ( $PPV = \frac{TP}{TP+FP}$ ) of these pairs on both datasets (presented in Figures 1a-b). As can be seen, for both datasets, *smi* and *chi* measures constantly have higher PPVs when same number of similar pairs are analyzed. For colon cancer dataset, when compared to Pearson correlation, on average *smi* and *chi* measures improved the PPVs 30% and 34% respectively. For the lung cancer dataset, *smi* and *chi* measures again produce higher PPVs (on average an increase by 11% and 10%) than the Pearson measure. On the other hand, for both datasets TOM does equivalently or poorly when compared to the Pearson measure. Our analyzes also show that *confidence* is not a robust similarity measure due to the fact that it only considers two genes co-occurrence without analyzing their independence. As a result, it is impossible to tell if two genes are *correlated*, *independent* or *complementary* based on their *confidence* scores. This leads to incorrect conclusions about two gene’s similarity as implied by the fluctuating pattern of the *confidence* measure in Figures 2a-b. These results also suggest that *mutual independence* based analysis generates more robust external similarity measures when compared to the *confidence* based analysis.

In the next experiment, we evaluate the PPV of top pairs for different values of  $\kappa$ . We re-run our analysis on colon cancer dataset for different  $\kappa$  thresholds (depicted in Figure 1a ( $\kappa = 0.5$ ) and Figures 2a-b ( $\kappa = 0.45$  and  $\kappa = 0.55$ )). In each case, pairs identified by our *extrinsic* measures have systematically higher PPVs than the other measures. As in the previous cases, confidence measure produces inconstant PPVs and TOM does equally well with the Pearson correlation. These results show that although  $\kappa$  threshold has an impact on the efficacy of extrinsic measures, within a reasonable range (can be chosen by considering the average size of neighborhood lists) of  $\kappa$  values, *extrinsic* measures would be better alternatives to *intrinsic* measures.

### 6.3 Effect on Similarity Networks

In this experiment, we construct association networks by connecting the top scoring gene pairs identified by each measure. To keep the same size for all networks, we only used the top 0.01% of ‘similar’ gene pairs in each case. Accordingly, from the colon cancer dataset a network of 12,438 edges and from the lung cancer dataset a network composed of 89,359 edges are constructed. To investigate the biological quality of these networks, we identify the ‘TP’ and ‘FP’ pairs (i.e., edges) in each network. Here, we again observe that the advantage of using extrinsic measures is two-fold as shown in the below table. First, they reduce the number of ‘FP’ edges and secondly they increase the number of ‘TP’ edges. As a result, for the colon cancer dataset PPV is increased by 18% and 20% when *smi* and *chi* measures are employed respectively. For the lung cancer dataset, both measures improve the PPV by 15 % when compared to the Pearson measure. Networks identified using the TOM, do not have higher PPVs than the networks generated by the Pearson correlation, implying that TOM fails to contribute to a standard intrinsic similarity measure. These results suggest that extrinsic measures are not only effective in reducing the false inferences, but they also introduce certified edges missed by the existing similarity measures. Given this, we believe that well-suited *extrinsic* measures, can give additional insight into the gene similarity networks which cannot be captured by an *intrinsic* measure.

	Colon Cancer			Lung Cancer		
	TP	FP	PPV	TP	FP	PPV
Pearson	427	548	0.44	3571	4027	0.47
TOM	420	539	0.44	2913	4125	0.41
Confidence	409	583	0.41	2881	3719	0.44
Smi	445	419	0.52	<b>4494</b>	3814	<b>0.54</b>
Chi	<b>449</b>	<b>395</b>	<b>0.53</b>	4309	<b>3702</b>	<b>0.54</b>

We also evaluate the effect of using different percentile cut-offs that are used to infer ‘TP’ and ‘FP’ pairs. For this purpose, we re-analyze the gene network generated from the colon cancer dataset by varying the percentile cut-offs. We vary upper tail percentile cut-offs between 0.05, 0.1 and 0.2 and correspondingly lower tail cut-offs between 0.95, 0.9 and 0.8. We then analyze the PPV of colon cancer ‘similarity’

