Pacific Symposium on Biocomputing 15:145-156(2010)

SUBSPACE DIFFERENTIAL COEXPRESSION ANALYSIS: PROBLEM DEFINITION AND A GENERAL APPROACH

GANG FANG, RUI KUANG, GAURAV PANDEY, MICHAEL STEINBACH, CHAD L. MYERS and VIPIN KUMAR

Department of Computer Science, University of Minnesota, Twin Cities
200 Union Street SE, Minneapolis, MN 55455, USA
E-mail: {gangfang, kuang, gaurav, steinbac, cmyers, kumar}@cs.umn.edu

In this paper, we study methods to identify differential coexpression patterns in case-control gene expression data. A differential coexpression pattern consists of a set of genes that have substantially different levels of coherence of their expression profiles across the two sample-classes, i.e., highly coherent in one class, but not in the other. Biologically, a differential coexpression patterns may indicate the disruption of a regulatory mechanism possibly caused by disregulation of pathways or mutations of transcription factors. A common feature of all the existing approaches for differential coexpression analysis is that the coexpression of a set of genes is measured on all the samples in each of the two classes, i.e., over the *full-space* of samples. Hence, these approaches may miss patterns that only cover a subset of samples in each class, i.e., *subspace patterns*, due to the heterogeneity of the subject population and disease causes. In this paper, we extend differential coexpression analysis by defining a subspace differential coexpression pattern, i.e., a set of genes that are coexpressed in a relatively large percent of samples in one class, but in a much smaller percent of samples in the other class. We propose a general approach based upon association analysis framework that allows exhaustive yet efficient discovery of subspace differential coexpression patterns. This approach can be used to adapt a family of biclustering algorithms to obtain their corresponding differential versions that can directly discover differential coexpression patterns. Using a recently developed biclustering algorithm as illustration, we perform experiments on cancer datasets which demonstrates the existence of subspace differential coexpression patterns. Permutation tests demonstrate the statistical significance for a large number of discovered subspace patterns, many of which can not be discovered if they are measured over all the samples in each of the classes. Interestingly, in our experiments, some discovered subspace patterns have significant overlap with known cancer pathways, and some are enriched with the target gene sets of cancer-related microRNA and transcription factors. The source codes and datasets used in this paper are available at http://vk.cs.umn.edu/SDC/.

Keywords: Differential coexpression; differential biclustering; differential network analysis; association analysis

1. Introduction

Diseases are often caused by perturbations in networks of genes or their products that are working together to keep a cell in a healthy state. DNA microarrays are one of the most popular technologies for studying these perturbations and understanding their effect on the expression of genes at a large scale, and eventually linking them to diseases. The genome-wide expression profiles of many types of diseases, particularly tumors, have been analyzed, and several associations have been identified between gene expression profiles and phenotypes corresponding to different stages of cancer.[30] Traditional analysis of gene expression data for this task focuses on the identification of (groups of) genes with substantially different expression values (up- or down-regulated) across sample-classes of interest, commonly known as differentially expressed (DE) genes (or patterns).[8] An example of such a group of differentially expressed genes is shown in Figure 1(a), where these genes have significantly higher expression levels in the disease class than in the control class.

However, given that diseases are often caused by the disruption of a system, or network, of genes, identifying only the individual differentially expressed genes may not be adequate for discovering the underlying mechanisms of all the diseases. An important example of such mechanisms is the dysregulation of signaling pathways in cancer.[14] A complementary view for studying these mechanisms is provided by a differential coexpression pattern (DC),[20,22,27,34] which is defined as a set of genes that have substantially different levels of coherence of their expression profiles in the two sample-classes, i.e., highly coherent in one class, but not in the other. An example of a DC pattern is shown in Figure 1(b), where the constituent genes are either all up-, down-, or neutrally-regulated for each sample in the control group (shown by the vertical streaks), but they do not follow any particular trend in the disease group. Biologically, a differential coexpression pattern may indicate the disruption of a regulatory mechanism possibly caused by the disregulation of a pathway[20] or a mutation of a transcription factor,[16,17] among other mechanism. Figure 1(c) illustrates one of these mechanisms, where the mutation of a regulator causes the disruption of the normal activity of a pathway. Specifically, $G0$ is a dominant regulator of

Pacific Symposium on Biocomputing 15:145-156(2010)

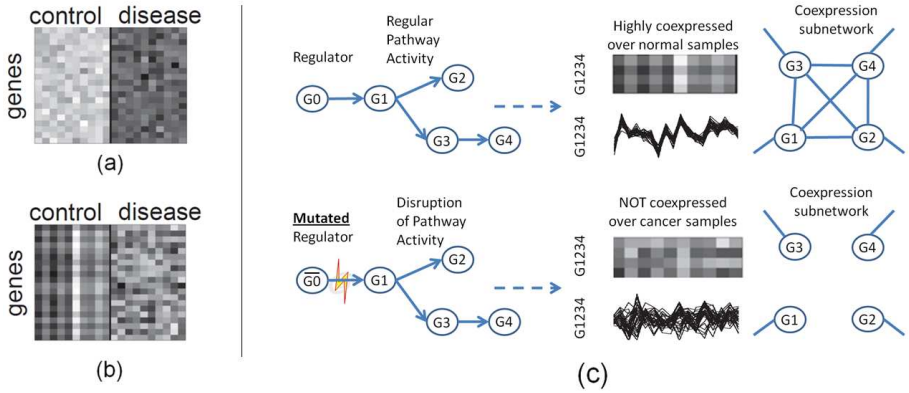


Fig. 1. Illustration of (a) a differential expression (DE) pattern, (b) a differential coexpression (DC) pattern, and (c) a possible mechanism for the occurrence of a DC pattern due to the mutation of a regulator. Note that, while $G0$ is a dominant regulator of $G1$-$G4$, the latter genes are also regulated by other *independent* regulators, that are not shown in the two pathway graphs for simplicity. ((a) and (b) taken from Kostka and Spang (2004).[20])

$G1$-$G4$ that leads to the coordinated and hence coherent expression of all of them. However, once mutated in the disease state, $G0$ is unable to regulate these genes, and their regulation may be taken over by other *independent* regulators that may only be active in the disease state. Now, since the regulation of G1-G4 is independent, they are no longer coordinated, thus leading to the disruption of the coherence of their expression. Therefore, DC patterns can serve as biomarker candidates for some diseases, e.g. cancer and its subtypes,[16,20,47] as well as for differentiating between evolutionarily-related species.[17] Furthermore, at a representation level, a DE pattern can be considered as a connected subgraph of a coexpression network, which is intact in one sample-class but not connected in the other. Such a convenient representation of these patterns can be very useful for their visualization and understanding, and we present some examples of this in Section 3.

