JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 3, 2011

Pp. 263–281
DOI: 10.1089/cmb.2010.0269

Subnetwork State Functions Define Dysregulated Subnetworks in Cancer

SALIM A. CHOWDHURY,[1] ROD K. NIBBE,[2,4] MARK R. CHANCE,[3,4]
and MEHMET KOYUTÜRK [1,4]

ABSTRACT

Emerging research demonstrates the potential of protein-protein interaction (PPI) networks in uncovering the mechanistic bases of cancers, through identification of interacting proteins that are coordinately dysregulated in tumorigenic and metastatic samples. When used as features for classification, such coordinately dysregulated subnetworks improve diagnosis and prognosis of cancer considerably over single-gene markers. However, existing methods formulate coordination between multiple genes through additive representation of their expression profiles and utilize fast heuristics to identify dysregulated subnetworks, which may not be well suited to the potentially combinatorial nature of coordinate dysregulation. Here, we propose a combinatorial formulation of coordinate dysregulation and decompose the resulting objective function to cast the problem as one of identifying subnetwork state functions that are indicative of phenotype. Based on this formulation, we show that coordinate dysregulation of larger subnetworks can be bounded using simple statistics on smaller subnetworks. We then use these bounds to devise an efficient algorithm, CRANE, that can search the subnetwork space more effectively than existing algorithms. Comprehensive cross-classification experiments show that subnetworks identified by CRANE outperform those identified by additive algorithms in predicting metastasis of colorectal cancer (CRC).

Key words: combinatorial optimization, computational molecular biology, machine learning.

1. INTRODUCTION

RECENT ADVANCES IN HIGH-THROUGHPUT SCREENING TECHNIQUES enable studies of complex phenotypes in terms of their associated molecular mechanisms. While genomic studies provide insights into genetic differences that relate to certain phenotypes, functional genomics (e.g., gene expression, protein expression) helps elucidate the variation in the activity of cellular systems (Schadt, 2005). However, cellular systems are orchestrated through combinatorial organization of thousands of biomolecules (Papin et al., 2005). This complexity is reflected in the diversity of phenotypic effects, which generally present themselves as weak signals in the expression profiles of single molecules. For this reason, researchers increasingly focus

[1]Department of Electrical Engineering and Computer Science, [2]Department of Pharmacology, [3]Department of Physiology and Biophysics, and [4]Center of Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, Ohio.

on identification of multiple markers that together exhibit differential expression with respect to various phenotypes (Ideker and Sharan, 2008; Rich et al., 2005).

1.1. Network-based approaches to identification of multiple markers

High-throughput protein-protein interaction (PPI) data (Ewing et al., 2007) provide an excellent substrate for network-based identification of multiple interacting markers. Network-based analyses of diverse phenotypes show that products of genes that are implicated in similar phenotypes are clustered together into "hot spots" in PPI networks (Goh et al., 2007; Rhodes and Chinnaiyan, 2005). This observation is exploited to identify novel genetic markers based on network connectivity (Franke et al., 2006; Karni et al., 2009; Lage et al., 2007). For the identification of differentially expressed subnetworks with respect to GAL80 deletion in yeast, Ideker et al. (2002) propose a method that is based on searching for connected subgraphs with high aggregate significance of individual differential expression. Variants of this method are shown to be effective in identifying multiple genetic markers in prostate cancer (Guo et al., 2007), melanoma (Nacu et al., 2007), diabetes (Liu et al., 2007), and others (Cabusora et al., 2005; Patil and Nielsen, 2005; Scott et al., 2005).

1.2. Coordinate/synergistic dysregulation

Network-based approaches are further elaborated to capture coordinate dysregulation of interacting proteins at a sample-specific resolution (Chowdhury and Koyutürk, 2010). Ulitksy et al. (2008) define dysregulated pathways as subnetworks composed of products of genes that are dysregulated in a large fraction of phenotype samples. Chuang et al. (2007) define subnetwork activity as the aggregate expression of genes in the subnetwork, quantify the dysregulation of a subnetwork in terms of the mutual information between subnetwork activity and phenotype, and develop fast algorithms to identify subnetworks that exhibit significant dysregulation. Subnetworks identified by this approach are also used as features for classification of breast cancer metastasis, providing significant improvement over single-gene markers (Chuang et al., 2007). Nibbe et al. (2009, 2010) show that this notion of coordinate dysregulation is also effective in integrating protein and mRNA expression data to identify important subnetworks in colorectal cancer (CRC). Anastassiou (2007) introduces the concept of synergy to delineate the complementarity of multiple genes in the manifestation of phenotype. While identification of multiple genes with synergistic dysregulation is intractable (Anastassiou, 2007), important insights can still be gained through pairwise assessment of synergy (Watkinson et al., 2008).

1.3. Contributions of this study

Despite significant advances, existing approaches to the identification of coordinately dysregulated subnetworks have important limitations, including the following: (i) additive formulation of subnetwork activity can only highlight the coordinate dysregulation of interacting proteins that are dysregulated in the same direction, overlooking the effects of inhibitory and other complex forms of interactions; (ii) simple heuristics that make greedy decisions may not be able to adequately capture the coordination between multiple genes that provide weak individual signals. In this article, with a view to addressing these challenges, we develop a novel algorithm, CRANE, for the identification of Combinatorially dysRegulAted subNEtworks. The contributions of the proposed computational framework include the following:

- We formulate coordinate dysregulation combinatorially, in terms of the mutual information between *subnetwork state functions* (specific combinations of quantized mRNA expression levels of proteins in a subnetwork) and phenotype (as opposed to additive *subnetwork activity*).
- We decompose combinatorial coordinate dysregulation into individual terms associated with individual state functions, to cast the problem as one of identifying state functions that are *informative* about the phenotype.
- Based on this formulation, we show that the information provided on phenotype by a state function can be bounded from above using statistics of subsets of this subnetwork state. Using this bound, we develop bottom-up enumeration algorithms that can effectively prune out the subnetwork space to identify informative state functions efficiently.
- We use subnetworks identified by the proposed algorithms to train neural networks for classification of phenotype, which are better suited to modeling the combinatorial relationship between the expression

levels of genes in a subnetwork, as compared to classifiers that require aggregates of the expression profiles of genes as features (e.g., Support vector machines [SVMs]).

We describe these algorithmic innovations in detail in Section 2.

1.4 Results

We implement CRANE in Matlab and perform comprehensive cross-classification experiments for prediction of metastasis in CRC. These experiments show that subnetworks identified by the proposed framework outperform subnetworks identified by additive algorithms in terms of accuracy of classification. We then conduct comprehensive experiments to evaluate the effect of parameters on the performance of CRANE. We also investigate the highly informative subnetworks in detail to assess their potential in highlighting the mechanisms of metastasis in human CRC. We present these results in Section 3 and conclude our discussion in Section 4.

2. METHODS

In the context of a specific phenotype, a group of genes that exhibit significant differential expression and whose products interact with each other 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 independent data 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 article, we refer to the collective differential expression of a group of genes as *coordinate dysregulation*. We call a group of coordinately dysregulated genes that induce a connected subnetwork in a PPI network a *coordinately dysregulated subnetwork.* The terminology and notation in this article are described in Table 1.

2.1. 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 (Quackenbush, 2002) 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 metastatic sample) and $C_j = 0$ indicates that s_j is a control sample (e.g., taken from a non-metastatic tumor sample). 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 $\boldsymbol{X}$ 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.

2.2. Additive 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 subnetwork of $\mathcal{G}$ with set of nodes $\mathcal{S} \subseteq \mathcal{V}$, Chuang et al. (2007) define the *subnetwork activity* of S as $E_{\mathcal{S}} = \sum_{g_i \in \mathcal{S}} E_i / \sqrt{|\mathcal{S}|}$, i.e., the aggregate expression profile of the genes in S. Then, the dysregulation of $\mathcal{S}$ is given by $I(E_{\mathcal{S}}; C)$, which is a measure of the reduction in uncertainty on phenotype C, due to knowledge of the aggregate expression level of all genes in S. In the following discussion, we refer to $I(E_{\mathcal{S}}; C)$ as the *additive coordinate dysregulation* of S.

