Pacific Symposium on Biocomputing 15:133-144(2010)

IDENTIFICATION OF COORDINATELY DYSREGULATED SUBNETWORKS IN COMPLEX PHENOTYPES

SALIM A. CHOWDHURY[1] AND MEHMET KOYUTÜRK[1,2]
[1] *Department of Electrical Engineering and Computer Science*
[2] *Center for Proteomics and Bioinformatics*
Case Western Reserve University, Cleveland, OH, USA
e-mail: {sxc426,koyuturk}@eecs.case.edu

In the study of complex phenotypes, single gene markers can only provide limited insights into the manifestation of phenotype. To this end, protein-protein interaction (PPI) networks prove useful in the identification of multiple interacting markers. Recent studies show that, when considered together, many proteins that are connected via physical and functional interactions exhibit significant differential expression with respect to various complex phenotypes, including cancers. As compared to single gene markers, these "coordinately dysregulated subnetworks" improve diagnosis and prognosis of cancer significantly and offer novel insights into the network dynamics of phenotype. However, the problem of identifying coordinately dysregulated subnetworks presents significant algorithmic challenges. Existing approaches utilize heuristics that aim to greedily maximize information-theoretic class separability measures, however, by definition of "coordinate" dysregulation, such greedy algorithms do not suit well to this problem. In this paper, we formulate coordinate dysregulation in the context of the well-known set-cover problem, with a view to capturing the coordination between multiple genes at a sample-specific resolution. Based on this formulation, we adapt state-of-the-art approximation algorithms for set-cover to the identification of coordinately dysregulated subnetworks. Comprehensive experimental results on human colorectal cancer (CRC) show that, when compared to existing algorithms, the proposed algorithm, NetCover, improves diagnosis of cancer and prediction of metastasis significantly. Our results also demonstrate that subnetworks in the neighborhood of known CRC driver genes exhibit significant coordinate dysregulation, indicating that the notion of coordinate dysregulation may indeed be useful in understanding the network dynamics of complex phenotypes.

1. Introduction

Variations among organisms occur in every aspect of biological systems, including their morphology, behavior, physiology, development and susceptibility to common diseases. Many of these phenotypes are controlled by multiple genetic and epigenetic factors and are therefore called complex phenotypes (multigenic traits), in contrast to phenotypes that are controlled by single genes (monogenic or Mendelian traits).[1]

In the past decade, genome-wide monitoring of gene expression, enabled by DNA microarray technology, has been commonly used as an important tool for the investigation of complex phenotypes, including human cancers. Differential analysis of gene expression facilitates identification of genes that are *dysregulated* with respect to the phenotype of interest; that is, genes that exhibit significant difference in the amount of mRNA transcripts present in a range of phenotype and control samples. To date, systematic analyses of differential gene expression has led to identification of genetic markers associated with many complex diseases, including leukemia,[2] breast cancer,[3] lung cancer[4] and prostate cancer,[5] as well as genes that are associated with tumor grade, metastasis, and disease recurrence .[6–9]

While investigation of differential gene expression for individual genes proves useful in identification of single-gene markers, it offers limited insights into the interplay between multiple interacting factors. Therefore, research on complex phenotypes rapidly shifts toward identification of multiple genes that are together differentially expressed. Such multiple markers are likely to shed light on the underlying molecular mechanisms of complex phenotypes. Knowledge of molecular interactions proves extremely useful in identification of multiple markers, in that it establishes the physical basis for understanding the dynamics of the interplay between multiple factors, through network models. Indeed, integration of genome-wide expression data with protein-protein interactions (PPIs) is shown to be useful in extracting subnetworks composed of genes with correlated expression profiles across diverse conditions.[10,11]

In the context of complex phenotypes, a range of algorithmic approaches are developed for the identification of phenotype-implicated subnetworks. Earlier studies quantify differential expression for each gene individually and search for subnetworks with significant aggregate differential expression.[12–15] While these

Pacific Symposium on Biocomputing 15:133-144(2010)

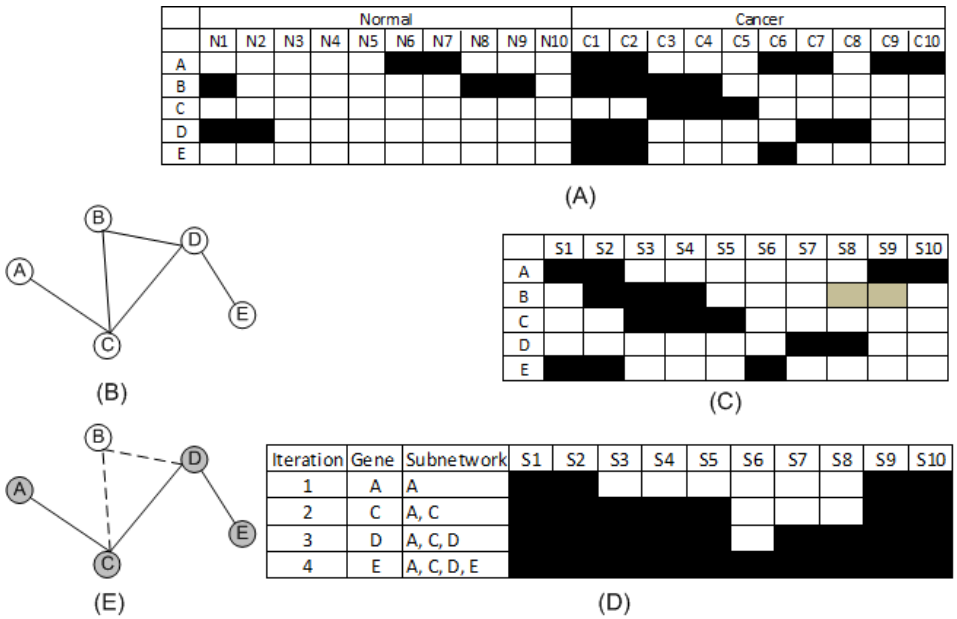

Fig. 1. Illustration of the proposed algorithm,NetCover, for identifying coordinately dysregulated subnetworks. (A) Binarized mRNA expression data with 10 paired samples and 5 genes. A light/dark square indicates that the respective gene is "expressed"/"not expressed" in the respective sample. (B) Part of human PPI network showing the interactions among the products of these genes. (C) *Differential expression profile* for each gene. A black/grey box indicates that the respective gene is up-regulated/down-regulated in the phenotype (with respect to control) for the respective paired sample (the gene *covers* the sample positively/negatively). (D) Progress of the algorithm showing the gene added to the subnetwork at each iteration and the set of samples covered by the corresponding subnetwork. (E) The resulting subnetwork, comprising of the highlighted genes and solid edges, covers all samples positively.