Owing to their definition, differential coexpression patterns cannot be discovered via univariate analysis, since the coherence of expression values of a group of genes has to be measured collectively. Corresponding to this need, several techniques for identifying DC patterns have been proposed in the literature, the first ones of which only searched for gene-pairs with sufficiently different correlations (or other statistical measures) between the two classes.[21–23,34] Extending this to larger groups of genes, differential coexpression has also been studied in the context of clustering[17,44] and coexpression networks,[7,11,12,27,47,48] where a cluster or a sub-network of genes is considered differentially coexpressed if they collectively have different pairwise coexpression across the sample-classes of interest. Some algorithms also employ differential coexpression measures collectively for a set of genes,[20,29] instead of only pairwise coexpression measures. Recently, some studies have adopted a related but different perspective and have proposed methodologies for identifying differentially coexpressed gene-pathway pairs[31] and pathway-pathway pairs.[6]

Despite the differences in the methodologies adopted by these approaches for finding DC patterns, a feature common to all of them is that the coexpression of a set of genes is measured over all the samples in each of the two classes, i.e., over the *full space* of samples. For instance, the example shown in Figure 1(b) is a full-space DC pattern. However, as pointed out for the discovery of differentially expressed genes,[39,41,46] the causes of diseases as well as the population affected by them, are often heterogeneous in nature. In such a scenario, full-space approaches may not always be appropriate and may ignore patterns that cover only a subset of the samples in each class, i.e., *subspace patterns*. For instance, a set of genes may only be coexpressed over 60% of the samples in the normal class, and may not be even slightly coexpressed over any of the samples in the disease class, thus qualifying to be a valid subspace pattern. However, this pattern may not be uncovered if the discovery algorithm requires the constituent genes to be coexpressed over all the samples in the normal class. Indeed, even if a pattern can be discovered by both full-space approaches and subspace approaches, the latter can better indicate the subgroup of samples on which the pattern is coexpressed, and thus may allow further study of the different causes of diseases and different demographics among subgroups of samples, which may potentially help personal diagnosis and treatment. These challenges call for the design of new approaches that

Pacific Symposium on Biocomputing 15:145-156(2010)

can discover patterns that only show differential coexpression over subsets of the samples in the two classes, and can also indicate these subsets as a companion to the patterns. Interestingly, similar challenges faced by traditional clustering approaches have motivated the design of a variety of biclustering algorithms.[5,25]

In this paper, we address these challenges by extending differential coexpression analysis to enable the discovery of subspace DC patterns. We define these patterns as sets of genes that are coexpressed over a relatively large percent of the samples in one class, but in a much smaller percent of samples in the other class. Following this definition, we propose a general approach based upon association analysis framework[1] that allows exhaustive[a] yet efficient discovery of subspace differential coexpression patterns. This approach can be used to adapt a family of biclustering algorithms that have antimonotonicity[24,28,43,49,51] to obtain their corresponding differential versions that can directly discover differential coexpression patterns. Specifically, we illustrate the features of our approach by extending a recently developed biclustering algorithm.[28] Experiments using this approach on lung cancer datasets demonstrate the existence of subspace differential coexpression patterns in real-life data. Permutation tests demonstrate the statistical significance for a large number of discovered subspace patterns, many of which can not be discovered if they are measured over all the samples in each of the classes. Interestingly, some discovered patterns also have a significant overlap with known cancer pathways, and some are enriched with the target gene sets of a cancer-related microRNA and a cancer-related transcription factor. These results suggest that subspace DC patterns may aid in developing new understanding about the mechanisms underlying cancer and other diseases.

2. Proposed Approach

In this section, we first extend differential coexpression analysis to subspace patterns, then we will describe a general approach for the discovery of subspace differential coexpression patterns.

2.1. *Subspace Differential Coexpression Analysis*

A subspace differential coexpression pattern is a set of genes that are highly coexpressed in a relatively large percent (not necessarily all) of samples in one class, but in a much smaller percent of samples in the other class. We formulate the problem of subspace differential coexpression pattern discovery as follows. Let D be a gene expression dataset with a set of p genes, $G = \{g_1, g_2, ..., g_p\}$, and two classes of samples, A and B, which can be considered as cases and controls of size N_A and N_B, respectively, i.e., $A = \{a_1, a_2, ..., a_{N_A}\}$ and $B = \{b_1, b_2, ..., b_{N_B}\}$. Let Ψ be a coexpression measure for a set of genes α ($\alpha \subseteq G$). To illustrate, this measure could be a test as to whether the minimum of the pairwise correlation of the expression profiles of the genes in α is above a particular threshold. We use $A_\Psi(\alpha)$ ($B_\Psi(\alpha)$) to denote the subset of samples in A (B) on which α is coexpressed, i.e., $A_\Psi(\alpha) \subseteq A$ and $B_\Psi(\alpha) \subseteq B$. The two ratios, $\frac{|A_\Psi(\alpha)|}{|A|}$ and $\frac{|B_\Psi(\alpha)|}{|B|}$ are respectively the percentage of samples in A and B on which α is coexpressed. They are denoted as $R_A^\Psi(\alpha)$ and $R_B^\Psi(\alpha)$, respectively. The absolute difference of these two ratios can be used to measure the subspace differential coexpression of α:

Definition 2.1. Subspace Differential Coexpression (SDC)

$$SDC^\Psi(\alpha) = |R_A^\Psi(\alpha) - R_B^\Psi(\alpha)| \quad (1)$$

Given a threshold d, a set of genes α ($\alpha \subseteq G$) is called *$d-$differentially coexpressed if* $SDC^\Psi(\alpha) \geq d$. Then, the problem of subspace differential coexpression pattern discovery with reference to a threshold d can be formulated as discovering all the *$d-$differentially coexpressed* patterns.

We will explain our approach for addressing this problem using Figure 2, which shows a number of types of subspace and full-space, differentiating and non-differentiating, coexpression patterns. Figure 2(a) shows a conceptual example of a differential full-space pattern, while Figure 2(b) shows a conceptual example of a differential subspace pattern. Figures 2(c) and 2(d) are examples of non-differential patterns. Although Figure 2(e) is a differential full-space pattern, it contains a redundant gene, i.e., the dashed curve.

[a]Given a threshold, an exhaustive search guarantees to discover all the patterns w.r.t. that threshold. Different from brute-force search, exhaustive search may avoid exploring the whole search space by pruning a large number of patterns that are guaranteed to disqualify the threshold.