2.3. Combinatorial coordinate dysregulation

Additive coordinate dysregulation is useful for identifying subnetworks that are composed of genes dysregulated in the same direction (either up- or down-regulated). However, interactions among genes and proteins can also be inhibitory (or more complex), and the dysregulation of genes in opposite directions can

TABLE 1. SUMMARY OF NOTATIONS

Symbol	*Meaning*
$\mathcal{V}$	Set of genes
$\mathcal{U}$	Set of samples
g_i	Individual member (i.e. gene product) of $\mathcal{V}$
$p\ (x)$	Probability that random variable X is equal to x
$\mathcal{G}$	Protein protein interaction network
$\mathcal{S}$	Subnetwork of $\mathcal{G}$
E_i	Normalized gene expression profile of g_i
$\hat{E}_i$	Binarized expression profile of g_i
$E_i\ (j)$	Relative expression of g_i in sample s_j
$E_{\mathcal{S}}$	*Subnetwork activity of* $\mathcal{S}$
$F_{\mathcal{S}}$	Combination of binary expression states of the genes in $\mathcal{S}$
$f_{\mathcal{S}}$	State function (observation of random variable $F_{\mathcal{S}}$)
C	Phenotype vector
C_j	Phenotype associated with sample s_j
$H\ (C)$	Shannon entropy of C
$H\ (C\|E_i)$	Entropy of C after knowledge of E_i
$I\ (E_i; C)$	Mutual information of E_i and C
$I\ (E_{\mathcal{S}}; C)$	*Additive coordinate dysregulation of* $\mathcal{S}$
$I\ (F_{\mathcal{S}}; C)$	*Combinatorial coordinate dysregulation of* $\mathcal{S}$
$J\ (f_{\mathcal{S}}; C)$	Information provided by $f_{\mathcal{S}}$ on C
$f_{\mathcal{S}} \sqsupseteq f_{\mathcal{R}}$	Superset, i.e. $\mathcal{R} \subseteq \mathcal{S}$ and $f_{\mathcal{R}}$ maps each gene in $\mathcal{R}$ to an expression level that is identical to the mapping provided by $f_{\mathcal{S}}$
$J_{\text{bound}}\ (f_{\mathcal{S}}, C)$	The maximum achievable $J(f_{\mathcal{S}}; C)$ for any $f_{\mathcal{R}} \sqsupseteq f_{\mathcal{S}}$
j^*	The threshold on information provided by a state function to be considered as informative
b	The number of extensions recorded for a state function during the bottom-up search for informative state functions
d	The maximum height of the search tree during bottom up enumeration of state functions

also be coordinated, as illustrated in Figure 1. Combinatorial formulation of coordinate dysregulation may be able to better capture such complex coordination patterns.

To define combinatorial coordinate dysregulation, we consider binary representation of gene expression data. 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 (Akutsu et al., 1999) to clustering (Koyutürk et al., 2004) and classification (Akutsu and Miyano,

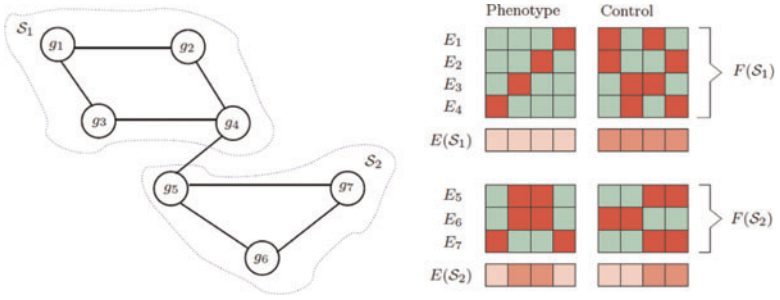

FIG. 1. Additive versus combinatorial coordinate dysregulation. Genes (g) are shown as nodes; interactions between their products are shown as edges. Expression profiles (E) of genes are shown by colormaps. Dark red indicates high expression (H); light green indicates low expression (L). None of the genes can differentiate phenotype and control samples individually. Aggregate *subnetwork activity* (average expression) for each subnetwork is shown in the row below its gene expression matrix. The aggregate activity of $\mathcal{S}_1$ can perfectly discriminate phenotype and control, but the aggregate activity of $\mathcal{S}_2$ cannot discriminate at all. For each subnetwork $\mathcal{S}_1$ and $\mathcal{S}_2$, each column of the gene expression matrix specifies the *subnetwork state* in the corresponding sample. The states of both subnetworks can perfectly discriminate phenotype and control (for $\mathcal{S}_2$, up-regulation of g_7 alone or g_5 and g_6 together indicates phenotype; we say *state functions* LLH and HHL are indicative of phenotype).

2001). Ulitsky et al. (2008) also use binary representation of differential expression to identify dysregulated pathways with respect to a phenotype. There are also many algorithms for effective binarization of gene expression data (Shmulevich and Zhang, 2002).

For our purposes, let $\hat{E}_i$ denote the binarized expression profile of gene g_i. We say that gene g_i has *high expression* in sample s_j if $\hat{E}_i(j) = H$ and *low expression* if $\hat{E}_i(j) = L$. Then, the *combinatorial coordinate dysregulation* of subnetwork $\mathcal{S}$ is defined as

$$I(F_{\mathcal{S}}; C) = H(C) - H(C|\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_m), \tag{1}$$

where $F_{\mathcal{S}} = \{\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_m\} \in \{L, H\}^m$ is the random variable that represents the combination of binary expression states of the genes in $\mathcal{S}$ and $m = |\mathcal{S}|$.

The difference between additive and combinatorial coordinate dysregulation is illustrated in Figure 1. Anastassiou (2007) also incorporates this combinatorial formulation to define the synergy between a pair of genes as $\psi(g_1, g_2) = I(\hat{E}_1, \hat{E}_2; C) - (I(\hat{E}_1; C) + I(\hat{E}_2; C))$. Generalizing this formulation to the synergy between multiple genes, it can be shown that identification of multiple genes with synergistic dysregulation is an intractable computational problem (Anastassiou, 2007). Here, we define combinatorial coordinate dysregulation as a more general notion than synergistic dysregulation, in that coordinate dysregulation is defined based solely on collective differential expression, whereas synergy explicitly looks for genes that cannot individually distinguish phenotype and control samples.

Subnetworks that exhibit combinatorial coordinate dysregulation with respect to a phenotype may shed light into the mechanistic bases of that phenotype. However, identification of such subnetworks is intractable, and due to the combinatorial nature of the associated objective function ($I(F_{\mathcal{S}}; C)$), simple heuristics may not suit well to this problem. This is because, as also demonstrated by the example in Figure 1, it is not straightforward to bound the combinatorial coordinate dysregulation of a subnetwork in terms of the individual dysregulation of its constituent genes or coordinate dysregulation of its smaller subnetworks. Motivated by these considerations, we propose to decompose the combinatorial coordinate dysregulation of a subnetwork into individual subnetwork state functions and show that information provided by state functions of larger subnetworks can be bounded using statistics of their smaller subnetworks.