approaches are useful in relating individual genes that are differentially expressed, they do not necessarily capture the coordination or synergy in the dysregulation of multiple genes (e.g., genes that do not exhibit significant differential expression when considered individually, but exhibit significant differential expression when considered together). Chuang et al.[16] address this problem by considering *subnetwork activity*, defined as the aggregate expression of gene products in the subnetwork in each sample. Assessment of differential expression with respect to subnetwork activity enables identification of subnetworks that are *coordinately dysregulated*; i.e., groups of interacting proteins with collective mRNA-level differential expression. As compared to single gene markers, such subnetwork markers are shown to provide better classification performance in the prediction of disease progression in breast cancer.[16] Similarly, the concept of synergistic differential expression captures the collective differential expression of a group of genes that are not individually dysregulated.[17]

The concept of coordinate dysregulation is quite promising in generating novel insights into the network dynamics of complex phenotypes. However, the problem of identifying subnetworks with significant coordinate dysregulation is intractable. Furthermore, since the objective function associated with this problem is combinatorial in nature,[17] bottom-up heuristics that grow subnetworks to greedily maximize the objective function may seriously lack global awareness. Motivated by these considerations, we formulate this problem as a variation of the well-known set-cover problem. The proposed approach is illustrated in Figure 1. As seen in the figure, we first quantize a gene expression dataset with paired samples into binary expression levels. Then, for each gene, we identify individual samples that are *covered* (can be discriminated as phenotype or control) by the expression level of that gene. Subsequently, we search the human protein-protein interaction (PPI) network for subnetworks composed of genes that together cover all samples in the dataset. Since the genes in such a subnetwork complement each other in discriminating phenotype and control, we expect that these subnetworks may have a modular role in the manifestation of phenotype.

In the next section, we formally introduce the proposed framework. We argue that this formulation better suits to the notion of coordinate dysregulation from a biological perspective, in that it captures the coordination between multiple genes at a sample-specific resolution. Furthermore, we theoretically establish the relationship between set-cover and information-theoretic formulation of coordinate dysregulation. We then adapt state-of-

Pacific Symposium on Biocomputing 15:133-144(2010)

the-art approximation algorithms for set-cover to the problem of identifying minimal subnetworks that cover all samples in a dataset. In Section 3, we evaluate the biological relevance of identified subnetworks in the context of diagnosis and prognosis of human colorectal cancer (CRC). Comprehensive experimental results show that, subnetworks identified by the proposed algorithm, NetCover, outperform subnetworks identified by existing algorithms in terms of accurate classification of tumorigenic and metastatic samples.

2. Methods

In this section, we first introduce the notion of coordinate dysregulation and provide the motivation for the proposed algorithmic approach. Then, based on an information-theoretic formulation of coordinate dysregulation, we demonstrate the relationship between the problem of identifying coordinately dysregulated subnetworks and the well-known set-cover problem. Finally, we propose a cover-based algorithm for the identification of coordinately dysregulated subnetworks and discuss how the subnetworks identified by our algorithm can be used for the diagnosis and prognosis of complex diseases.

2.1. *Coordinately Dysregulated Subnetworks*

In the context of a specific phenotype, a group of genes that exhibit significant differential expression and whose products are connected to each other through physical and functional interactions may be useful in understanding the network dynamics of the phenotype. This is because, the patterns of (i) collective differential expression and (ii) connectivity in PPI network are derived from orthogonal sources (sample-specific mRNA expression and generic protein-protein interactions, respectively). Thus, they provide corroborating evidence indicating that the corresponding subnetwork of the PPI network may play an important role in the manifestation of phenotype. In this paper, we refer to the collective differential expression of a group of genes as *coordinate dysregulation*. A group of coordinately dysregulated genes that induce a connected subnetwork in a PPI network is thus called a *coordinately dysregulated subnetwork*.

Dysregulation of a gene with respect to a phenotype. For a set $\mathcal{V}$ of genes and $\mathcal{U}$ of samples, let $E_i \in R^{|\mathcal{U}|}$ denote the properly normalized[18] gene expression vector for gene $g_i \in \mathcal{V}$, where $E_i(j)$ denotes the relative expression of g_i in sample $s_j \in \mathcal{U}$. Assume that the phenotype vector C annotates each sample as phenotype or control, such that $C_j = 1$ indicates that sample s_j is associated with the phenotype (e.g., taken from a tumor tissue) and $C_j = 0$ indicates that s_j is a control sample (e.g., taken from a normal tissue). Then, the mutual information $I(E_i; C) = H(C) - H(C|E_i)$ of E_i and C is a measure of the reduction of uncertainty about phenotype C due to the knowledge of the expression level of gene g_i. Here, $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ denotes the Shannon entropy of discrete random variable X with support $\mathcal{X}$. The entropy $H(E_i)$ of the expression profile of gene g_i is computed by quantizing E_i properly. Clearly, $I(E_i; C)$ provides a reasonable measure of the dysregulation of g_i, since it quantifies the power of the expression level of g_i in distinguishing phenotype and control samples.

Coordinate dysregulation. Now let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a PPI network where the product of each gene $g_i \in \mathcal{V}$ is represented by a node and each edge $g_i g_j \in \mathcal{E}$ represents an interaction between the products of g_i and g_j. For a given subnetwork of $\mathcal{G}$ with set of nodes $S \subseteq \mathcal{V}$, Chuang *et al.*[16] define the *subnetwork activity* of S as $E_S = \frac{1}{\sqrt{|S|}} \sum_{g_i \in S} E_i$, that is the aggregate expression profile of the genes in S. Then, naturally, the dysregulation of subnetwork S is given by $I(E_S; C)$, which provides a measure of the reduction of uncertainty about phenotype C due to the knowledge of the aggregate expression level of the genes in S. In the following discussion, we refer to $I(E_S; C)$ as the *coordinate dysregulation* of S.

Identification of coordinately dysregulated subnetworks. Clearly, identification of subnetworks with maximal $I(E_S; C)$ is an intractable computational problem. Simple greedy approaches to this problem grow subnetworks by starting from a single protein and adding to the subnetwork the proteins in its network neighborhood. At each step, the protein in the neighborhood that maximally increases $I(E_S; C)$ is added to the subnetwork. While such algorithms are useful in identifying subnetworks with reasonably high coordinate dysregulation, they are biased toward identifying subnetworks with very few genes that exhibit significant

Pacific Symposium on Biocomputing 15:133-144(2010)

individual dysregulation. Consider, for example, a group of genes that are marginally dysregulated with respect to the phenotype, when considered individually. Assume that these genes exhibit significant dysregulation when considered together (i.e., have a large $I(E_S; C)$). This group of genes is not likely to be identified by such an algorithm since the individual contribution of each gene to the dysregulation of the subnetwork will not be apparent at any stage of the algorithm until all genes are added to the subnetwork. Indeed, in our experiments on a gene expression dataset (*GSE8671*) with 8987 genes and 32 samples of colorectal adenomas paired with those of normal mucosa, such an algorithm assigns 84% of the genes to subnetworks composed of at most two genes. However, for effective investigation of the systems biology of complex phenotypes, larger subnetworks with weaker individual, but stronger coordinate dysregulation are very interesting since they offer insights beyond what single gene markers can provide. In the following discussion, we propose a novel framework that utilizes biological insights into the dysregulation of genes at a sample-specific resolution, to develop algorithms for more effective discovery of such coordinately dysregulated subnetworks.