Pacific Symposium on Biocomputing 15:145-156(2010)

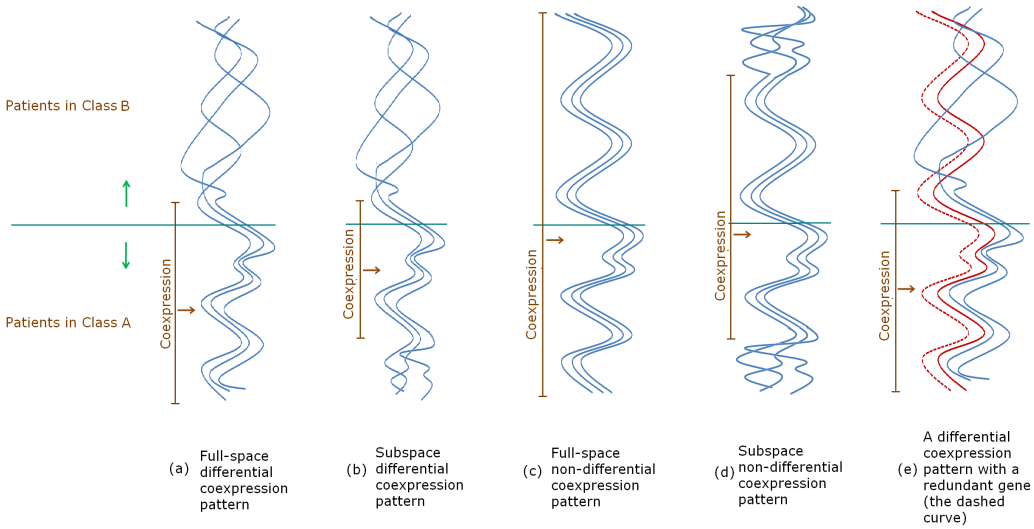


Fig. 2. Different types of full-space and subspace, differential and non-differential coexpression patterns. Each curve denotes the expression values of a gene on all the samples. The horizontal line separates all the samples into class A and B. Five patterns are illustrated, with a brown line indicating the samples on which the set of genes are coexpressed.

Given Definition 2.1, an effective mining algorithm is expected to discover patterns like (a), (b) and (e) in Figure 2, but not the patterns that are equally coexpressed in the two classes (as shown in Figure 2(c) and (d)). However, if we take a further look at pattern (e), we can observe that, although the four genes together have differential coexpression, two genes are coexpressed in both of the two classes (the two red curves). Only one of the two coexpressed genes is enough to form a differential coexpression pattern with the other two genes (the two blue curves). Indeed, including both genes would lead to redundancy that is conceptually unappealing and increases the number of patterns, without improving the SDC measure given in Definition 2.1. We considered the patterns like (e) as redundant ones that can be represented by their subsets. Therefore, for the control of redundant genes as well as for efficient pattern discovery, patterns like (e) will also be pruned together with the non-differential ones like (c) and (d) in the pattern mining process.

A common property of the five patterns in Figure 2 is that, they are all coexpressed in a large percent of samples in class A. In the meanwhile, the common property of the three patterns that are expected to be pruned, namely (c),(d) and (e), is that they all have at least one pair of genes that are coexpressed in a large percent of samples in class B. Motivated by these two observations, we refine the target of subspace differential coexpression pattern mining as those sets of genes that are coexpressed in a relatively large percent of samples in one class, while all of the pairs of genes in the set are coexpressed in a much smaller percent of samples in the other class. Mathematically, we define a measure for this refined criteria as follows:

Definition 2.2. Refined Definition of Subspace Differential Coexpression ($\widetilde{SDC}$) (Assume $R_A^\Psi(\alpha) \geq R_B^\Psi(\alpha)$)

$$\widetilde{SDC}^\Psi(\alpha) = R_A^\Psi(\alpha) - max_{i,j\in\alpha}(R_B^\Psi(\{i,j\})) \tag{2}$$

$\widetilde{SDC}$ is computed as the difference between the percent of samples in class A on which α is coexpressed and the maximal percent of samples in class B on which a size-2 subset of α is coexpressed. A large value for $\widetilde{SDC}$ indicates that a set of genes, α is coexpressed on a much larger percent of samples in class A compared to the coexpression of any size-2 subset of α in class B. Therefore, given a proper threshold, d, $\widetilde{SDC}$ can differentiate interesting subspace differential coexpression patterns like patterns (a) and (b) from uninteresting patterns like patterns $(c) - (e)$.

Mathematically, for some coexpression measures, $\widetilde{SDC}$ has another property called antimonotonicity, which basically means that $\widetilde{SDC}(\alpha)$ is guaranteed to be no less than the $\widetilde{SDC}$ of any superset of α. For $\widetilde{SDC}(\alpha)$ to have the antimonotonicity property, it is sufficient that the coexpression measure used to define $\widetilde{SDC}(\alpha)$ is antimonotonic (A formal proof is given in Fang et al.[9]). Indeed, the coexpression measures used in several

Pacific Symposium on Biocomputing 15:145-156(2010)

existing association-based and subspace clustering based biclustering algorithms have this property.[24,28,43,49,51] This antimonotonicity property guarantees that, given a threshold, d, $\widetilde{SDC}$ can be used in a systematic yet efficient pattern mining framework, namely Apriori,[1] to discover all and only the patterns with $\widetilde{SDC} \geq d$. We briefly describe the computational algorithm for this approach in Section 2.2.

2.2. *Computation Algorithm*

The Apriori framework is essentially a bottom-up exhaustive combinatorial search framework initially designed for association analysis on binary data. Different from brute-force search, given an antimonotonic measure M and a threshold m, the Apriori search algorithm can avoid exploring the whole search space of all sets of items (genes in our case) by pruning a large number of candidates that are guaranteed to disqualify the threshold based on the antimonotonicity of M.

The process of searching patterns with $\widetilde{SDC} \geq d$ in the Apriori framework can be viewed as the generation of a level-wise pattern tree. Every level of the tree contains patterns with the same number of genes. If the level is increased by one, the pattern size (number of genes in each pattern) is also increased by one. Every pattern has a branch (sub-tree) which contains all the supersets of this pattern. The search is breadth-first. We first check all the patterns at the second level, since the elemental component of differential coexpression analysis is a pair of genes. If a pattern does not satisfy the user-specified $\widetilde{SDC}$ threshold d, the whole branch corresponding to this pattern can be pruned without the need of further checking. This is guaranteed by the antimonotone property of $\widetilde{SDC}$ measures.[9] Following this approach, the pattern tree grows level-by-level until all the qualified patterns have been discovered. This algorithm is systematic yet efficient for handling large-scale datasets. Note that, in Definition 2.2, it is assumed that $R_A^\Psi(\alpha) \geq R_B^\Psi(\alpha)$. In practice, the algorithm will be run twice, one time to find patterns for which $R_A^\Psi(\alpha) \geq R_B^\Psi(\alpha)$, and the other to find patterns for which $R_B^\Psi(\alpha) \geq R_A^\Psi(\alpha)$. Use of the general measure $\widetilde{SDC}$ in the Apriori framework allows the effective pruning of non-differential coexpression patterns like (c) and (d), and also controls gene redundancy in patterns like (e). $\widetilde{SDC}$ also provides the antimonotonicity that allows exhaustive yet efficient discovery of differential coexpression patterns like (a) and (b) (Figure 2) in the Apriori framework. We will use $\widetilde{SDC}$-*Apriori* to denote the the approach of using the general measure $\widetilde{SDC}$ in the Apriori algorithm.