2.4. Subnetwork state functions informative of phenotype

Let $f_{\mathcal{S}} \in \{\mathrm{H}, \mathrm{L}\}^m$ denote an observation of the random variable $F_{\mathcal{S}}$, i.e., a specific combination of the expression states of the genes in $\mathcal{S}$. By definition of mutual information, we can write the combinatorial coordinate dysregulation of $\mathcal{S}$ as

$$I(F_{\mathcal{S}}; C) = \sum_{f_{\mathcal{S}} \in \{\mathrm{H}, \mathrm{L}\}^m} J(f_{\mathcal{S}}; C) \tag{2}$$

where

$$J(f_{\mathcal{S}}; C) = p(f_{\mathcal{S}}) \sum_{c \in \{0, 1\}} p(c|f_{\mathcal{S}}) \log \left(p(c|f_{\mathcal{S}})/p(c)\right). \tag{3}$$

Here, $p(x)$ denotes $P(X = x)$, that is the probability that random variable X is equal to x (similarly, $p(x|y)$ denotes $P(X = x|Y = y)$). In biological terms, $J(f_{\mathcal{S}}; C)$ can be considered a measure of the information provided by subnetwork *state function* $f_{\mathcal{S}}$ on phenotype C. Therefore, we say a state function $f_{\mathcal{S}}$ is *informative* of phenotype if it satisfies the following conditions:

- $J(f_{\mathcal{S}}; C) \geq j^*$, where j^* is an adjustable threshold.
- $J(f_{\mathcal{S}}; C) \geq J(f_{\mathcal{R}}; C)$ for all $f_{\mathcal{R}} \sqsubseteq f_{\mathcal{S}}$. Here, $f_{\mathcal{R}} \sqsubseteq f_{\mathcal{S}}$ denotes that $f_{\mathcal{R}}$ is a substate of state function $f_{\mathcal{S}}$, that is $\mathcal{R} \subseteq \mathcal{S}$ and $f_{\mathcal{R}}$ maps each gene in $\mathcal{R}$ to an expression level that is identical to the mapping provided by $f_{\mathcal{S}}$.

Here, the first condition ensures that the information provided by the state function is considered high enough with respect to a user-defined threshold. The second condition ensures that informative state functions are non-redundant, that is, a state function is considered informative only if it provides more information on the phenotype than any of its substates can. This restriction ensures that the expression of each gene in the subnetwork provides additional information on the phenotype, capturing the synergy

between multiple genes to a certain extent. For a given set of phenotype and control samples and a reference PPI network, the objective of our framework is to identify all informative state functions.

The following theorem shows that for any $\mathcal{S} \subseteq \mathcal{V}$, $0 \leq J(f_{\mathcal{S}}; C) \leq \max\{-q \log p(c), -(1-q)\log(1-q)\} = j_{\max}(q)$ where q denotes the fraction of phenotype samples among all available samples.

Theorem 1. *For a given gene expression dataset, let the fraction of phenotype samples be* $q = p(1) = P(C=1)$. *Then, for any subnetwork* $\mathcal{S} \subseteq \mathcal{V}$,

$$0 \leq J(f_{\mathcal{S}}, C) \leq \max\{-q \log q, -(1-q)\log(1-q)\}. \tag{4}$$

We use the following conventions for notational convenience:

- z denotes $p(f_{\mathcal{S}}) = P(F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$ in a given sample.
- s denotes $p(1|f_{\mathcal{S}}) = P(C=1|F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability that a sample with state $f_{\mathcal{S}}$ for the genes in $\mathcal{S}$ is associated with the phenotype of interest.

Proof. Assume that q and z are fixed. Then we can write $J(f_{\mathcal{S}}; C)$ as a function of s:

$$J(s) = z(s \log(s/q) + (1-s)\log((1-s)/(1-q))). \tag{5}$$

Taking the derivative of this function with respect to s, we obtain

$$J'(s) = z \log \frac{s(1-q)}{q(1-s)}. \tag{6}$$

Observe that $J'(s)$ assumes its zero at $s = q$. Furthermore, for $s > q$, since $s/q > 0$ and $(1-s)/(1-q) < 0$, $J'(s)$ is always positive and J is an increasing function of s. Similarly, for $s < q$, $J'(s)$ is always negative and J is a decreasing function of s. Consequently, $J(s)$ is always non-negative and it assumes its maximum at one of the boundaries of the range of values that s can take. Therefore, for fixed q, if we bound $J(s)$ at the boundaries that are enforced by z, we can write the bound on J as a function of z. The maxima of this function over all values of z will provide a bound on J over all possible values of z and s for fixed q. We analyze the cases $z \geq q$ and $z \leq q$ separately.

Case A: $z \geq q$, that is the state function is observed at least as commonly as the phenotype of interest. In this case, since the number of phenotype samples in which the state function is observed can be at most equal to the number of all phenotype samples, we have $s \leq q/z$. On the other hand, if $z \leq 1-q$, then it is possible that none of the samples that exhibit the state function are associated with the sample, and therefore $s \geq 0$. Finally, when $z \leq 1-q$ (which is only possible if $z \geq 1/2$, the s will be minimized if all samples that are not associated with the phenotype exhibit the state function, and therefore we have $s \geq 1-(1-q)/z$. Consequently, we have three boundary cases for s:

1. $s = q/z$, subject to $q \leq z \leq 1$.
2. $s = 0$, subject to $q \leq z \leq 1-q$.
3. $s = 1-(1-q)/z$, subject to $\max\{q, 1-q\} \leq z \leq 1$.

We consider each of these boundary cases separately.

Case A1: Letting $s = q/z$ in (5), we obtain

$$J_{A1}(z) = z\left(\frac{q}{z}\log\frac{1}{z} + \frac{z-q}{z}\log\frac{z-q}{z(1-q)}\right), \tag{7}$$

and therefore $J_{A1}(z) = (z-q)\log((z-q)/(1-q)) - z\log z$. Consequently, $J'_{A1}(z) = \log((z-q)/(z(1-q))) \leq 0$ for $q \leq z \leq 1$ and therefore $J_{A1}(z) \leq J(q) = -q \log q$, proving the bound for this case.

Case A2: Letting $s = 0$ in (5), we obtain $J_{A2}(z) = -z\log(1-q)$ and therefore $J_{A2}(z) \leq J(1-q) = (1-q)\log(1-q)$ for $q \leq z \leq 1-q$, proving the bound for this case.

Case A3: Letting $s = 1-(1-q)/z$ in (5), we obtain

$$J_{A3}(z) = z\left(\frac{q+z-1}{z}\log\frac{q+z-1}{qz} + \frac{1-q}{z}\log\frac{1}{z}\right) \tag{8}$$

and therefore $J_{A3}(z) = (q+z-1)\ \log((q+z-1)/q) - z\ \log\ z$. Consequently, $J'_{A3}(z) = \log((q+z-1)/qz)$. $J'_{A3}(z)$ assumes its zero at $z=1$, corresponding to a minimum at $J(1)=0$. Therefore, if $q \le 1-q$, then $J_{A3}(z)$ attains its maximum at $z=1-q$, which gives $J_{A3}(z) \le J_{A3}(q) = -(1-q)\ \log(1-q)$. Otherwise ($q > 1-q$ and hence $q > 1/2$), $J_{A3}(z)$ attains its maximum at $z=q$, which gives $J_{A3}(z) \le J_{A3}(1-q) = (2q-1)\ \log((2q-1)/q) - (1-q)\ \log(1-q) \le -(1-q)\ \log(1-q)$ since $(2q-1)/q \le 1$ for $1/2 \le q \le 1$. This proves the bound for this case.

Case B: $z \le q$, that is the state function is observed at most as commonly as the phenotype of interest. In this case, s can attain the value 1 if all samples that exhibit the state function are associated with the phenotype of interest, thus $s \le 1$. On the other hand, for $z \le 1-q$, s can be as low as 0 if all samples that exhibit the state function are samples that are not associated with the phenotype. Finally, if $z \ge 1-q$, then s has to be at least $1-(1-q)/z$ since at most this fraction of samples that exhibit the state function can be samples that are not associated with the phenotype. Consequently, we have three boundary cases for s:

1. $s=1$, subject to $0 \le z \le q$.
2. $s=0$, subject to $0 \le z \le \min\{1-q, q\}$.
3. $s=1-(1-q)/z$, subject to $1-q \le z \le q$.

We consider each of these boundary cases separately.

Case B1: Letting $s=1$ in (5), we obtain $J_{B1}(z) = -z\ \log\ q$ and therefore $J_{B1}(z) \le -q\ \log\ q$ for $0 \le z \le q$, proving the bound for this case.

Case B2: Letting $s=0$ in (5), we obtain $J_{B2}(z) = -z\ \log(1-q)$ and therefore $J_{B2}(z) \le -(1-q)\ \log(1-q)$ for $q \le z \le 1-q$, proving the bound for this case.