2.2. *Coordinate Dysregulation and Set Cover*

We now show that the problem of identifying coordinately dysregulated subnetworks can be formulated as a variation of the set-cover problem. For this purpose, consider a binary representation of the gene expression data of interest. Binary representation of gene expression is commonly utilized for several reasons, including removal of noise, algorithmic considerations, and tractable biological interpretation of identified patterns. Such approaches are shown to be effective in the context various problems, ranging from genetic network inference[19] to clustering[20] and classification.[21] There are also many algorithms for effective binarization of gene expression data.[22] For our purposes, let $\hat{E}_i$ denote the binarized expression profile of gene g_i (we discuss how we binarize a gene expression dataset in our experiments in Section 3). We say that gene g_i is *expressed* (or "on") in sample s_j if $\hat{E}_i(j) = 1$ and *not expressed* (or "off") if $\hat{E}_i(j) = 0$.

In order to illustrate the relationship between coordinate dysregulation and set-cover, we introduce various concepts that provide insights into the dysregulation of genes at a sample-specific resolution. For this purpose, we assume that the gene expression data is paired; that is, there is one-to-one correspondence between phenotype and control samples. This is indeed the case for many available gene expression datasets that monitor complex phenotypes, e.g., for each sample taken from a cancerous tissue of a patient, a control sample is also taken from the part of the tissue without the lesion. This approach controls for the noise and bias that might be introduced by the biological variability among different tissues or individuals. Formally, we assume that $|\mathcal{U}| = 2n$ is even, and for each $1 \leq j \leq n$, the pair of samples s_j and s_{j+n} are phenotype and control samples that are associated with each other (i.e., they come from the same individual or tissue). A sample instance of binary gene expression data with paired samples is shown in Figure 1(A). We can now define the positive and negative cover for a gene.

Definition 1. Positive and Negative Cover Set of a Gene. A gene g_i is said to cover a sample s_j *positively/negatively* if it is up-regulated/down-regulated in the phenotype sample with respect to control ($\hat{E}_i(j) = 1$ and $\hat{E}_i(j+n) = 0$/ $\hat{E}_i(j) = 0$ and $\hat{E}_i(j+n) = 1$). The set of samples that are covered positively/negatively by g_i is called the *positive/negative cover set* of g_i and denoted $\mathcal{P}_i = \mathcal{P}(g_i)/\mathcal{N}_i = \mathcal{N}(g_i)$.

Note here that the notion of up- or down-regulation of a gene with respect to a sample depends on the procedure for binarization of expression levels. For this reason, we systematically evaluate the effect of binarization on the performance of our algorithms in Section 3.

The positive and negative cover sets of the genes in Figure 1(A) are shown in Figure 1(C). As seen in the figure, gene B covers S2 positively since it is expressed in C2 while it is not expressed in N2. Overall, the positive cover set of gene B is given by $\mathcal{P}(B) = \{S2, S3, S4\}$ and its negative cover set is given by $\mathcal{N}(B) = \{S8, S9\}$. Observe that, since gene B is dysregulated with respect to samples S2, S3, and S4, it can be used to distinguish phenotype and control samples based on its expression in a given sample. However, clearly, the statistical power (or reliability) of gene B in distinguishing phenotype and control samples depends on the number of samples that it covers. Moreover, the samples S8 and S9, which are covered negatively by gene B

Pacific Symposium on Biocomputing 15:133-144(2010)

interfere with its power in distinguishing phenotype and control samples, since the signals provided by these two groups of samples are conflicting with each other.

Based on these observations, we postulate that genes that can distinguish different sets of samples may be complementary of each other in the manifestation of phenotype. In other words, the dysregulation of genes involved in similar processes may have similar effects on phenotype, and since such genes are expected to be functionally related, they are likely to be in close proximity of each other in a network of interactions. Consequently, if we can identify subnetworks composed of genes that together cover all samples consistently (i.e., either positively or negatively), then the products of these genes may indeed have a coordinate effect on the manifestation of the phenotype. These subnetworks may be useful as features for classification of phenotype and they may reveal targets for theurapetic intervention. Motivated by these considerations, we formulate the problem of identifying coordinately dysregulated subnetworks as one of identifying minimal groups of interacting genes that cover all samples either positively or negatively. To provide a theoretical foundation for this approach, we first show that the cardinality of the cover set of a gene is related to a sound measure of class separability, namely the mutual information between the expression profile of the gene and the phenotype vector.

Theorem 1. *For any two genes $g_i, g_j \in V$, if $|\,|\mathcal{P}_i| - |\mathcal{N}_i|\,| > |\,|\mathcal{P}_j| - |\mathcal{N}_j|\,|$, then $I(\hat{E}_i; C) > I(\hat{E}_j; C)$.*

Proof. By definition of mutual information, we have $I(\hat{E}_i; C) - I(\hat{E}_j; C) = H(C|\hat{E}_j) - H(C|\hat{E}_i)$. Therefore, it will suffice to show that $|\,|\mathcal{P}_i| - |\mathcal{N}_i|\,| > |\,|\mathcal{P}_j| - |\mathcal{N}_j|\,|$ implies $H(C|\hat{E}_i) < H(C|\hat{E}_j)$. First note that

$$H(C|\hat{E}_i) = P(\hat{E}_i = 0)H(C|\hat{E}_i = 0) + P(\hat{E}_i = 1)H(C|\hat{E}_i = 1). \tag{1}$$

We will show that $H(C|\hat{E}_i = 0)$ and $H(C|\hat{E}_i = 1)$ both decline with growing $|\,|\mathcal{P}_i| - |\mathcal{N}_i|\,|$, to conclude that $H(C|\hat{E}_i)$ also declines with growing $|\,|\mathcal{P}_i| - |\mathcal{N}_i|\,|$, since $P(\hat{E}_i = 0)$ and $P(\hat{E}_i = 1)$ are both positive.