The coexpression measures used in several existing association-analysis-based and subspace-clustering-based biclustering studies have the antimonotonicity property[24,28,43,49,51] and can be adapted to yield their corresponding differential versions that can directly discover differential coexpression patterns. Because of the complementarity of biclustering algorithms (i.e. they may discover patterns in common with each other, as well as some unique to their formulation), their corresponding differential versions are also complementary to each other.

2.3. *DiffRange: an illustration of* $\widetilde{SDC}$

In this paper, we shall use a specific instance of this approach based on a recently proposed antimonotonic coexpression measure, namely range-support.[28] This measure is intended for the discovery of constant-row biclusters[b][25] in the Apriori framework. Conceptually, a range-support pattern is a set of genes that are coexpressed (the expression value of the set of genes fall within a close range) over a set of conditions in a gene expression data matrix. Let $RangeSup_A^r(\alpha)$ denote the range-support of α in class A (an instantiation of $R_A^\Psi(\alpha)$), i.e. the percentage of samples in class A that fall within the predefined range threshold r. From Definition 2.2($\widetilde{SDC}$), the corresponding differential range-support measure *DiffRange* (Differential Range-support) can be adapted:

Definition 2.3. Given a range threshold r, the *DiffRange* of a subset of genes α ($\alpha \subseteq G$) on class A and B

$$DiffRange(\alpha) = RangeSup_A^r(\alpha) - max_{i,j \in \alpha}(RangeSup_B^r(\{i,j\})) \quad (3)$$

3. Experimental Results

In this section, we describe the experimental design for the analysis of the subspace differential coexpression patterns discovered by *DiffRange*. The, we present experimental results which demonstrate that the proposed

[b]In a constant-row bicluster, the set of genes have similar expression values on each condition/sample.

Pacific Symposium on Biocomputing 15:145-156(2010)

general approach discovers statistically significant and biologically relevant subspace differential coexpression patterns in real-life data. .

3.1. *Datasets and Preprocessing*

In the experiments, three lung cancer datasets[3,35,36] are used, which are all generated with Affymetrix microarrays.[c] To have a larger sample size for better illustration of the existence of subspace patterns and of their statistical significance[d], we combined the three datasets resulting in 102 samples with lung cancer and 67 normal samples (total 169 samples). Across the three datasets, 8787 genes are common. We preprocessed the three datasets with RMA-normalization.[18] Additional cross-platform normalization algorithms[2,32] were also tested and gave similar results, so only RMA normalized results are included here. The effect of different normalization methods on differential coexpression pattern mining will be studied in future work.

Instead of normalized gene expression data, we used rank-converted values, i.e., the expression values are converted to expression ranks ranging from 1 to 169 (number of samples) separately for each gene (similar as used in *Spearman's rank correlation*). Our analysis shows that rank-conversion can allow the discovery of patterns containing genes with different ranges of expression values but still showing differential coexpression. Thus, we focus on the analysis of the patterns discovered on rank-converted data, on which more patterns are discovered. The patterns discovered on the data with expression value are presented on our website. Note that, rank-transformation is especially useful for *DiffRange*, since it is based on the biclustering algorithm[28] designed to find constant-row patterns.[25] Such rank-transformation may not be required in the $\widetilde{SDC}$*-Apriori* framework for other biclustering algorithms that are able to find coherent additive, coherent multiplicative or coherent evolution biclusters.[25]

Higgins et al.[15e] collected a list of genes that are shown to be related to cancer. Out of the 8787 genes in the dataset, 1975 are on the cancer gene list. In the following experiments, we analyze the subspace patterns discovered on these 1975 genes (rank-converted data, and denoted as dataset D'), because cancer genes are more likely to have disregulated patterns and based on the existing knowledge of these cancer genes, the evaluation on the patterns discovered from these genes can better illustrate the biological relevance of subspace differential coexpression patterns. Note that although these 1975 genes are known to be related to cancer, the subspace differential coexpression patterns discovered on them can provide new insights about their relationship with cancer, e.g., by identifying the interactions among individual cancer genes.

3.2. *Pattern Discovery*

With parameters $r = 0.2$[f], and $d = 0.2$, *DiffRange* is used in the Apriori framework to discover subspace DC patterns on D'. Most patterns are of size-2 (gene-pairs), but there are also size-3 and size-4 patterns (no larger size patterns are discovered for the selected parameters). To control the redundancy of genes among size-3 and size-4 patterns,[g] we order them by decreasing SDC value and sequentially select a subset of the patterns in which none of the pairs of patterns have greater than 25% overlap of genes. This compact set has 95 patterns (88 size-3 patterns and 7 size-4 patterns). Figure 3(a) shows the size and SDC value for each discovered DC pattern.

3.3. *Are the discovered subspace differential coexpression patterns statistically significant?*

Due to the issues of low sample size and high-dimensionality for data sets used for problems such as biomarker discovery, many patterns may be falsely associated with the class label by random chance, especially when a large number of combinations of genes are searched. This raises the multiple-hypothesis testing problem.[33] In this paper, a permutation test is used to evaluate the statistical significance of the discovered subspace DC patterns. Specifically, the original class labels are randomly shuffled 1000 times. For each random labeling, the same

[c]The first two use platform HG-U95A, while the other uses platform HG-U133A

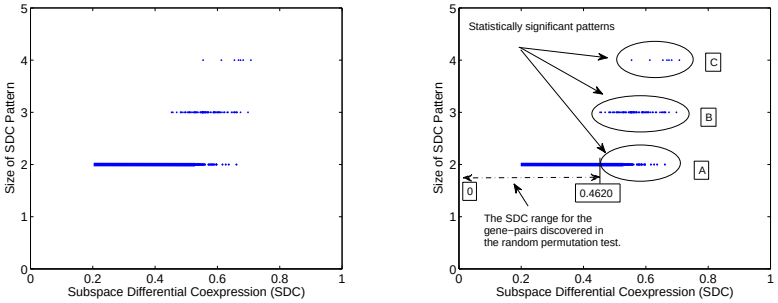
[d]The patterns discovered from the three datasets separately are not statistically significant in the permutation test (refer to Section 3.3 for details), due to the low sample size of each individual datasets.

[e]In this paper, we union the two lists respectively downloaded in October 2008 and June 2009, with a total of 2622 genes

[f]In the rank-converted data, this means k genes have coherent expression if the rank difference of their expression is less than 20% of the 169 samples, i.e. 33.

[g]Note that in the discussion of pattern (e) in Fig. 2, the redundancy is within a pattern rather than among the patterns like here.

Pacific Symposium on Biocomputing 15:145-156(2010)



(a) Size and SDC for the DC patterns discovered with the original class labels (Each dot is a pattern).

(b) Statistically significant DC patterns according to the permutation test.