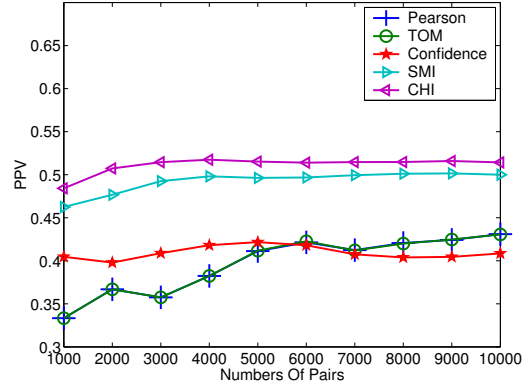
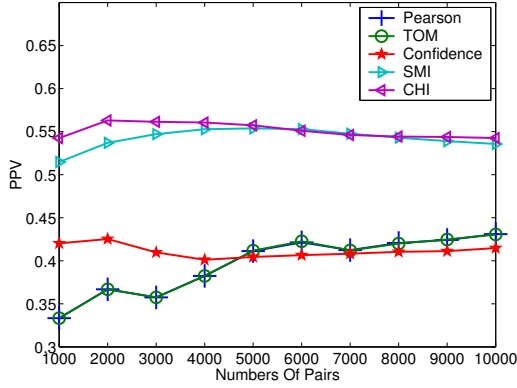


Figure 2: PPV of the top ‘similar’ gene pairs identified from Colon cancer dataset for different values of  $\kappa$  (a)0.45 and (b)0.55.

networks using these varying cut-offs (depicted in Figure 3). As can be seen from this figure, although changing the cut-

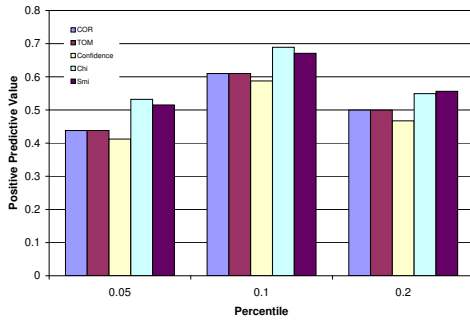


Figure 3: Evaluation of colon cancer network for various percentile cut-offs.

offs effect the mere value of PPVs, networks generated from *extrinsic* measures do consistently better than their intrinsic counterparts for any cut-off setting. However, we also note that when a wider (lower and upper) tail is considered for our analysis, the improvement of *extrinsic* measures over *intrinsic* measures decreases. For example when we compare *smi* measure with Pearson, the increase in PPV decreases from 18% to 12% when the 20<sup>th</sup> (and 80<sup>th</sup>) percentile is used instead of the 5<sup>th</sup> (and 95<sup>th</sup>) percentile. This can be attributed to the existence of missing information in the GO database. As expected, inference based on wider tails are more severely affected by the partial information than the inference based on extreme tails.

#### 6.4 Effect on Network Clusters

In this experiment, we examine the quality of clusters extracted from different gene similarity networks. Extracting groups of genes that are tightly connected in a co-expression network is important for the inference of functional annotation [10, 21, 3]. However, it is not yet clear which clustering/partitioning method is the most useful one for this purpose. To identify dense regions from our networks, we employ the most commonly used clustering algorithm, i.e., hierarchical clustering with UPGMA. To our knowledge, no entirely reliable method exists for identifying the correct number of clusters (i.e.,  $k$ ) in a dataset. That is why, we

perform hierarchical clustering for a range of different numbers of clusters ( $100 \leq k \leq 1000$ ). Modularity measure proposed by Newman et al [14] is used to estimate the correct number of clusters for each network. As suggested by the modularity analysis, colon and lung cancer networks are initially partitioned into 500 and 400 clusters respectively. Each clustering arrangement is validated using the cluster validation measure (*CPPV*). We then eliminate the clusters with zero *CPPV* values and plot *CPPV* of the remaining ones (depicted in Figures 4a-b). As can be observed from these figures, *smi* and *chi* networks produce more clusters with high *CPPV* values for both datasets. These results confirm that networks generated based on external similarity notions are better sources for obtaining biologically more meaningful clusters.

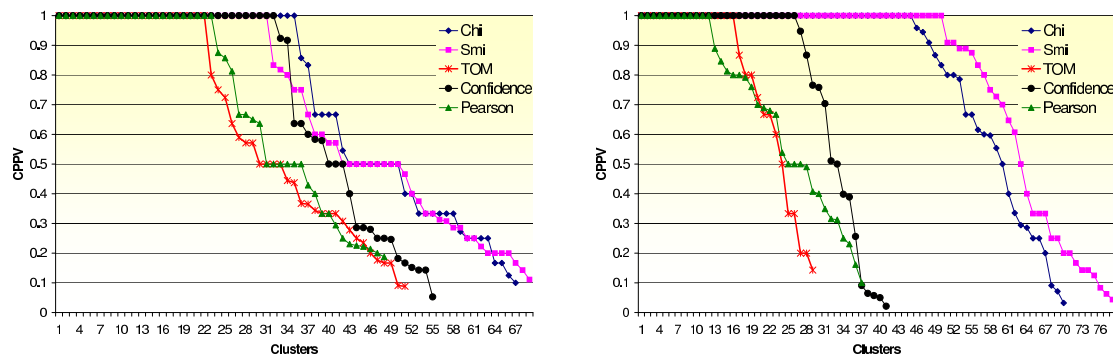
We next investigate the importance of identifying biologically sound groupings for reaching a better understanding of cancer and consequently developing new treatments.