Now, for $x, y \in \{0, 1\}$, let $n_i^{(x,y)}$ denote the number of samples with phenotype x in which the binary expression of g_i is y; e.g., $n_i^{(1,0)}$ is the number of phenotype samples in which gene g_i is not expressed. Then, clearly

$$H(C|\hat{E}_i = 0) = h(p_i^{(0)}) \text{ and } H(C|\hat{E}_i = 1) = h(p_i^{(1)}). \tag{2}$$

Here, $h(p) = -p \log p - (1 - p) \log(1 - p)$ denotes the entropy of a Bernoulli random variable with success probability p, $p_i^{(0)} = \frac{n_i^{(0,0)}}{n_i^{(0,0)} + n_i^{(1,0)}}$, and $p_i^{(1)} = \frac{n_i^{(0,1)}}{n_i^{(0,1)} + n_i^{(1,1)}}$. Let $n_i^+ = |\mathcal{P}_i|$, $n_i^- = |\mathcal{N}_i|$ and let m_i^+ / m_i^- denote the number of samples in which g_i is expressed/not expressed in both phenotype and control. Then, we can write:

$$n_i^{(0,0)} = n_i^+ + m_i^-,\ n_i^{(0,1)} = n_i^- + m_i^+,\ n_i^{(1,0)} = n_i^- + m_i^-,\ \text{and } n_i^{(1,1)} = n_i^+ + m_i^+. \tag{3}$$

Consequently,

$$p_i^{(0)} = \frac{n_i^+ + m_i^-}{n_i^+ + n_i^- + 2m_i^-}, \tag{4}$$

and therefore we have $p_i^{(0)} - 1/2 = \frac{n_i^+ - n_i^-}{2(n_i^+ + n_i^- + 2m_i^-)}$. Since $p_i^{(0)} - 1/2$ assumes its zero at $n_i^+ = n_i^-$ and its derivative with respect to n_i^+ is always positive, $|p_i^{(0)} - 1/2|$ grows with growing $|n_i^+ - n_i^-|$. Since the entropy function $h(p_i^{(0)})$ is maximized at $p_i^{(0)} = 1/2$ and declines with growing $|p_i^{(0)} - 1/2|$, we conclude that $h(p_i^{(0)})$ (similary, $h(p_i^{(1)})$) declines with growing $|\,|\mathcal{P}_i| - |\mathcal{N}_i|\,|$, completing the proof. □

This theorem establishes that the number of paired samples for which a gene is consistently up- or down-regulated is directly associated with the information its expression profile provides on the phenotype. It can be seen from the construction of the proof that this result can also be generalized to the aggregate expression profile E_S of a subnetwork S (where E_S is quantized properly). Motivated by this observation, we generalize the notion of positive and negative cover sets to subnetworks and conjecture that a subnetwork with a larger consistently positive or negative cover set provides more information on the phenotype.

Pacific Symposium on Biocomputing 15:133-144(2010)

Definition 2. Positive and Negative Cover Set of a Subnetwork. For a given subnetwork $S \subseteq \mathcal{V}$, the *positive* and *negative* cover sets of S are respectively defined as $\mathcal{P}(S) = \bigcup_{g_i \in S} \mathcal{P}_i$ and $\mathcal{N}(S) = \bigcup_{g_i \in S} \mathcal{N}_i$.

2.3. *Minimal Covering Subnetwork Problem*

We now formulate the coordinately dysregulated subnetwork identification problem as one of identifying minimal subnetworks that cover all samples either positively or negatively. Similar set-cover based approaches are also shown to be effective in feature selection.[23] Here, rather than searching for the "best" subnetwork in the entire network, we look for the "best" subnetwork associated with a given gene. This is because, from a biological perspective, it is not necessarily true that a single subnetwork that maximizes an objective criterion over the entire network will be the only relevant subnetwork. On the contrary, for various applications including identification of subnetwork markers for classification, multiple subnetworks are useful. Furthermore, in understanding the relationship between the manifestation of phenotype at different levels of cellular control (e.g., genomic sequences, gene expression, protein expression), researchers may be interested in finding subnetworks associated with known genetic markers. Indeed, in Section 3.5, we demonstrate that subnetworks associated with known genetic markers ("driver genes") of colorectal cancer (CRC) are more effective in classification of CRC as compared to subnetworks associated with random genes. Furthermore, proteomic targets that are identified based on differential protein expression, when used as seeds for identification of dysregulated subnetwork markers, are shown to provide significant insights into the systems biology of complex phenotypes.[24]

Definition 3. Minimal Covering Subnetwork Associated with a Gene. For a set of paired samples $\mathcal{U}$, a set of genes $\mathcal{V}$ with binary expression profiles $\hat{E}_i \in \{0,1\}^{|\mathcal{U}|}$, a PPI network $G = (\mathcal{V}, \mathcal{E})$ and a gene $g_i \in \mathcal{V}$, the minimal covering subnetwork associated with g_i is defined as a subnetwork $S_i \subseteq \mathcal{V}$ satisfying the following conditions:

(1) $g_i \in S_i$.
(2) S_i is a local subnetwork, i.e., $\forall\ g_j \in S_i,\ \exists\ g_k \in S_i$ such that $\delta(g_j, g_k) \leq \ell$, where $\delta(g_j, g_k)$ denotes the network distance between g_j and g_k, and ℓ denotes an adjustable threshold that specifies the desired locality of the subnetwork (if $\ell = 1$, the subnetwork is connected).
(3) S_i covers all samples either positively or negatively, i.e., $\mathcal{P}(S_i) = \mathcal{U}$ or $\mathcal{N}(S_i) = \mathcal{U}$.
(4) If $\mathcal{P}(S_i) = \mathcal{U}$ ($\mathcal{N}(S_i) = \mathcal{U}$), then $|\mathcal{N}(S_i)|$ ($|\mathcal{P}(S_i)|$) is minimum over all subnetworks that satisfy the above three conditions.
(5) S_i is minimal, i.e., $\forall g_j \in S_i$, subnetwork $S_i \setminus \{g_j\}$ does not satisfy the above conditions.

Condition (1) ensures that the subnetwork is indeed associated with the gene of interest. Condition (2) ensures that the genes in the subnetwork are functionally associated with each other. Condition (3) ensures that for each paired sample, there is at least one gene in the subnetwork that can distinguish phenotype and control. Condition (4) ensures that the noise introduced by the genes that are dysregulated in the opposite direction is minimal. Finally, condition (5) ensures that there are no redundant genes in S_i.

Note that the minimal covering subnetwork problem is similar to the *minimum connected cover* (MCC) problem introducted by Ulitsky *et al.*[25] However, there is a fundamental conceptual difference between the two problems. MCC searches for subnetworks in which multiple genes are dysregulated in each phenotype sample. On the contrary, minimal covering subnetwork problem explicitly looks for subnetworks composed of genes that are *complementary* of each other in distinguishing phenotype and control samples.