Case B3: Observe that $J_{B3}(z) = J_{A3}(z)$. As we know from case A3, $J_{B3}(z)$ is a decreasing function of z and $J_{A3}(1-q) \le -(1-q)\ \log(1-q)$, so $J_{B3}(1-q) \le -(1-q)\ \log(1-q)$, proving the bound for this case. ■

Based on this result, we allow the user to specify a threshold j^{**} in the range [0, 1] in practice and adjust it as $j^* = j^{**} j_{\max}(p(c))$, to make the scoring criterion interpretable and uniform across all datasets.

2.5 Algorithms for the identification of informative state functions

Since the space of state functions is very large, the problem of discovering all informative state functions is intractable. Here, we address this challenge by utilizing a bound on the value of J to effectively prune the search space. Our approach is inspired by a similar result by Smyth and Goodman (1992) on information-theoretic identification of association rules in databases. In the following theorem, we show that the information that can be provided by all superstates of a given state function can be bounded based on the statistics of that state function, without any information about the superstate.

Theorem 2. *Consider a subnetwork $\mathcal{S} \subseteq \mathcal{V}$ and associated state function $f_{\mathcal{S}}$. For any $f_{\mathcal{R}} \sqsupseteq f_{\mathcal{S}}$ the following bound holds:*

$$J(f_{\mathcal{R}}; C) \le p(f_{\mathcal{S}}) \max_{c \in \{0,1\}} \left\{ p(c|f_{\mathcal{S}}) \log \frac{1}{p(c)} \right\} = J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{9}$$

The proof of this theorem is based on a more general result by Smyth and Goodman (1992) in the context of association rule mining. We first prove two lemmas necessary for the proof of Theorem 2.

Lemma 1. *For* $0 \le x < a < b \le 1$, $\dfrac{(a-x)}{(b-x)} \le \dfrac{a}{b}$.

Proof. Let $x_1 < x_2$. Since $b-a > 0$, we have $x_1(b-a) < x_2(b-a)$. Adding $x_1x_2 + ab$ to both sides of the inequality, we obtain $(a-x_2)(b-x_1) < (a-x_1)(b-x_2)$. Consequently, $x_1 < x_2$ implies

$$\frac{a-x_2}{b-x_2} < \frac{a-x_1}{b-x_1}, \tag{10}$$

and therefore the maximum of $\dfrac{(a-x)}{(b-x)}$ for the interval $0 \le x < a$ occurs at $x=0$, which is equal to $\dfrac{a}{b}$. ■

Lemma 2. *For* $0 \le b < a < x \le 1$, $\dfrac{(x-a)}{(x-b)} \le \dfrac{1-a}{1-b}$.

Proof. Let $x_1 > x_2$. Since $a - b > 0$, we have $x_1(a-b) > x_2(a-b)$. Adding $x_1x_2 + ab$ to both sides of the inequality, we obtain $(x_1 - a)(x_2 - b) > (x_1 - b)(x_2 - a)$. Consequently, $x_1 > x_2$ implies

$$\frac{x_1 - a}{x_2 - a} > \frac{x_1 - b}{x_2 - b}, \tag{11}$$

and therefore the maximum of $\frac{(x-a)}{(x-b)}$ for the interval $a < x \leq 1$ occurs at $x = 1$, which is equal to $\frac{1-a}{1-b}$. ■

To prove Theorem 2, we use the following conventions for notational convenience:

- r denotes $p(1|f_{\mathcal{R}}) = P(C = 1|F_{\mathcal{S}} = f_{\mathcal{R}})$, that is the probability that a sample with state $f_{\mathcal{R}}$ for the genes in $\mathcal{R}$ is associated with the phenotype of interest.
- $\mathcal{T}$ denotes $\mathcal{R} \setminus \mathcal{S}$ and $f_{\mathcal{T}}$ denotes the state of $\mathcal{T}$ that is consistent with $f_{\mathcal{R}}$.
- γ denotes $P(F_{\mathcal{T}} = f_{\mathcal{T}}|F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability of observing state $f_{\mathcal{T}}$ for subnetwork $\mathcal{T}$, given that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$.
  θ denotes $P(C = 1|F_{\mathcal{S}} = f_{\mathcal{S}}, F_{\mathcal{T}} \neq f_{\mathcal{T}})$, that is the probability that a sample is associated with the phenotype of interest, given that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$, but subnetwork $\mathcal{T}$ is not in state $f_{\mathcal{T}}$ in that sample.

Proof of Theorem 2. We can write $J(f_{\mathcal{R}}; C)$ and $J_{\text{bound}}(f_{\mathcal{S}}, C)$ as follows:

$$J(f_{\mathcal{R}}; C) = z\gamma(r \log(r/q) + (1-r) \log((1-r)/(1-q))). \tag{12}$$

$$J_{\text{bound}}(f_{\mathcal{S}}; C) = z \max\{s \log(1/q), (1-s) \log(1/(1-q))\}. \tag{13}$$

We will show that, for fixed $\mathcal{S}$ and $f_{\mathcal{S}}$, the maximum value that $J(f_{\mathcal{R}}; C)$ attains cannot exceed $J_{\text{bound}}(f_{\mathcal{S}}; C)$ (for any choice of $\mathcal{T}$ and $f_{\mathcal{T}}$). First, by definition of conditional probability, we note the following equality:

$$s = \gamma r + (1-\gamma)\theta. \tag{14}$$

Since s is fixed, this equation represents a constraint that must be satisfied by r, γ, and θ. Thus, we will bound $J(f_{\mathcal{R}}; C)$ subject to this constraint. Note also that we can write this constraint as

$$\gamma = \frac{s-\theta}{r-\theta} = \frac{\theta - s}{\theta - r}. \tag{15}$$

Without loss of generality, we assume $s > q$, that is the observation of state function $f_{\mathcal{S}}$ increases the probability of a sample being associated with the phenotype ($f_{\mathcal{S}}$ "indicates" phenotype). Since we consider only two classes for the samples (phenotype or control), if the assumption does not hold (i.e., if $f_{\mathcal{S}}$ "indicates" control), then the following arguments still hold if we simply interchange the labels of sample classes.

Given that $s > q$, five different cases are possible: (i) $q < s < r$, (ii) $q < s = r$, (iii) $q < r < s$, (iv) $q = r < s$, and (v) $r < q < s$. We consider each case separately.

Case (i): $q < s < r$. In this case, the probability of phenotype given the state of the larger subnetwork is larger than the probability of phenotype given the state of the smaller subnetwork (and thus the additional part of the larger subnetwork provides additional evidence indicating that the sample might be associated with phenotype).

Since $s < r$, we have $r > \gamma r + (1-\gamma)\theta$ from (14) and thus $r > \theta$. Therefore, since $0 \leq \theta < s < r \leq 1$, we can write by Lemma 1 that $\gamma \leq s/r$, without putting any additional constraint on r. Consequently, from (12), we obtain

$$J(f_{\mathcal{R}}; C) \leq z\frac{s}{r}(r \log(r/q) + (1-r) \log((1-r)/(1-q))) \tag{16}$$

and thus

$$J(f_{\mathcal{R}}; C) \leq zs(\log(r/q) + (1/r - 1) \log((1-r)/(1-q))). \tag{17}$$

Since $q < r \leq 1$, the second term in parenthesis is negative. Consequently, noting $r \leq 1$, we obtain

$$J(f_{\mathcal{R}}; C) \leq zs \log(1/q) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{18}$$

This proves the theorem for case (i).

Case (ii): $q < s = r$. In this case, the probability of phenotype given the state of the larger subnetwork is equal to the probability of phenotype given the state of the smaller subnetwork (and thus the additional part of the larger subnetwork does not provide additional information).

Noting $\lambda \leq 1$ and replacing r with s, we can write

$$J(f_{\mathcal{R}}; C) \leq z(s \log(s/q) + (1-s) \log((1-s)/(1-q))). \tag{19}$$

Since $1 - s < 1 - q$, the second term in parentheses is negative, so we have

$$J(f_{\mathcal{R}}; C) \leq z(s \log(s/q)) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{20}$$

This proves the theorem for case (ii).