## 7. 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 make use of the colon cancer dataset which is composed of tumorous and non-tumorous tissues of the human colon and rectum. As being the third most common cancer and the second leading cause of cancer-related death in US, a better understanding of the development and progression of this disease can be crucial for determining novel targets and strategies for its treatment.

Our experimental results show that by using *extrinsic* similarity notions, we obtain clusters with higher *CPPV* implying pairwise similarities of genes in the same cluster. However, pairwise similarities do not prove that the cluster is composed of many genes that are involved in the same pathway or molecular function. We further analyze the extracted clusters to investigate the ones that are functionally coherent. For this purpose, we employ an enrichment analysis that signifies the statistical value of a cluster’s functional homogeneity. We calculate an enrichment score (i.e., p-value) which is defined as the chance of observing that particular grouping, or better, given the background distribution<sup>5</sup>.

<sup>5</sup>All three ontologies are employed. For more details please refer to our previous work [24].



**Figure 4: Distribution of CPPV for clusters extracted from (a) Colon cancer ( $k = 500$ ) and (b) Lung cancer datasets ( $k = 400$ ).**

Among all clusters, the ones that are significantly enriched in genes from the same functional group are determined and presented in the following table. Recommended cut-off of 0.05 is used for all our validations. A more detailed analysis of these significant clusters is 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.

Several of the clusters extracted from the *chi* network, are annotated with the GO terms related to the *Signal Transduction Pathway* (i.e., *receptor signaling protein activity*, *signal transducer activity*, *scavenger receptor activity*). This is an important pathway targeted for colorectal cancer treatment [7]. Thus, studying these clusters might be important for understanding the role of signal transduction in colorectal cancer, and accordingly introducing promising molecular targets, and strengthening the existing therapeutic approaches. An additional use of these clusters might be to understand the interactions between various functional groups that initiate and maintain colorectal cancer. One can study the edges between clusters in order to reveal this information. Other measures cannot disclose the biological signal regarding the role of *Signal Transduction Pathway* in colon cancer from our test data.

From the *smi* network, we extract a cluster that is composed of genes associated with the GO term *cytoskeleton*. 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 (and less error-prone) clusters are identified, relations between these clusters can be used to reveal the function level interactions vital for understanding the cause of some diseases.

Besides revealing pathways and functional groups associated with the colon cancer, significant clusters can also be employed for function prediction. Determining the functions of genes is a central problem in biology [21, 5, 13]. An unannotated gene that is located into a cluster with a significant functional annotation can be predicted to be part of this

same functional module. Our hypothesis is that clusters that are functionally more coherent are better sources for function prediction. As an example, one of the *smi* clusters is associated with the GO term *tRNA metabolism*. In this group, a gene (H05910) does not have a known annotation. This suggests that the unknown gene might have an unrevealed task in this biological process. Using other similarity measures the same gene is located into clusters that are not enriched in any functional gene groups which provides no information for function prediction and identification.