Fig. 3. Patterns discovered with the original class labels (a) and indication of the statistically significant ones (b).

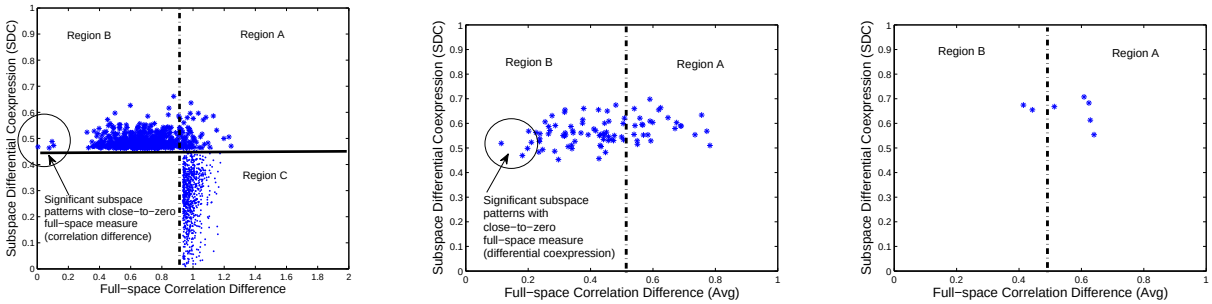
DiffRange parameters ($r = 0.2$ and $d = 0.2$) are used to discover a set of patterns. With the 1000 randomized labels, only size-2 patterns are discovered, with SDC values in the range of $[0, 0.4620]$. Figure 3(b) indicates this range by overlaying a double-arrow on top of the patterns discovered with the original class labels (as shown in Figure 3(a)). Considering 0.4620 as a statistical significance cutoff, Ellipse A indicates the 560 statistically significant gene-pairs whose SDC value was never exceeded by any random pattern. In addition, since there are no size-3 and size-4 patterns discovered with randomized labels in the permutation test, the 88 size-3 patterns (Ellipse B) and 7 size-4 pattern (Ellipse C) are also considered statistically significant. Note that, although a differential coexpression pattern can be highly coexpressed in either the cancer class or the normal class,[21,22] all the statistically significant patterns in the three ellipses are highly coexpressed in the normal class while less coexpressed in the cancer class.

3.4. *How differentially coexpressed are the discovered subspace patterns when measured over full-space?*

In this experiment, we measure the full-space differential coexpression for the statistically significant subspace patterns selected based on the above permutation test, i.e., the 560 gene-pair patterns and the 88 size-3 patterns and the 7 size-4 patterns. We will show that there are subspace patterns that have close-to-random differential coexpression when considered as full-space patterns. A variety of full-space differential coexpression measures are proposed in existing work as discussed in section 1. As used in several studies,[20,34,44] we will use the correlation difference of a pattern between the two classes for illustration purpose. For a gene-pair pattern, correlation difference is just the difference of the two correlations respectively in the two classes. For a pattern of size greater than 2, we compute the difference between the average pair-wise correlation in the normal and cancer class to measure the correlation difference.

The three subfigures in Figure 4 plot the correlation difference and SDC for the statistically significant size-2, size-3, and size-4 patterns, respectively. The three dashed lines indicate the statistical significance cutoff of correlation difference for size-2, size-3 and size-4 patterns (0.9361, 0.5176 and 0.4953), respectively, which is also decided via permutation test. For the gene-pair patterns (Figure 4(a)), several observations can be made: (i) some patterns are considered statistically significant by both correlation difference and SDC (region A); (ii) some gene-pairs are considered significant only by SDC but not by correlation difference (region B). Among these patterns, several pairs have close-to-zero correlation difference(within the circle), which means they show very little differential coexpression when considered as full-space patterns; and (iii) there are also 801 gene-pairs that are only considered significant in terms of correlation difference but not by SDC (region C). This is as expected since many factors can affect the discovery of DC patterns, e.g. different coexpression measures, different mining algorithms, and the parameters used in the algorithms. Our highlight is the existence of subspace differential coexpression patterns that show close-to-random differential coexpression when considered as full-space patterns. Similar observation can also be made in Figures 4(b) and 4(c) which respectively plot the correlation difference for the size-3 and size-4 patterns.

Pacific Symposium on Biocomputing 15:145-156(2010)



(a) The 560 gene-pair (stars). (b) The 88 size-3 patterns (c) The 7 size-4 patterns.

Fig. 4. Illustration of the full-space differential coexpression (correlation difference) for the discovered statistically significant subspace differential coexpression patterns. The dashed lines in $(a) - (c)$ indicate the statistical significance cutoffs for correlation difference for size-2, size-3 and size-4 patterns respectively. The solid line in (a) is the statistical significance cutoff for SDC (0.4620). There are no corresponding lines in (b) and (c) because all the patterns of size 3 and 4 are statistically significant in terms of SDC as discussed in section 3.3. Region A contains patterns that are considered significant by both correlation difference and SDC; Region B has patterns that are not significant as full-space patterns, several of which have close-to-zero correlation difference (within the circle); and Region C shows the significant full-space patterns that are not discovered by SDC.

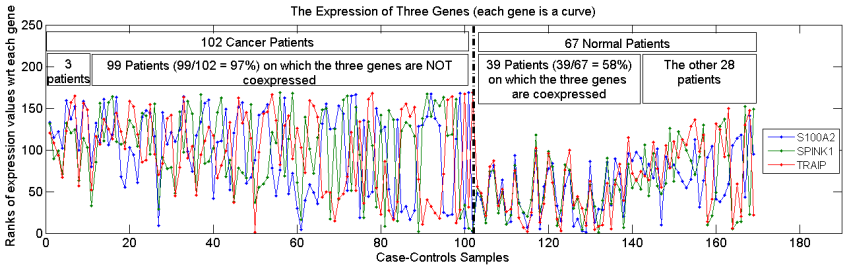


Fig. 5. A statistically significant subspace differential coexpression pattern, for which the minimal pairwise correlation in the normal class is only 0.28. For better visualization, samples are sorted by increasing range of expression ranks separately in the two classes (similar for Figures 6(a), 6(b) and 7(a)).

In Figure 5, we illustrate a subspace DC pattern with very small correlation difference (0.19). This pattern is coexpressed in only 58% of the normal samples [h], and the minimal pairwise correlation of the genes in this pattern over all the normal samples is only 0.28. For this pattern, it is not reasonable to assume that the genes are coexpressed on all the normal samples. Furthermore, discovering the pattern as a subspace DC pattern can explicitly show the subgroup of samples on which the three gene show coexpression, i.e., the 39 normal samples and the 3 cancer samples in the cancer class. This allows further analysis of the difference between the 39 normal samples with the pattern and the 28 without it (e.g., different demographic characteristics), which may help personalized diagnosis and treatment.