Case (iii): $q < r < s$. In this case, the observation of the state of the larger subnetwork increases the probability of phenotype compared to background, but not to the extent that the smaller subnetwork does.

The proof here is very similar to that in case (ii). Let $y(x) = x \log(x/q) + (1-x) \log((1-x)/(1-q))$. Then we have $y'(x) = \log(x/q) - log((1-x)/(1-q))$. Therefore, for $x > q$, since $x/q > 1$ and $(1-x)/(1-q)$, $y'(x)$ is always positive and y is an increasing function of x. Consequently, for $q < r < s$, we have:

$$r \log(r/q) + (1-r) \log((1-r)/(1-q)) \leq s \log(s/q) + (1-s) \log((1-s)/(1-q)). \tag{21}$$

Once this inequality is etablished, the rest of the proof for case (iii) follows the proof for case (ii).

Case (iv): $q = r < s$. In this case, the probability of phenotype given the state of the larger subnetwork is equal to background, thus the additional part of the larger subnetwork takes away all the evidence provided by the smaller subnetwork in favor of phenotype.

By definition of $J(.)$, $J(f_{\mathcal{R}}; C) = 0$ (both r/q and $(1-r)/(1-q)$ are equal to 1 in (12)). Thus, $J(f_{\mathcal{R}}; C)$ trivially satisfies the bound, proving the theorem for this case.

Case (v): $r < q < s$. In this case, the additional part of the larger subnetwork reverses the direction of evidence provided by the smaller subnetwork, that is the state function of the larger subnetwork increases the probability of the sample being associated with control.

The proof in this case is very similar to that for case (i). Since $r < s$, using Equation 14 we have $r < \gamma r + (1-\gamma)\theta$ and thus $r < \theta$. Therefore, since $0 \leq r < s < \theta \leq 1$, we can write by Lemma 2 (see below) that $\gamma \leq (1-s)/(1-r)$, without putting any additional constraint on r. Consequently, from (12), we obtain

$$J(f_{\mathcal{R}}; C) \leq z \frac{1-s}{1-r} (r \log(r/q) + (1-r) \log((1-r)/(1-q))) \tag{22}$$

and thus

$$J(f_{\mathcal{R}}; C) \leq z(1-s)(\log((1-r)/(1-q)) + (r/(1-r)) \log(r/q)). \tag{23}$$

Since $r < q$, the second term in parenthesis is negative and also $1 - r \leq 1$; therefore

$$J(f_{\mathcal{R}}; C) \leq z(1-s) \log(1/(1-q)) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{24}$$

This proves the theorem for case (v). ■

Note that this theorem does not state that the J-value of a state function is bounded by the J-value of its smaller parts, it rather provides a bound on the J-value of the larger state function based on simpler statistics of its smaller parts. Using this bound, we develop an algorithm, CRANE, to efficiently search for informative state functions. CRANE enumerates state functions in a bottom-up fashion, by pruning out the search space effectively based on the following principles:

1. A state function $f_{\mathcal{S}}$ is said to be a *candidate* state function if $|\mathcal{S}| = 1$ or $J(f_{\mathcal{S}}; C) \geq J(f_{\mathcal{S} \setminus \{g_i\}}; C)$ for all $g_i \in \mathcal{S}$.
2. A candidate state function $f_{\mathcal{S}}$ is said to be *extensible* if $J_{\text{bound}}(f_{\mathcal{S}}; C) \geq j^*$. This restriction enables pruning of larger state functions using statistics of smaller state functions.

3. An extension of state function $f_{\mathcal{S}}$ is obtained by adding one of the H or L states of a gene $g_i \in \mathcal{V}\backslash\mathcal{S}$ such that $g_i g_j \in \mathcal{E}$, where g_j is the most recently added gene to $f_{\mathcal{S}}$. This ensures network connectivity of the subnetwork associated with the generated state functions.
4. For an extensible state function, all possible extensions are considered and among those that qualify as candidate state functions, the top b state functions with maximum $J(.)$ are selected as candidate state functions. Here, b is an adjustable parameter that determines the breadth of the search and the case $b = 1$ corresponds to a greedy algorithm.
5. An extensible state function $f_{\mathcal{S}}$ is not extended if $|\mathcal{S}| = d$. Here, d is an adjustable parameter that determines the depth of the search.

CRANE enumerates all candidate state functions that qualify according to these principles, for given j^*, b, and d. At the end of the search process, the candidate state functions that are not superceded by another candidate state function (the leaves of the enumeration tree) are identified as informative state functions, if their J-value exceeds j^*. A detailed pseudo-code for this procedure is given as Algorithm 1.

Algorithm 1: CRANE-EXTENDSTATEFUNCTION $((\mathcal{S}, f_{\mathcal{S}}), \mathcal{T}, j^*, b, d)h$: Extends a subnetwork and associated state function. Invoked for each $g_i \in \mathcal{V}$ and $\hat{E}_i \in \{0, 1\}$ as CRANE-EXTENDSTATEFUNCTION, $((\{g_i\}, \{\hat{E}_i\}), \emptyset, j^*, b, d)$, where j^*, b, and d are user-defined.

Global: $\triangleright \mathcal{V}$: Set of genes, C: Phenotype vector
$\triangleright$ E: Gene expression associated with $\mathcal{V}$ and C, $\mathcal{E}$: PPI dataset associated with $\mathcal{V}$
Inputs: $\triangleright$ $(\mathcal{S}, f_{\mathcal{S}})$: Subnetwork/state-function pair to be extended
$\triangleright$ j^*: Threshold on information provided by a state function on phenotype
$\triangleright$ d: Maximum number of genes in a subnetwork
$\triangleright$ b: Maximum number of immediate extensions of a subnetwork/state-function pair
Input/Output: $\triangleright$ $\mathcal{T}$: Set of state functions informative of phenotype.

if $|\mathcal{S}| = = d$ **then**
  if $(J(f_{\mathcal{S}}; C) \geq j^*$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\mathcal{S}, f_{\mathcal{S}})\}$
  end if
  return
end if
$\mathcal{Q} \leftarrow \emptyset$; $g_i \leftarrow$ most recently added gene to $\mathcal{S}$
for each $g_k : (g_i, g_k) \in \mathcal{E}$ **and** $g_k \notin \mathcal{S}$ **do**
  $\mathcal{S}' \leftarrow \mathcal{S} \cup \{g_k\}$
  for each $\hat{E}_k \in \{0, 1\}$ **do**
    $f_{\mathcal{S}'} \leftarrow f_{\mathcal{S}'} \cup \hat{E}_k$; *redundant* $\leftarrow$ **false**
    for $g_j \in \mathcal{S}'$ **do**
      if $(J(f_{\mathcal{S}'\backslash\{g_j\}}; C) \geq J(f_{\mathcal{S}'}; C)$ **then**
        redundant $\leftarrow$ **true**; **break**
      end if
    end for
    if (**not** *redundant*) **and** $((J_{\max}(f_{\mathcal{S}'}; C) > j^*))$ **then**
      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(\mathcal{S}', f_{\mathcal{S}'})\}$
    end if
  end for
end for
if $\mathcal{Q} = \emptyset$ **then**
  if $J(f_{\mathcal{S}}; C) \geq j^*$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\mathcal{S}, f_{\mathcal{S}})\}$
  end if
  return
end if
$\mathcal{Q}_b \leftarrow$ set of top b subnetwork/state-function pairs in Q with respect to $J(f_{\mathcal{S}'}; C)$
for each $(\mathcal{S}', f_{\mathcal{S}'}) \in \mathcal{Q}_b$ **do**
  CRANE-EXTENDSTATEFUNCTION $((\mathcal{S}', f_{\mathcal{S}'}), \mathcal{T}, j^*, b, d)$
end for

2.6. Using state functions to predict metastasis in cancer

An important application of informative state functions is that they can serve as features for classification of phenotype. Since the genes that compose an informative state function are by definition highly discriminative of phenotype and control when considered *together*, they are expected to perform better than single-gene features (Chuang et al., 2007). Note here that CRANE discovers specific state functions that are informative of phenotype, as opposed to subnetworks that can discriminate phenotype or control. However, by Equation 2, we expect that a high $J(f_S, C)$ for a specific state function f_S is associated with a potentially high $I(F_S, C)$ for the corresponding subnetwork S. Therefore, for the application of CRANE in classification, we sort the subnetworks that are associated with discovered state functions based on their combinatorial coordinate dysregulation $I(F_S, C)$ and use the top K disjoint (non-overlapping in terms of their gene content) subnetworks with maximum $I(F_S, C)$ as features for classification. In the next section, we report results of classification experiments for different values of K.