2.4. *Algorithm for the Identification of Minimal Covering Subnetworks*

There is a clear conceptual and mathematical similarity between the minimal covering subnetwork problem and the well-known set-cover problem. An instance of the set-cover problem consists of a finite set X and a family $\mathcal{F}$ of subsets of X, such that the union of all the sets in $\mathcal{F}$ constitute X. The set-cover problem asks for a minimum size subset $\mathcal{C} \subseteq \mathcal{F}$ such that the union of all sets in $\mathcal{C}$ is equal to X. Clearly, the minimal covering subnetwork problem is a special case of this NP-hard problem.[26] Here, the selection of sets in $\mathcal{C}$

Pacific Symposium on Biocomputing 15:133-144(2010)

(which corresponds to S_i) are further constrained by (i) the network locality of the genes in S_i and (ii) a second collection of associated sets, union of which is to be minimized. We here adapt a polynomial-time approximation algorithm for this well-studied problem, which works by picking, at any stage, the set that covers the maximum number of remaining uncovered elements.[27]

The proposed algorithm for the identification of minimal covering subnetworks, NetCover, is illustrated in Figure 1. For a given gene g_i, assume without loss of generality that $|\mathcal{P}_i| > |\mathcal{N}_i|$. Then, NetCover identifies the minimal covering subnetwork associated with g_i as follows:

(1) Initialize subnetwork: $S_i \leftarrow \{g_i\}$
(2) Initialize set of uncovered samples: $\mathcal{T} \leftarrow \mathcal{U} \setminus \mathcal{P}_i$
(3) Initialize set of neighboring genes: $\mathcal{Q} \leftarrow \{g_j \in \mathcal{V} : \delta(g_i, g_j) \leq \ell\}$
(4) For all genes $g_j \in \mathcal{Q}$, compute $\mathcal{P}'_j \leftarrow \mathcal{P}_j \cap \mathcal{T}$
(5) Find the genes in $\mathcal{Q}$ with maximum $|\mathcal{P}'_j|$ and let g_k be the gene among these genes with minimum $|\mathcal{N}_j|$
(6) Update subnetwork: $S_i \leftarrow S_i \cup \{g_k\}$
(7) Update set of uncovered samples: $\mathcal{T} \leftarrow \mathcal{T} \setminus \mathcal{P}'_k$
(8) Update set of neighboring genes: $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{g_j \in \mathcal{V} : \delta(g_k, g_j) \leq \ell\} \setminus \{g_k\}$
(9) If $\mathcal{T} = \emptyset$ or $\mathcal{Q} = \emptyset$, return S_i; otherwise, go to step (4)

Since the cardinality of $\mathcal{Q}$ is bounded by $|\mathcal{V}|$ and the loop (4)-(9) can be repeated at most $|\mathcal{V}|$ times, the worst-case running time of this algorithm is $O(|\mathcal{V}|^2)$. However, in our experiments on several colorectal cancer datasets, we observe that the algorithm successfully identifies covering subnetworks that are fairly small (i.e., the loop exists with $\mathcal{T} = \emptyset$ after a few iterations).

2.5. *Using Minimal Covering Subnetworks for Classification*

To quantify the discriminative potential (hence, relevance to the manifestation of phenotype) of discovered subnetworks, we use these subnetworks to build classifiers for diagnosis and prognosis of the phenotype. For this purpose, for a given gene expression dataset, we first discover the minimal covering subnetwork for all genes in the human PPI network. Then, since the minimal covering subnetworks for more than one gene might be identical, we eliminate the subnetworks that are redundant. Subsequently, we score each of the remaining subnetworks according to their coordinate dysregulation ($I(E_S, C)$). Then we select the K non-redundant subnetworks with maximum $I(E_S, C)$. Here, we consider a subnetwork redundant if it shares a gene with a subnetwork that is already selected. The number of selected subnetwork features, K, is designated as an adjustable parameter. Finally, we use the aggregate expression profiles (E_S) of these selected subnetworks to compute feature vectors for each sample. We then use these feature vectors to train and test classifiers for the prognosis and diagnosis of the phenotype of interest, as we discuss in the next section.

3. Results and Discussion

In this section, we comprehensively evaluate the performance of the proposed algorithm in the context of human colorectal cancer (CRC) and compare its performance with the greedy algorithm by Chuang *et al.*,[16] which aims to directly maximize the additive mutual information by greedily growing subnetworks. We then investigate the biological relevance of coordinately dysregulated subnetworks in the network neighborhood of genes that are implicated in CRC according to gene association studies.

3.1. *Human Colorectal Cancer (CRC)*

Colorectal Cancer (CRC) is the third most common cancer and second leading cause responsible for cancer related death in the western world.[28] CRC generally starts with a simple, benign tumor. Most often, these growths go undetected, even for years, before they develop into malignancies. Therefore, until the symptoms of cancer are developed, diagnosis of cancer is rather difficult. Monitoring of differential gene expression is therefore useful for diagnosis of cancer at early stages, as well as prognosis of the development of the tumor and therapeutic outcome. Furthermore, since most deaths are related to diagnosis of CRC in late stages, understanding of the network dynamics of CRC thorough its various stages may be very useful in development

Pacific Symposium on Biocomputing 15:133-144(2010)

of more effective therapeutic intervention strategies. To this end, if coordinately dysregulated subnetworks that are identified with respect to dysregulation at a particular stage can successfully classify disease at another stage, such subnetworks may be identified as those that are important in the development of disease. For this reason, in our experimental studies, we use a particular dataset to identify coordinately dysregulated subnetworks and use these subnetworks to develop classifiers for the diagnosis and prognosis of CRC with respect to other datasets.

3.2. *Datasets*

In our experiments, we use three CRC-related microarray datasets obtained from GEO (Gene Expression Omnibus, `http://www.ncbi.nlm.nih.gov/geo/index.cgi`). These datasets, which are referenced here by their accession number in the GEO database, are the following:

- *GSE8671* contains the expression profiles of 8,987 genes across 32 prospectively collected adenomas paired with those of normal mucosa.[29]
- *GSE10950* contains the expression profiles of 18,171 genes across 24 normal and tumor pairs.[30]
- *GSE6988* contains the expression profiles of 17,104 genes across the following tissue samples: paired tissues of 25 normal colorectal mucosa, 27 primary colorectal tumors, 13 normal liver and 27 liver metastasis, and 20 primary colorectal tumors without liver metastasis.[31]

The human protein-protein interaction data used in our experiments is obtained from the Human Protein Reference Database (HPRD `http://www.hprd.org`). This dataset contains 35023 binary interactions among 9299 proteins, as well as 1060 protein complexes consisting of 2146 proteins. The binary interactions contain *in vivo*, as well as *in vitro* interactions obtained via high-throughput screening. We integrate the binary interactions and protein complexes using a matrix model (i.e., each complex is represented as a clique of the proteins in the complex), to obtain a PPI network composed of 42781 binary interactions among 9442 proteins.