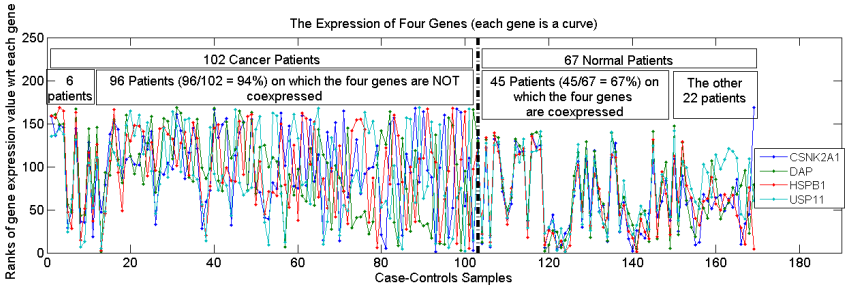
The existence of subspace patterns that show small and insignificant differential coexpression when considered as full-space patterns demonstrates the potential usefulness of subspace differential coexpression analysis. Next, we will evaluate the biological relevance of the discovered subspace patterns.

3.5. *Are the discovered subspace differential coexpression patterns biologically relevant?*

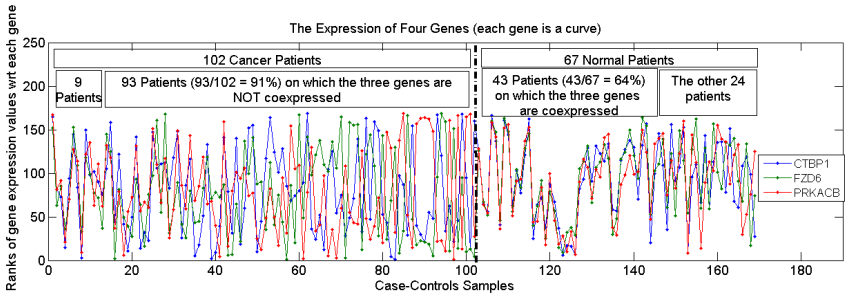
Quantitatively, two enrichment experiments are used to evaluate the biological relevance of the discovered subspace differential coexpression patterns: (i) enrichment with ten known cancer-related signaling pathways [i],[15] (ii) enrichment with the 5452 gene sets in the Molecular Signature database (MSigDB)[j].[37] Since patterns of size-2

[h]This is with respect to the *DiffRange* parameters used to discover the patterns, $r = 0.2$. Similar for Figures 6(a) and 6(b)
[i]http://cbio.mskcc.org/CancerGenes/Select.action
[j]Specifically, MSigDB contains 386 positional gene sets, 1892 curated gene sets, 837 motif gene sets, 883 computational gene sets, and 1454 GO gene sets. http://www.broadinstitute.org/gsea/msigdb/

Pacific Symposium on Biocomputing 15:145-156(2010)



(a) The pattern that is enriched with the $TNF\alpha/NF\kappa B$ signaling pathway (enrichment p-value 0.0011).



(b) The pattern that is enriched with the WNT signaling pathway (enrichment p-value 0.0042).

Fig. 6. Two patterns that are respectively enriched with the $TNF\alpha/NF\kappa B$ and the WNT signaling pathway.

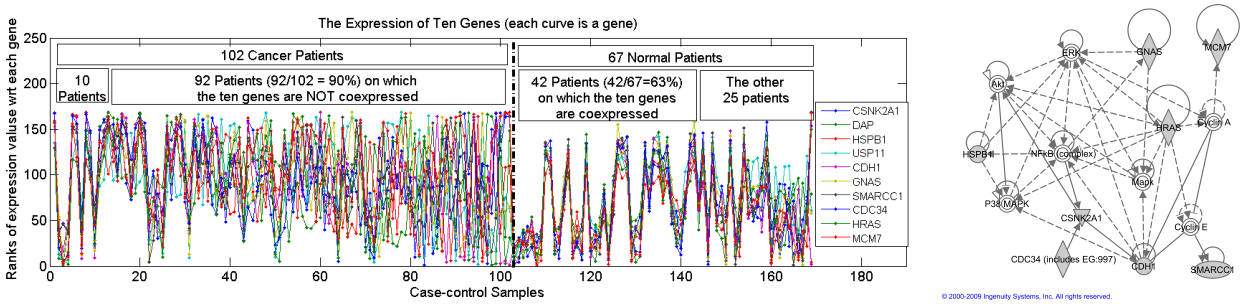
are difficult to assess in terms of enrichment, we perform the two biological evaluations only for the 95 patterns of size 3 or 4. Briefly, (i) six patterns have an overlap of 2 or more genes with one of the ten known cancer-related pathways, (ii) in the MSigDB enrichment, 40 patterns have enrichment p-value less than 0.001, among which five have p-value less than 0.0001. Detailed enrichment results can be found on the paper website.

Note that, due to the limited knowledge about differentially coexpressed patterns, the current stage of differential coexpression pattern mining is still hypothesis generation rather than hypothesis verification, as discussed in Kostka and Spang.[20] Indeed, since all the 95 patterns are statistically significant in the permutation test, and all the genes contained in the 95 patterns are known cancer-related genes[15] (as described in section 3.1), they can be considered as hypotheses that may lead to new understanding of the interactions among them, and of the relationship between differential coexpression and cancer mechanism. Therefore, in addition to the above standard enrichment analyses, we will illustrate and discuss several interesting patterns that are enriched with known cancer pathways, or target sets of cancer-related microRNAs and transcription factors.

Figure 6 displays two patterns that are enriched with the $TNF\alpha/NF\kappa B$ signaling pathway and the WNT signaling pathway respectively. Several observations can be made from these two figures. Firstly, they both show strong differential coexpression, i.e. they are both highly coexpressed in the normal class, and much less coexpressed in the cancer class. Secondly, both patterns are subspace differential coexpression patterns, i.e., they show coexpression in only 67% and 64% of the normal samples respectively. Similar to the pattern shown in Figure 5, these two patterns are coexpressed in only about two-third of the normal samples. Discovering them as subspace patterns also points out the subgroup of samples covered by them. This allows further study of the different causes of diseases and the different demographics among subgroups of samples. Finally, both the $TNF\alpha/NF\kappa B$ and WNT signaling pathways have been shown to be related to lung cancer.[19,40] Discovering the differential coexpression patterns enriched with these pathways may shed new light on the understanding of the two pathways and their relationships to cancer mechanism.

Among the six patterns that are enriched with at least one cancer pathway, three are enriched with the $TNF\alpha/NF\kappa B$ pathway. In Figure 7(a), the union of the three patterns, containing ten genes, are plotted. All the ten genes are known cancer-related genes.[15] Out of the ten genes, six overlap with the $TNF\alpha/NF\kappa B$

Pacific Symposium on Biocomputing 15:145-156(2010)



(a) Visualization of the pattern. (b) Enriched IPA subnetwork.