Deriving representative features for subnetworks is a challenging task. Using simple aggregates of individual expression levels of genes along with traditional classifiers (e.g., regression or SVMs) might not be adequate, since such representations may not capture the combinatorial relationship between the genes in the subnetwork. For this reason, we use neural networks that incorporate subnetwork states (F_S) directly as features. The proposed neural network model is illustrated in Figure 2. In the example of this figure, two subnetworks are used to build the classifier. Each input is the expression level of a gene and the inputs that correspond to a particular subnetwork are connected together to an input layer neuron. All input layer neurons, each representing a subnetwork, are connected to a single output layer neuron, which produces the output. Each layer's weights and biases are initialized with the Nguyen-Widrow layer initialization method (provided by Matlab's initnw parameter). Then for a given gene expression dataset for a range of control and phenotype samples (which, in our experiments, is identical to that used for identification of informative state functions), the network is trained with Levenberg-Marquardt backpropagation (using Matlab's trainlm parameter), so that, given expression profiles in the training dataset, the output of the second layer matches the associated phenotype vector within minimal mean squared error. This learned model is then used to perform classification tests on a different gene expression dataset for the same phenotype.

Since Neural Networks show stochastic behavior, we train 30 independent NNs with the same training data and use the following voting scheme to consolidate these 30 different runs. For each run, we feed both the training and test samples as separate test data to the trained neural network and collect the real valued predictions for all training and test sample cases. For each neural network, we convert the quantitative NN outputs to binary predictions based on the NN outputs of training samples. We then predict the discrete class label of each test sample using output of NN as features. This procedure generates 30 separate class labels for each sample, each predicted by a particular NN. Then the final class label of each sample is determined based on the majority of the predicted class labels for that sample (i.e., for a particular sample, if more than 50% of the output labels represent phenotype, we declare it as a phenotype sample).

3. RESULTS AND DISCUSSION

In this section, we evaluate the performance of CRANE in identifying state functions associated with metastasis of CRC. We first compare the classification performance of the subnetworks associated with these state functions against single gene markers and subnetworks identified by an algorithm that aims to

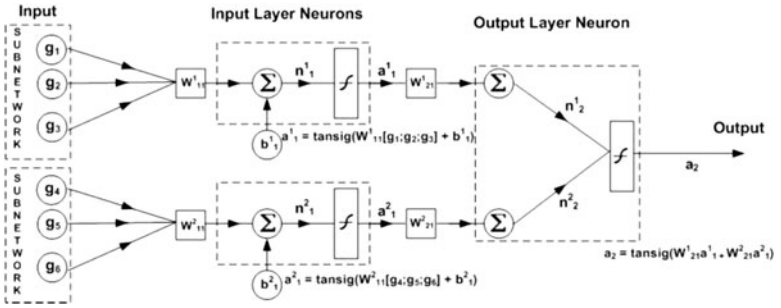

FIG. 2. Neural network model used to utilize subnetworks identified by CRANE for classification. Each subnetwork is represented by an input layer neuron, and these neurons are connected to a single output layer neuron.

maximize additive coordinate dysregulation. We then present comprehensive experimental results to evaluate the effect of parameters on the performance of CRANE. Subsequently, with a view to investigating the benefits of pruning the subnetwork search space, we compare CRANE's performance with a version that does not use the bound on $J(.)$ value to prune the search space. Finally, we inspect the subnetworks that are useful in classification, and discuss the insights these subnetworks can provide into the metastasis of CRC.

3.1. Datasets

In our experiments, we use two CRC-related microarray datasets obtained from GEO (Gene Expression Omnibus; *http://www.ncbi.nlm.nih.gov/geo/index.cgi*). These datasets, referenced by their accession number in the GEO database, include the following relevant data:

- *GSE6988* contains expression profiles of 17,104 genes across 29 *vs.* 51 colorectal tumor samples with and without liver metastasis, respectively (Ki et al., 2007).
- *GSE3964* contains expression profiles of 5,845 genes across 28 *vs.* 18 colorectal tumor samples with and without liver metastasis, respectively (Graudens et al., 2006).

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. We integrate the binary interactions and protein complexes using a matrix model (e.g., 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.

3.2. Experimental design

For each of the datasets mentioned above, we discover informative state functions (in terms of discriminating tumor samples with or without metastasis) using CRANE. While state functions that are indicative of either metastatic or non-metastatic phenotype can have high $J(.)$ values, we use only those that are indicative of (i.e., knowledge of which increases the likelihood of) metastatic phenotype for classification and further analyses, since such state functions are directly interpretable in terms of their association with metastasis.

In the experiments reported here, we set $b = 10$. d is set at 3 for *GSE3964* and at 6 for *GSE6988*. The value of j^{**} is set to and 0.15 and 0.40 for discovery of subnetworks on *GSE3964* and *GSE6988* respectively. The top five non-overlapping subnetworks discovered on *GSE6988* by CRANE using these parameter settings are shown in Table 2. Note that these parameters are used to balance the trade-off between computational cost of subnetwork identification and classification accuracy. The reported values are those that provide reasonable performance by spending a reasonable amount of time on subnetwork identification (a few hours in Matlab for each dataset). The effect of different values of these parameters on CRANE's performance are presented later in this section.

To binarize the gene expression datasets, we first normalize the gene expression profiles so that each gene has an average expression of 0 and standard deviation 1. Then we set the top α fraction of the entries in the normalized gene expression matrix to H (high expression) and the rest to L (low expression). In the reported experiments, we use $\alpha = 0.25$ (25% of the genes are expressed on an average) as this value is found to optimize the classification performance.

TABLE 2. FIVE NON-OVERLAPPING SUBNETWORKS THAT ARE ASSOCIATED WITH THE MOST INFORMATIVE STATE FUNCTIONS DISCOVERED ON GSE6988 WITH $D = 6$ AND THE FUNCTIONAL ENRICHMENT OF THESE SUBNETWORKS

Rank	*Proteins*	*Comb. coor. dysregulation*	*Most significantly enriched process*	*Enrichment* p-*value*
1	SERPINA3, KLK3, EPOR, GNB2L1, RASA1, RAF1	0.94	Inflammation	1×10^{-3}
2	E2F4, CCNE1, GSK3B, HNRPD, SF3B2, RPL13	0.85	Cell movement	1×10^{-3}
3	DMTF1, CCND2, AKAP8, DDX5, FN1, CRP	0.85	Cell migration	1×10^{-4}
4	ANXA11, PLSCR1, ESWR1, PTK2B, ITGB2, HP	0.81	Cell adhesion	1×10^{-4}
5	SKP1A, CCNA2, CDKN1A, GADD45G, EEF1G, RGL2	0.81	Inflammation	1×10^{-4}

3.3. Implementation of other algorithms

We identify *single gene markers* by running CRANE with $d = 1$ (i.e., by searching for subnetworks composed of one gene). We also identify coordinately dysregulated subnetworks using an *additive algorithm*, that is an algorithm that aims to maximize additive coordinate dysregulation (Chuang et al., 2007). The additive algorithm identifies a subnetwork associated with each gene in the network by seeding a greedy search process from that gene. It grows subnetworks by iteratively adding to the subnetwork a network neighbor of the genes that are already in the subnetwork. At each iteration, the neighbor that maximizes the coordinate dysregulation of the subnetwork is selected to be added. Once all subnetworks are identified, we sort these subnetworks according to their coordinate dysregulation ($I(E_S; C)$ or ($I(F_S; C)$ and use the top K disjoint subnetworks to train and test classifiers, for different values of K. While quantizing E_S to compute ($I(E_S; C)$, as suggested in Chuang et al. (2007), we use $\lfloor \log_2(|\mathcal{U}|) \rfloor + 1$ bins where $|\mathcal{U}|$ denotes the number of samples. Note that, in Chuang et al. (2007), the subnetworks identified by the greedy algorithm are filtered through three statistical tests. In our experiments, these statistical tests are not performed for the subnetworks discovered by the additive algorithm and CRANE. This is because, testing of statistical significance based on multiple runs on permuted instances is computationally expensive, since CRANE performs an almost exhaustive search of the subnetwork space. It should be noted that this is currently an important limitation of CRANE. In this respect, development of efficient algorithms for testing statistical significance of subnetworks identified by such exhaustive algorithms remains an important problem.