GO Term	Measure	p-value
receptor signaling protein activity	<i>Chi</i>	.000291
signal transducer activity	<i>Chi</i>	.000091
scavenger receptor activity	<i>Chi</i>	.000278
immunological synapse	<i>Chi</i>	.000590
Ras GTPase binding	<i>Chi</i>	.000209
phosphoprotein binding	<i>Chi</i>	.000160
mRNA metabolism	<i>Chi</i>	.000480
protein homooligomerization	<i>Chi</i>	.000217
regulation of metabolism	<i>Chi</i>	.000049
positive regulation of I-kappaB kinase/NF-kappaB cascade	<i>Chi</i>	.000062
secretion	<i>Chi</i>	.000250
general RNA polymerase II transcription factor activity	<i>Smi</i>	.000761
phosphatase regulator activity	<i>Smi</i>	.000965
secretory granule	<i>Smi</i>	.000309
leading edge	<i>Smi</i>	.000189
non-membrane-bound organelle	<i>Smi</i>	.000359
cytoskeleton	<i>Smi</i>	.000453
cation channel activity	<i>Smi</i>	.000096
DNA-directed RNA polymerase activity	<i>Smi</i>	.000603
hematopoietin/interferon-class cytokine receptor activity	<i>Smi</i>	.000965
FAD binding	<i>Smi</i>	.000774
translation initiation factor activity	Pearson	.000500
synaptic transmission	Pearson	.000031
obsolete molecular function	Pearson	.000283
synaptic transmission	TOM	.000030
protein N-terminus binding	TOM	.000217
acetyl-CoA C-acyltransferase activity	Conf.	.000279
helicase activity	Conf.	.000025
golgi apparatus	Conf.	.000339

## 8. CONCLUSION

In this paper, we have introduced the notion of *mutual independence* of genes based on their relations with their common neighbors. We have presented suitable *extrinsic* similarity measures for microarray analysis that make use of the *mutual independence analysis*. We have investigated the efficacy of the proposed measures and run thorough analysis to compare them with other measures available in the literature. Our experimental results prove that using the *extrinsic* measures it is possible to identify gene pairs that are biologically more relevant. In addition, association networks generated based on these measures are shown to contain more ‘TP’ edges and less ‘FP’ edges.

Our analysis also shows that different similarity notions can reveal different aspects of a microarray dataset as implied by the diverse annotations extracted from different networks. Previously, we have studied different ensemble techniques to improve clustering results on a scale-free protein interaction network [2]. We believe that an ensemble approach in integrating different aspects of a dataset captured by different similarity measures could work well in microarray analysis. In the future, we plan to investigate this. As an extension, we would also like to work on characterizing the group level interactions among genes and gene products using the multivariate information analysis.

## 9. ACKNOWLEDGMENTS

The authors would like to thank Sitaram Asur for his valuable comments.

## 10. REFERENCES

- [1] U. Alon and 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] S. Asur, D. Ucar, and S. Parthasarathy. 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] G. Bader and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.
- [4] D. Beer and et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 9:816, 2002.
- [5] A. Butte and I. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 5:418–429, 2000.
- [6] S. Carter, C. Brechbiler, M. Griffin, and A. T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20:14:2242–2250, 2004.
- [7] S. J. Cohen, R. B. Cohen, and N. J. Meropol. Targeting signal transduction pathways in colorectal cancer—more than skin deep. *Journal of Clinical Oncology*, 23:5374–5385, 2005.
- [8] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 23–29, 1998.
- [9] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. *Report C-1997-66, University of Helsinki, Department of Computer Science*, October 1997.
- [10] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci*, 95:25:14863–14868, 1998.
- [11] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. Int’l Conf. Research in Computational Linguistics, ROCKLING X*, 1997.
- [12] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th Int’l Conf. Machine Learning*, 1998.
- [13] T. Murali, C. Wu, and S. Kasif. The art of gene function prediction. *Nature Biotechnology*, 24:1474–1475, 2006.
- [14] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
- [15] I. Näthke. Cytoskeleton out of the cupboard: colon cancer and cytoskeletal changes induced by loss of *apc*. *Nature Reviews Cancer* 6, pages 967–974, 2006.
- [16] B. Ostel. Statistics in research basic concepts and techniques for research workers. *Iowa State University Press, Ames, Iowa, USA*, 1963.
- [17] R. P. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1:448–453, 1995.
- [18] C. Palmer and C. Faloutsos. Electricity based external similarity of categorical attributes. *7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2003)*, 2003.
- [19] E. Ravasz and et al. Hierarchical organization of modularity in metabolic networks. *Science*, 297:5586:1551–1555, 2002.
- [20] J. L. Sevilla and et al. Correlation between gene expression and go semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:4, 2005.
- [21] B. Snel, P. Bork, and M. Huynen. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci*, 99:5890–5895, 2002.
- [22] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100:21, 2003.
- [23] J. Stuart, E. Segal, D. Koller, and S. Kim. A gene coexpression network for global discovery of conserved genetic modules. *Science*, 302:5643:249–255, 2003.
- [24] D. Ucar, S. Asur, U. V. Catalyurek, and S. Parthasarathy. Improving functional modularity in protein-protein interactions graphs using hub-induced subgraphs. *PKDD*, pages 371–382, 2006.
- [25] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:1, 2005.