Fig. 7. The union of the three patterns that are all enriched with the $TNF\alpha/NF\kappa B$ signaling pathway. There are ten genes in this combined pattern (all are known cancer related genes), out of which six are included in the $TNF\alpha/NF\kappa B$ signaling pathway (enrichment p-value 1.4024×10^{-5}).

pathway (enrichment p-value 1.4024×10^{-5}). This may suggest that the other four genes may also participate in or interact with this cancer pathway.[31] Figure 7(b) shows the enriched IPA[k] subnetwork, containing 8 of the ten genes (the 8 genes are shaded). Interestingly, IPA also shows that the connecting components (*NFκB complex, ERK, Mapk and Cyclin E*) are also known to be related to cancer.

Specifically among those patterns that are not considered significant by correlation difference (Region Bs in Figure 4), some are enriched with the target gene sets of cancer-related microRNAs and transcription factors. For example, the first two genes in the pattern (*PIK3C2B,TSC22D1,AKAP12*) are among the set of target genes of miR-101 (p-value 0.001), a small non-coding RNA that regulates gene expression. miR-101 was shown to be down-regulated in cancer.[10] This agrees with the loss of coexpression of its target genes (*PIK3C2B,TSC22D1*) in the cancer class. Furthermore, the miR-101 targets are enriched for several signaling pathways, and the third gene *AKAP12* is a known regular of protein kinase A (PKA), a central signaling pathway involved in cell growth and proliferation. This may lead to the differential coexpression of (*PIK3C2B,TSC22D1*) and *AKAP12* together as a DC pattern. In another example, the first two genes in the pattern (*ETV4,PTHLH,CBX5*) are among the set of genes with promoter regions $[-2kb, 2kb]$ around the transcription start site containing the motif *TTACGTAA* which matches the binding site for the transcription factor ATF2 (p-value 2.5119×10^{-4}). Mutations of ATF2 was shown to be related to cancer,[45] which agrees with the loss of coexpression of its target genes (*ETV4,PTHLH*) in the cancer class. In addition, the ATF2 targets show enrichment for transcription regulation (repression), and *CBX5* is component of heterochromatin, an epigenetic factor in the regulation of gene expression. This may lead to the differential coexpression of (*ETV4,PTHLH*) and *CBX5* together as a DC pattern.

4. Conclusions

In this paper, we studied methods to identify disease-related change of coexpression subnetworks, i.e. differential coexpression analysis. Specifically, we extended differential coexpression analysis to subspace patterns and proposed an approach based upon association analysis framework[1] that allows exhaustive yet efficient discovery of subspace differential coexpression patterns. This approach can be used to adapt a family of biclustering algorithms to obtain their corresponding differential versions that can directly discover differential coexpression patterns. We illustrated the general approach on a recently-developed biclustering algorithm, and presented the results of experiments on lung cancer datasets using this algorithm. The results showed the existence of meaningful subspace differential coexpression patterns in real-life data. Permutation tests demonstrated the statistical significance for a large number of discovered patterns, many of which can not be discovered if they are measured over all the samples in each of the classes. Interestingly, some discovered patterns also have a significant overlap with known cancer pathways, and some are enriched with the target gene sets of a cancer-related microRNA and

[k]Ingenuity Pathway Analysis: http://www.ingenuity.com/

Pacific Symposium on Biocomputing 15:145-156(2010)

a cancer-related transcription factor. These results suggest that subspace DC patterns may aid in developing new understanding about the mechanisms underlying cancer and other diseases.

5. Limitations and Future Work

In this section, we discuss several limitations of the proposed approach, possible solutions and future work.

(1) **Size of patterns**: Due to the fixed thresholds imposed on $\widetilde{SDC}$ in the Apriori framework, there may be some larger patterns that do not satisfy the thresholds and are split into smaller ones. This limitation of association analysis is usually addressed by pattern summarization,[13] in which smaller size patterns are merged into larger ones under some criteria. For example, the size-10 pattern in Figure 7 is obtained by merging three smaller patterns as described in Section 3.5. More sophisticated summarization approaches[13] can be exploited in future work.

(2) **Enhancing scalability**: The scalability of the approach depends on the mining algorithm, as well as the permutation test. Generally, the algorithm itself takes about ten minutes for 2000 genes, several hours for 4000 genes and more than a day for all the 8787 genes[l], which is acceptable. However, the real challenge comes from the permutation test in which the mining algorithm is called 1000 times, the total time of which is unacceptable on all the 8787 genes. Thus, to have a comprehensive evaluation of the discovered patterns, we limited the pattern discovery and the follow-up statistical and biological analysis to the subset of genes that are known to be related to cancer. In future work, for the efficiency of the mining algorithm, more effective pruning schemes should be studied together with preprocessing procedures such as standard deviation based gene filtering[m]. For the scalability in the context of permutation test, efficiency could possibly be improved by reusing the calculation over the large number of permutations as studied by Zhang et al.[50]

(3) **Modifying other biclustering algorithms**: In this paper, *DiffRange* is presented as an illustration of the general approach, $\widetilde{SDC}$*-Apriori*, for modifying a biclusteirng algorithm to its differential version. As discussed in Section 2.1, $\widetilde{SDC}$*-Apriori* can also be applied to modify other biclustering algorithms[24,43,49,51] with the antimonotonicity property, and their corresponding differential versions are expected to complement *DiffRange* for discovering differential coexpression patterns.

(4) **Differential biclustering**: Differential coexpression patterns can essentially be considered as biclusters that exist mostly in one class but not in the other. Indeed, such type of biclusters have already been observed in several studies,[26,38,51] where a set of biclusters are discovered in the first step and then the ones that are unique to a single class are selected in the second step. Such a two-step approach can also be used to discover differential coexpression patterns. However, the general approach proposed in this paper, $\widetilde{SDC}$*-Apriori*, can be considered as an initial effort towards a more general *differential biclustering* problem, where more efficient discovery of differential biclusters are possible by making use of class labels within the biclustering process. Similar problems can also be formulated as differential/discriminative co-clustering and differential/discriminative subspace clustering in the data mining community.

(5) **Pattern-based classification**: Since a subspace differential coexpression pattern explicitly captures the subgroups of samples it covers, it will also be interesting to investigate the predictive power of subspace differential coexpression patterns in a pattern-based classification framework,[4,42] where the combination of traditional differentially expressed genes and subspace differential coexpression patterns may provide more accurate disease diagnosis.

Acknowledgments

The authors would like to thank Ba Ryun Hwang for dataset preprocessing, and thank the anonymous reviewers for the constructive comments. This work was supported by NSF grants #CRI-0551551, #IIS-0308264, and #ITR-0325949. Access to computing facilities was provided by the Minnesota Supercomputing Institute.

[l]The experiments presented here were run on a Linux machine with Intel(R) Xeon(R) CPU (E5310 @ 1.60GHz) and 16GB memory
[m]A gene with small variation across samples is less likely to constitute a differential coexpression pattern.