In order to reduce the effect of systematic experimental bias in high-throughput microarray experiments, the expression profile of each gene in each dataset is normalized to have a mean value of zero and standard deviation of one across all the samples. For the identification of subnetworks, these normalized expression values are then binarized by setting the largest $\alpha\%$ of the normalized expression values to 1 and all of the rest to 0. Here, α is an adjustable parameter that specifies the fraction of "expressed" genes in the dataset. This tunable binarization scheme is chosen with a view to investigating the effect of binarization on the performance of NetCover (related experimental results are provided at the end of this section). In the experimental results reported in the following discussion, α is set to 25% since this value is found to be optimal for the performance of NetCover in all experiments. Note here that binary expression levels are used only for the identification of dysregulated subnetworks via NetCover. On the other hand, coordinate dsyregulation of subnetworks is computed by quantizing the aggregate expression profiles into eight bins, since this value is found to be optimal for the performance of Chuang *et al.*'s greedy algorithm. Therefore, the performance methods other than NetCover does not depend on α.

3.3. *Experimental Design*

NetCover is implemented in Matlab. Using this implementation, two different sets of coordinately dysregulated subnetworks are generated based on the datasets *GSE8671* and *GSE10950*. In order to evaluate the classification performance of these subnetworks, extensive experiments are performed by using various types of classifiers and different classification problems. Three classification problems are considered for this purpose:

- *Diagnosis of samples in GSE10950.* Subnetworks discovered using *GSE8671* are used to predict the class (cancerous *vs.* non-cancerous) of samples in *GSE10950*.
- *Diagnosis of samples in GSE6988.* Sets of subnetworks discovered using *GSE8671* and *GSE10950* are used to predict the class (cancerous *vs.* non-cancerous) of samples in *GSE6988*.
- *Prognosis of samples in GSE6988.* Subnetworks that are discovered using *GSE8671* and *GSE10950* are used to classify the 27 colorectal tumors with liver metastasis and 20 without liver metastasis in *GSE6988* into metastatic *vs.* non-metastatic classes.

Pacific Symposium on Biocomputing 15:133-144(2010)

For each of these problems, two types of classification tests are performed:

- *Cross-classification* (CC): The classifier is trained on one dataset and tested on another dataset (note that this is not applicable to the prognosis of samples in *GSE6988*, since other datasets do not contain metastasis information).
- *Leave-one-out cross-validation* (LOOCV): For each sample in a dataset, one sample is left out and the classifier is trained using the remaining samples in that dataset, which is then used to classify the corresponding sample. The performance of the classifier is evaluated by repeating this procedure for all samples.

For each instance, two types of classifiers are used: (i) a quadratic regression model, provided by Matlab's `classify` function and (ii) a support vector machine (SVM), provided by Matlab's `svmclassify` function.

3.4. *Classification Performance in Diagnosis and Prognosis of Colorectal Cancer*

In this section, we report systematic experimental results on the diagnostic and prognostic performance of the subnetwork markers identified on *GSE10950* and *GSE8671* datasets on classifying samples in the *GSE10950* and *GSE6988* datasets. The performance of each of these sets of subnetwork markers is compared against (i) the subnetwork markers identified by Chuang *et al.*'s greedy approach[16] (also implemented in Matlab for comparison purposes), and (ii) single gene markers. NetCover takes about 10 minutes for the identification of subnetworks associated with all genes in each of *GSE10950* and *GSE8671*, while the greedy algorithm takes about 30 minutes to complete the same task. The subnetwork markers that are identified by Chuang *et al.*'s algorithm are ranked and used for classification in the same way as the subnetwork markers identified by NetCover. Similar to subnetwork markers, single gene markers are ranked based on their individual mutual information with the phenotype and the top k markers are used for classification for $1 \leq k \leq K$.

The number of (subnetwork or single gene) markers, denoted K, is used as a free parameter for building, training, and testing the performance of classifiers. For each $1 \leq K \leq 50$, the performance of the classification is measured using 'Area Under ROC Curve'(AUC) criterion. AUC is a measure of the overall performance of a classifier, which accounts for the trade-off between the precision (selectivity) and recall (sensitivity) achieved by the classifier. Here, precision is defined as the fraction of true positives among all samples classified as phenotype by the classifier, while recall is defined as the fraction of true positives among all true phenotype samples. AUC is a measure of the average precision across varying values of recall (or vice versa) and an AUC of 1.0 indicates that the classifier provides perfect precision without sacrificing recall. Therefore, a value of AUC closer to 1.0 indicates better performance of a classifier. Note that, when prediction of the classifier does not depend on an adjustable threshold (i.e., there is only a single point on the ROC curve), the AUC value returned by Matlab is equal to the arithmetic mean of precision and recall. In those cases, other measures such as the harmonic mean of precision and recall (known as F-measure) are considered more reliable; however, in our experiments, AUC and F-measure provided similar results while comparing different classifiers.

The overall performance of the subnetwork and single gene markers is evaluated in Figure 2. In this figure, to systematically evaluate the performance of the subnetwork and single gene markers across different values of K, we report (i) average AUC across all values of K, ranging from 1 to 50 (average performance of the classifier) and (ii) maximum AUC across this range of K (optimal performance of the classifier that can be obtained by adjusting the number of markers accurately). As seen in the figure, in all three experiments (diagnosis of *GSE10950*, diagnosis and prognosis of *GSE6988*), subnetworks identified by NetCover demonstrate better classification performance compared to subnetworks discovered by greedy algorithm and single gene markers for both classification procedures and classifier types used. This observation indicates that NetCover discovers subnetworks that are more relevant in terms of the network dynamics of the progression of CRC, in that subnetworks identified in one dataset can distinguish samples in another data set better, as compared to single gene markers or subnetworks identified by another state-of-the-art algorithm.