For the subnetworks with additive coordinate dysregulation, we compute the subnetwork activity E_S for each subnetwork, and use these as features to train and test two different classifiers: (i) a SVM using Matlab's svmtrain and svmclassify functions (this method is not applicable to combinatorial coordinate dysregulation), (ii) feed-forward neural networks, in which each input represents the subnetwork activity for a subnetwork and these inputs are connected to hidden layer neurons. For the single-gene markers, we rank all genes according to the mutual information of their expression profile with phenotype ($I(E_i; C)$) and use the expression level of K genes with maximum $I(E_i; C)$ as features for classification.

3.4. Classification performance

We evaluate the cross-classification performance of the subnetworks in the context of predicting metastasis of CRC. Namely, we use subnetworks discovered on the *GSE6988* dataset to train classifiers and we test the resulting classifiers on all the samples of *GSE3964*. Similarly, we use subnetworks discovered on *GSE3964* to train classifiers using the same dataset and perform testing of these classifiers on 28 metastatic and 20 randomly selected non-metastatic samples of *GSE6988*. The cross-classification performance of subnetworks discovered by an algorithm is not only indicative of the power of the algorithm in discovering subnetworks that are descriptive of phenotype, but also the reproducibility of these subnetworks across different datasets.

The classification performances of the subnetworks identified by CRANE, the additive algorithm, and single gene markers are compared in Figure 3. In the figure, for each $1 \leq K \leq 10$, the precision and recall achieved by each classifier are reported. These performance criteria are defined as follows:

$$\text{precision} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}} \tag{25}$$

$$\text{recall} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}} \tag{26}$$

Here, a true positive is defined as a metastatic sample that is correctly predicted as a metastatic sample, while a false positive is a non-metastatic sample that is incorrectly predicted as metastatic. A false negative is a metastatic sample that is incorrectly predicted as non-metastatic. Therefore, precision quantifies the fraction of true positives among all samples predicted as metastatic by the classifier, while recall quantifies the fraction of true positives among all metastatic samples.

As seen in Figure 3, subnetworks identified by CRANE outperform the subnetworks identified by other algorithms in predicting metastasis of colorectal cancer. In fact, in both cases, CRANE has the potential to deliver very high accuracy using very few subnetworks (maximum precision of 100% on both *GSE6988* and *GSE3964*, maximum recall of and 95% and 86% for classification of samples in *GSE6988* and

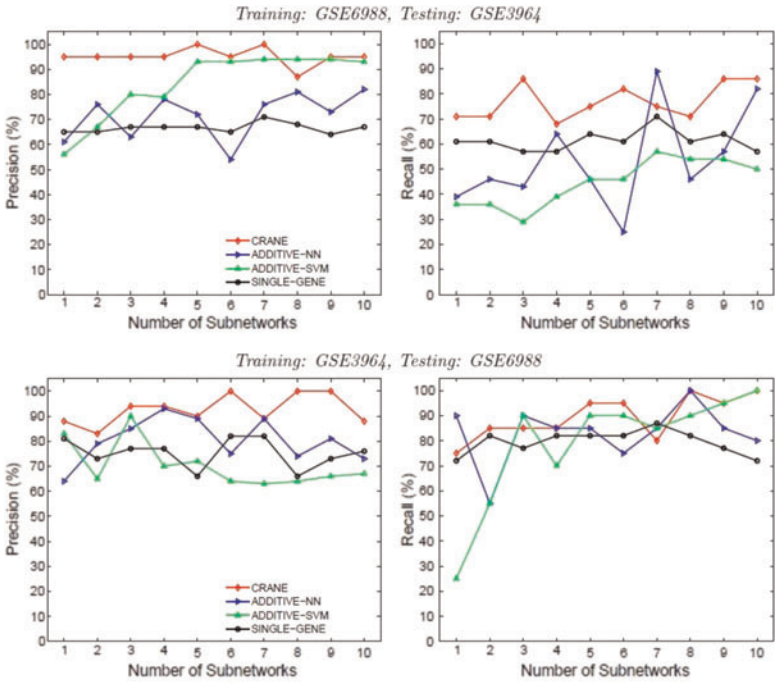

FIG. 3. Classification performance of subnetworks identified by CRANE in predicting colon cancer metastasis, as compared to single gene markers and subnetworks identified by algorithms that aim to maximize additive coordinate dysregulation. Subnetworks identified by CRANE are used to train neural networks (NNs), while those identified by the additive algorithm are used to train NNs, as well as support vector machines (SVMs). In the graphs, horizontal axes show the number of disjoint subnetwork features (with maximum combinatorial or additive coordinate dysregulation) used in classification, and vertical axes show the precision and recall achieved by the classifier.

GSE3964, respectively). While we use a simple feature selection method here for purposes of illustration, the performance of CRANE subnetworks are quite consistent, suggesting that these performance figures can indeed be achieved by developing elegant methods for selection of subnetwork features. These results are rather impressive, given that the best performance that can be achieved by the additive algorithm is 82%/93% precision and 89%/100% recall for the classification of *GSE3964* and *GSE6988*, respectively. Note that, while the performance of other algorithms is improved by increasing number of subnetwork features, the performance of CRANE appears to decline. This is likely because CRANE represents subnetwork features as multi-dimensional state functions. Therefore, while a few subnetworks each containing a few genes provide sufficient information for accurate classification, the accuracy declines as more subnetworks are incorporated because of the growth in dimensionality.

3.5. *Effect of pruning*

An important feature of CRANE is the use of a theoretical bound on $J(.)$ to prune out the search space. In order to verify the effectiveness of this feature in improving the efficiency of CRANE, as well as its ability to discover informative subnetworks, we compare CRANE with a version that does not apply pruning using the bound on $J(.)$. The results of this comparison are shown in Figure 4. These experiments are performed on *GSE*6988, by fixing $b = 10$, $j^{**} = 0.45$, $\alpha = 0.25$, and running CRANE and its version without pruning for d ranging from 1 to 8.

The runtimes of the CRANE and the algorithm without pruning are compared in Figure 4. As seen in the figure, the algorithm without pruning does not scale well with increasing d. This is expected, since the algorithm performs exhaustive search with a breadth of $b = 10$, making the runtime exponential in d. However, by pruning this search space using the bound on $J(.)$, CRANE reduces this runtime drastically, providing orders of magnitude improvement for larger values of d. Note that, if $b = \infty$, both CRANE and its version without pruning are guaranteed to discover all subnetworks with $J(.) \geq j^*$. However, since the breadth of search is limited by parameter b, both algorithms may miss some subnetworks. In the experiments reported here, CRANE is able to identify all subnetworks that are identified by the version without pruning; i.e., CRANE achieves the drastic improvement in runtime without compromising sensitivity. These results clearly demonstrate the value of using the theoretical bound on $J(.)$ value while searching for informative subnetworks.