Pacific Symposium on Biocomputing 15:145-156(2010)

References

1. R. Agrawal and R. Srikant. In *Proc. Very Large Data Bases*, pages 487–499, 1994.
2. M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. Perou, and J. Marron. *Bioinformatics*, 20(1):105–114, 2004.
3. A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al. *PNAS*, 21(15):3301–3307, 2005.
4. H. Cheng, X. Yan, J. Han, and C.-W. Hsu. In *Proceedings. Intl Conf. Data Engineering*, pages 716–725, 2007.
5. Y. Cheng and G. Church. In *Proceedings of ISMB*, pages 8:93–103, 2000.
6. S. Cho, J. Kim, and J. Kim. *BMC Bioinformatics*, 10(1):109, 2009.
7. J. Choi, U. Yu, O. Yoo, and S. Kim. *Bioinformatics*, 21(24):4348–4355, 2005.
8. X. Cui and G. Churchill. *Genome Biology*, 4(4):210, 2003.
9. G. Fang, G. Pandey, M. Gupta, M. Steinbach, and V. Kumar. Tech Report 09-011, Department of Computer Science, University of Minnesota, 2009.
10. J. Friedman, G. Liang, C. Liu, E. Wolff, Y. Tsai, W. Ye, X. Zhou, and P. Jones. *Cancer Research*, 69(6):2623, 2009.
11. T. Fuller, A. Ghazalpour, J. Aten, T. Drake, A. Lusis, and S. Horvath. *Mammalian Genome*, 18(6):463–472, 2007.
12. P. Gargalovic, M. Imura, B. Zhang, N. Gharavi, M. Clark, J. Pagnon, W. Yang, A. He, A. Truong, S. Patel, et al. *PNAS*, 103(34):12741, 2006.
13. J. Han, H. Cheng, D. Xin, and X. Yan. *Data Mining and Knowledge Discovery*, 15:55–86, 2007.
14. D. Hanahan and R. Weinberg. *Cell*, 100(1):57–70, 2000.
15. M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. *Nucl. Acids Res.*, 35(suppl 1):D721–726, 2007.
16. N. Hudson, A. Reverter, and B. Dalrymple. *PLoS Computational Biology*, 5(5), 2009.
17. J. Ihmels, S. Bergmann, J. Berman, N. Barkai, and L. Kruglyak. *PLoS Genetics*, 1(3):e39, 2005.
18. R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, et al. *Biostatistics*, 4(2):249–264, 2003.
19. D. Kim, S. Koo, K. Jeon, M. Kim, J. Lee, C. Hong, and S. Jeong. *Cancer Research*, 63:621-626, 2003.
20. D. Kostka and R. Spang. *Bioinformatics*, 20(1):194–199, 2004.
21. Y. Lai, B. Wu, L. Chen, and H. Zhao. *Bioinformatics*, 20(17):3146–3155, 2004.
22. K. Li. *PNAS*, 99(26):16875–16880, 2002.
23. K. Li, C. Liu, W. Sun, S. Yuan, and T. Yu. *PNAS*, 101(44):15561–15566, 2004.
24. G. Liu, J. Li, K. Sim, and L. Wong. In *Proc. Intl Conf. Data Engineering*, 1250–1254, 2007.
25. S. Madeira and A. Oliveira. *IEEE/ACM Trans on Compu Bio and Bioinfo*, 1(1):24–45, 2004.
26. T. Murali and S. Kasif. In *Proc. Pacific Symposium on Biocomputing 8:77-88*, 2003.
27. M. Oldham, S. Horvath, and D. Geschwind. *PNAS*, 103(47):17973, 2006.
28. G. Pandey, G. Atluri, M. Steinbach, C. L. Myers, and V. Kumar. In *Proc. ACM Conf. on Kownlege Discovery and Data Mining*, pages 677-686, 2009.
29. C. Prieto, M. Rivas, J. Sanchez, J. Lopez-Fidalgo, and J. De Las Rivas. *Bioinformatics*, 22(9):1103–1110, 2006.
30. J. Quackenbush. *The New England journal of medicine*, 354(23):2463, 2006.
31. B. Rosemary, C. Leslie, and P. Giovanni. *BMC Bioinformatics*, 9:488, 2008.
32. A. Shabalin, H. Tjelmeland, C. Fan, C. Perou, and A. Nobel. *Bioinformatics*, 24(9):1154, 2008.
33. J. Shaffer. *Annual Review of Psychology*, 46(1):561–584, 1995.
34. C. Silva, M. Silva, L. Faccioli, R. Pietro, S. Cortez, et al. *Clinical and Experimental Immunology*, 101(2):314, 1995.
35. R. Stearman, L. Dwyer-Nield, L. Zerbe, S. Blaine, Z. Chan, P. Bunn, G. Johnson, F. Hirsch, D. Merrick, W. Franklin, et al. *Am J Pathol.*, 167(6):1763–75, 2005.
36. L. Su, C. Chang, Y. Wu, K. Chen, C. Lin, et al. *BMC Genomics*, 8(1):140, 2007.
37. A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. *PNAS*, 102(43):15545–15550, 2005.
38. A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. *PNAS*, 101(9):2981–2986, 2004.
39. S. Tomlins, D. Rhodes, S. Perner, S. Dhanasekaran, R. Mehra, X. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, et al. *Science*, 310(5748):644–648, 2005.
40. K. Uematsu, B. He, L. You, Z. Xu, F. McCormick, and D. Jablons. *Oncogene*, 22(46):7218–7221, 2003.
41. I. Ulitsky, R. Karp, and R. Shamir. In *Proc. RECOMB*, 4955:347, 2008.
42. M. van Vliet, C. Klijn, L. Wessels, and M. Reinders. *PLoS ONE*, 2(10):1047, 2007.
43. H. Wang, W. Wang, J. Yang, and P. Yu. In *Proc. ACM Conf. on Management of Data*, pages 394–405. 2002.
44. M. Watson. *BMC Bioinformatics*, 7(1):509, 2006.
45. I. Woo, T. Kohno, K. Inoue, S. Ishii, and J. Yokota. *International Journal of Oncology*, 20(3):527–531, 2002.
46. B. Wu. *Biostatistics*, 8(3):566, 2007.
47. M. Xu, M. Kao, J. Nunez-Iglesias, J. Nevins, M. West, and X. Zhou. *BMC genomics*, 9(Suppl 1):S12, 2008.
48. B. Zhang and S. Horvath. *Stat Appl in Genet and Mol Bio*, 4(1):1128, 2005.
49. X. Zhang, F. Pan, and W. Wang. In *Proc. Intl Conf. Data Engineering*, 130-139, 2008.
50. X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang. In *Proceeding of RECOMB*, volume 5541, pages 253–269, 2009.
51. L. Zhao and M. Zaki. *IEEE Intelligent Systems*, 20(6):40–49, 2005.