3.5. *Coordinately Dysregulated Subnetworks Associated with Hallmarks of CRC*

In recent years, comparative genomic studies of CRC have revealed many genes with mutations that may be associated with colorectal cancer. These "Hallmarks of CRC" include *APC, CTNNB1, KRAS, HRAS,*

Pacific Symposium on Biocomputing 15:133-144(2010)

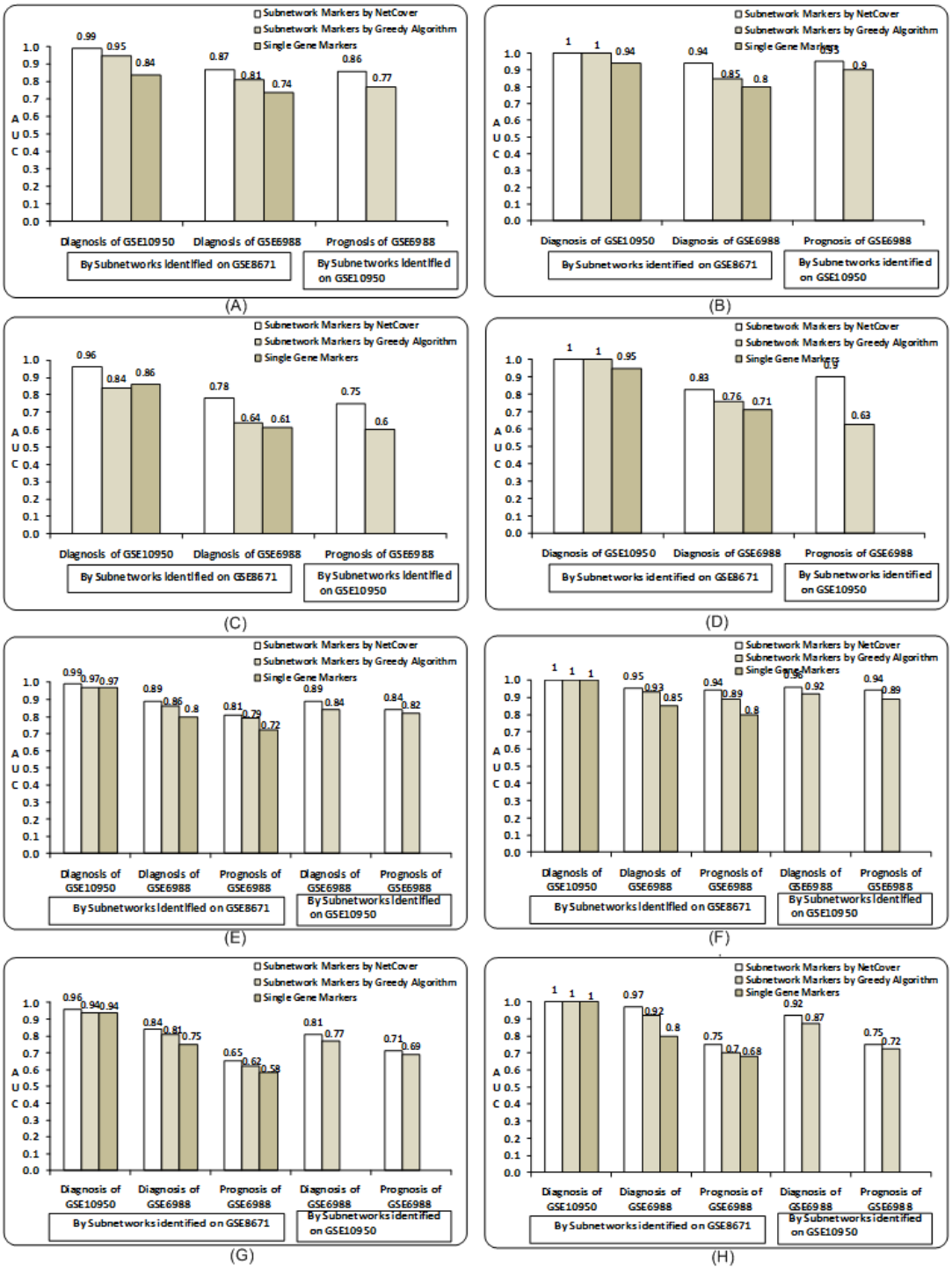

Fig. 2. Comparison of the classification performance of subnetwork markers identified by NetCover against subnetwork markers identified by Chuang *et al.*'s algorithm and single gene markers. In each figure, the AUC for diagnosis and prognosis of samples in *GSE8671*, *GSE10950*, and *GSE6988* datasets is shown for different combinations of classification and performance evaluation methods. For each configuration, average and maximum AUC (i.e., the best classification that can be obtained by accurately adjusting the number of features) are measured across number of markers ranging from 1 to 50. (A) Average, (B) maximum AUC for cross-classification (CC) using support vector machines (SVM). (C) Average, (D) maximum AUC for CC using Quadratic Regression (QR). (E) Average, (F) maximum AUC for leave-one-out cross-validation (LOOCV) using SVM. (G) Average, (H) maximum AUC for LOOCV using QR.

Pacific Symposium on Biocomputing 15:133-144(2010)

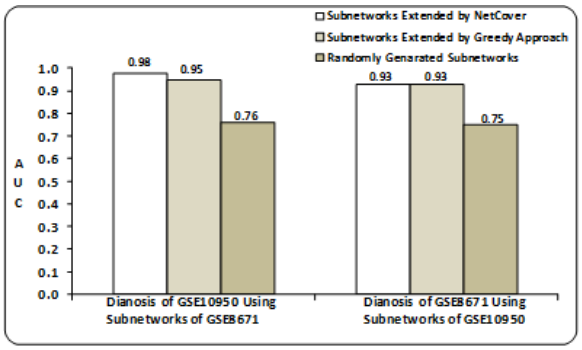

Fig. 3. Classification performance of coordinately dysregulated subnetworks associated with genes that are known to be susceptible for CRC, as compared to subnetworks associated with randomly selected genes.

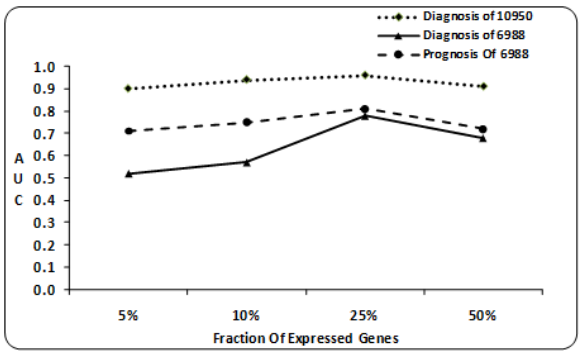

Fig. 4. Effect of binarization on the performance of NetCover. The graph shows the area under ROC curve for three instances with respect to fraction (α) of genes that are considered as "expressed".

SMAD4, TGFBR2, MYC, SRC and *DCC*.[28] To investigate whether the neighborhood of these genes contain coordinately dysregulated subnetworks that may be relevant in the manifestation of CRC, we discover coordinately dysregulated subnetworks associated with these genes using both NetCover and Chuang *et al.*'s greedy algorithm, on datasets *GSE8671* and *GSE10950*. Then, we use the subnetworks identified on each dataset to classify the samples in the other dataset. In Figure 3, the AUC provided by these classifiers is compared against the AUC provided by the same number of subnetworks associated with randomly selected genes. As can be seen in the figure, the subnetworks associated with the hallmarks of CRC demonstrate very high classification accuracy for both datasets. Furthermore, NetCover identifies larger subnetworks with better classification performance, as compared to the greedy algorithm. These results indicate that the coordinately dysregulated subnetworks associated with the CRC driver genes can be used to understand the effect of the mutations in the driver genes on the dysregulation of the other genes within the context of network dynamics. These subnetworks may also be useful in discovering proteins that are related to disease-causing genes but may not exhibit significant individual dysregulation.[32]

3.6. *Effect of Binarization*

In order to investigate the effect of binarization on the performance of the proposed algorithm and to choose the cutoff level at which to binarize the gene expression data, we fix α, which specifies the fraction of "expressed" genes in the dataset, at values of 5%, 10%, 25% and 50%. Then we use NetCover to discover the dysregulated subnetworks at each level of α and perform systematic experiments whose results are reported in Figure 4. It can be observed from the figure that the subnetworks discovered at $\alpha = 25\%$ show best performance for all instances. For this reason, we choose the cutoff level at which 25% of the genes are "expressed" to binarize the gene expression data.

4. Conclusion

In this paper, we propose a novel computational approach for identifying coordinately dysregulated subnetworks in a complex phenotype like cancer. Our algorithm integrates protein-protein interaction data with clinical gene expression data to capture the coordinated dysregulation of multiple interacting genes. Application of our algorithm on human colorectal cancer (CRC) datasets shows its potential in identifying subnetworks with high relevance to disease progression. However, our current algorithm is defined in terms of paired gene expression data; that is, datasets where there is a one-to-one relationship between control and phenotype samples. Extension of this approach to unpaired datasets will broaden the application of the proposed framework. Furthermore, the subnetwork identification algorithm implemented here is based on an approximation algorithm for the set-cover problem. Development of algorithms for more effective search of the subnetwork space may lead to identification of more biologically relevant subnetworks. Finally, detailed investigation of the subnetworks identified by NetCover may shed light into the network dynamics of CRC.

Pacific Symposium on Biocomputing 15:133-144(2010)

Acknowledgments

We would like to thank Rod Nibbe and Mark Chance for useful discussions on the systems biology of colorectal cancer. This work was supported, in part, by the National Institutes of Health Grant, UL1-RR024989 Supplement, from the National Center for Research Resources (Clinical and Translational Science Awards).

References

1. A. M. Glazier, J. H. Nadeau and T. J. Aitman, *Science* **298**, 2345(Dec 2002).
2. T. R. Golub, D. K. Slonim, P. Tamayo, M. G. C. Huard, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science* **286**, 531(Oct 1999).
3. C. M. Perou, T. Srlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, ystein Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lnning, A.-L. Brresen-Dale, P. O. Brown and D. Botstein, *Nature* **406**, 747(Aug 2000).
4. D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. T. Michelle, L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer and S. Hanash, *Nat. Med.* **8**, 816 (2002).
5. J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A. M. DeMarzo, R. Tibshirani, D. Botstein, P. O. Brown, J. D. Brooks and J. R. Pollack, *PNAS* **101**, 811(Jan 2004).
6. S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin and A. M. Chinnaiyan, *Nature* **412**, 822 (2001).
7. D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub and W. R. Sellers, *Cancer Cell* **1**, 203(Mar 2002).
8. E. LaTulippe, J. Satagopan, A. Smith, H. Scher, P. Scardino, V. Reuter and W. L. Gerald, *Cancer Res.* **62**, 4499(Aug 2002).
9. S. Ramaswamy, K. N. Ross, E. S. Lander and T. R. Golub, *Nat. Genet.* **33**, 49(Jan 2002).
10. R. Jansen, D. Greenbaum and M. Gerstein, *Genome Res.* **12**, 37 (2002).
11. E. Segal, H. Wang and D. Koller, *Bioinformatics* **19**, I264 (2003).
12. T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics* **18**, 233 (2002).
13. D. Rajagopalan and P. Agarwal, *Bioinformatics* **21**, 788 (2005).
14. L. Cabusora, E. Sutton, A. Fulmer and C. Forst, *Bioinformatics* **21**, 2898 (2005).
15. S. Nacu, R. Critchley-Thorne, P. Lee and S. Holmes, *Bioinformatics* **23**, 850 (2007).
16. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee and T. Ideker, *Molecular Systems Biology* **3** (2007).
17. D. Anastassiou, *Molecular Systems Biology* **3** (2007).
18. J. Quackenbush, *Nat Genet* **32 Suppl**, 496(December 2002).
19. T. Akutsu, S. Miyano and S. Kuhara, *Pacific Symposium on Biocomputing.* , 17 (1999).
20. M. Koyutürk, W. Szpankowski and A. Grama, Biclustering gene-feature matrices for statistically significant dense patterns, in *in: IEEE Computer Society Bioinformatics Conf*, 2004.
21. T. Akutsu and S. Miyano, Selecting informative genes for cancer classification using gene expression data, in *Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 2001.
22. I. Shmulevich and W. Zhang, *Bioinformatics* **18**, 555 (2002).
23. M. Dash, Feature selection via set cover, in *KDEX '97: Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop*, (IEEE Computer Society, Washington, DC, USA, 1997).
24. R. K. Nibbe, R. Ewing, L. Myeroff, M. Markowitz and M. Chance, *Mol Cell Prot* **9**, 827 (2009).
25. I. Ulitsky, R. M. Karp and R. Shamir, Detecting disease-specific dysregulated pathways via analysis of clinical expressio profiles, in *Proceedings of 12th Int'l Conf. Research in Comp. Molecular Biology (RECOMB'08)*, 2008.
26. M. R. Garey and D. S. Johnson, in *Computers and Intractability, A Guide to the Theory of NP-Completeness*, (W.H. Freeman, 1979)
27. V. Chvatal, *Mathematics of Operations Research* **4**, 233 (1979).
28. F. Macdonald, C. H. J. Ford and A. G. Casson, Colorectal cancer, in *Molecular Biology of Cancer*, 2005
29. J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Makee, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers and G. Marra, *Molecular Cancer Res.* **5(12)**, p. 1263(Dec 2007).
30. X. Jiang, J. Tan, J. Li, S. Kivime, X. Yang, L. Zhuang, P. L. Lee, M. T. Chan, L. W. Stanton, E. T. Liu, B. N. Cheyette and Q. Yu, *Cancer Cell* **13(6)**, 529(Jun 2008).
31. D. H. Ki, H. C. Jeung, C. H. Park, S. H. Kang, G. Y. Lee, W. S. Lee, N. K. Kim, H. C. Chung and S. Y. Rha, *International Journal of Cancer* **121**, 2005 (2007).
32. N. Turner, A. Tutt and A. Ashworth, *Nat Rev Cancer* **4**, 814 (2004).