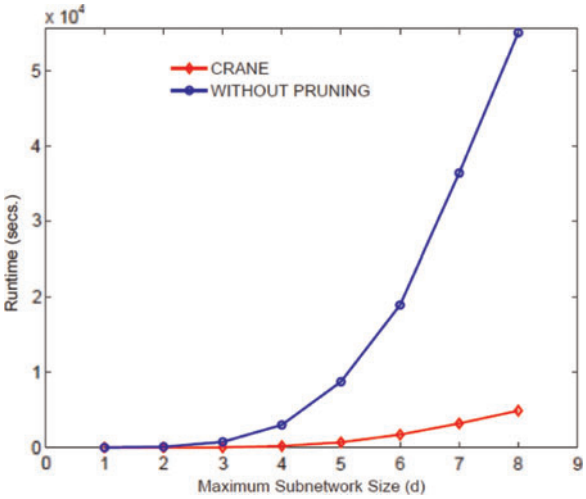

FIG. 4. Comparison of the runtimes of CRANE and its version that does not prune the subnetwork search space using the theoretical bound on $J(.)$. Note that CRANE identifies all subnetworks that are identified by the algorithm without pruning.

3.6. *Effect of parameters*

We also investigate the effect of parameters used to configure CRANE on classification performance of identified subnetworks, by fixing all but one of the parameters to the above-mentioned values and varying the remaining parameter. The tuneable parameters of CRANE are the following:

- d: d is the maximum size of a subnetwork. CRANE stops extending a subnetwork when the number of genes in the subnetwork reaches d. In other words, d determines the depth of the search.
- b: b is the number of state functions selected by CRANE at each iteration with maximum $J(.)$ value. Thus, b determines the breadth of the search.
- j^{**}: j^{**} is the minimum $J(.)$ value of a subnetwork state function to be considered informative.
- α: α is the fraction of the entries in the normalized gene expressin matrix that is set to H (high expression). The rest of the $(1-\alpha)$ entries of the gene expression matrix is set to L (low expression).

The results of our experiments on the effect of these parameters on the performance of CRANE are shown in Figure 5. In this figure, for each configuration of the parameters, we report the average F-measure across different values of the number of subnetworks used in classification, ranging from 1 to 10. Here, F-measure is defined as the harmonic mean of precision and recall, i.e.,

$$\text{F - measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{27}$$

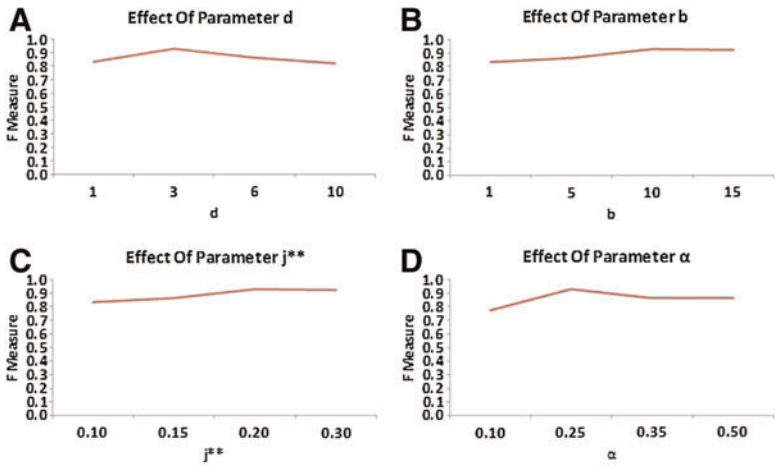

FIG. 5. The effect of parameters on the classification performance of subnetworks discovered by CRANE. For all experiments, subnetworks are discovered on *GSE3964* and tested on samples of *GSE6988*. The F-measure of a classifier is computed by increasing the number of subnetworks gradually from 1 to 10 and average F-Measure in these experiments is reported. (A–D) Performance of CRANE with respect to parameters d, b, j^{**}, and α, respectively. (**A**) $b=10$, and $j^{**}=0.15$. (**B**) $d=3$, and $j^{**}=0.15$. (**C**) $b=10$, and $d=3$. (**D**) $b=10$, $j^{**}=0.15$, and $d=3$.

We observe that classification performance is quite robust against variation in α ranging from 10% to 50%, while best performance is observed at $\alpha = 25\%$. As expected, classification performance improves by increasing j^{**}. Increasing the breadth of search (b) improves classification performance in general, which is also expected since larger values of b enable exploration of the search space further. Note that the special case with $b = 1$ is algorithmically equivalent to the additive algorithm with a different objective function (combinatorial coordinate dysregulation as opposed to additive coordinate dysregulation). We observe that CRANE outperforms the additive algorithm with $b = 1$ as well, indicating that the combinatorial formulation of coordinate dysregulation is potentially more useful than the additive formulation for classification.

As seen in Figure 5, increasing d improves performance as would be expected; however this improvement saturates for $d > 3$ and performance declines for larger subnetworks. This observation can be attributed to curse of dimensionality, since the number of possible values of random variable F (expression state of a subnetwork) grows exponentially with increasing subnetwork size. We also investigate the effect of parameter d on CRANE's ability to discover larger subnetworks. For this purpose, we compare the subnetworks identified by CRANE on *GSE*6988 using $d = 7$ and $d = 8$ with those identified using $d = 6$. The top five non-overlapping subnetworks identified using $d = 7$ and $d = 8$ are shown in Table 3. Comparison of the subnetworks in Tables 2 and 3 shows that, while there is some overlap in subnetworks discovered using different values of d, some subnetworks that can be discovered for larger values of d can be missed if a smaller value of d is used. Note, however, that this does not mean that smaller subnetworks of these subnetworks are not discovered by CRANE. Rather, such subnetworks are often eliminated because of their overlap with subnetworks that have higher combinatorial coordinate dysregulation. Indeed, comprehensive comparison of subnetworks shows that many of the subnetworks composed of seven genes, which are discovered using $d = 7$, are identified as different six-gene combinations when d is set 6. In other words, if d is set to a smaller value, then a larger "naturally occurring" subnetwork can be "truncated" into smaller subnetworks. For this reason, the parameter d needs to be set carefully, possibly by using different values of d and inspecting the size and gene content of subnetworks discovered for each d.

3.7. Subnetworks and state functions indicative of metastasis in CRC

Cancer metastasis involves the rapid proliferation and invasion of malignant cells into the bloodstream or lymphatic system. The process is driven, in part, by the dysregulation of proteins involved in cell adhesion and motility (Paschos et al., 2009), the degradation of the extracellular matrix (ECM) at the invasive front of the primary tumor (Zucker and Vacirca, 2004), and is associated with chronic inflammation (McConnell and Yang, 2009). An enrichment analysis of the top five subnetworks identified on *GSE6988* reveals that all of these subnetworks are highly significant for the network processes underlying these phenotypes (Table 2).

Further, as CRC metastasis is our classification endpoint, we wanted to evaluate our subnetworks in terms of their potential to propose testable hypotheses. In particular, to highlight the power of our model approach, we choose a subnetwork for which at least one gene was expressed in the state function

Table 3. Five Non-Overlapping Subnetworks that Are Associated with the Most Informative State Functions Discovered on GSE6988 for d = 7 and d = 8

Rank	*Proteins*	*Combinatorial coordinate dysregulation*
$d = 7$		
1	MYC, GJB5, RPL35A, TRPM1, HYOU1, ARF3, CDK10	0.81
2	FLJ20487, ATP5F1, GNG3, RBM17, RCN1, ENO3	0.79
3	TCF7, NDUFV2, ARMCX1, LRP16, PPT1, KIF13B, MSL3L1	0.79
4	SMARCA4, ATP5D, BTBD2, FGFBP1, PPP2R5D, PYCARD, PPP2R3A	0.76
5	MPG, NT5C2, RCL1, SH3BP5, MYT1, EEF1G	0.76
$d = 8$		
1	RARS, ARMCX1, SDF2, ARF3, HYOU1, F11R, LOC284361, ENO3	0.92
2	C18orf10, ATP5F1, VAV1, FBXW11, ASRGL1, EFHD1, KIAA0182, NDUFA1	0.89
3	SURB7, BRCA1, EXOSC4, PCBP2, RAD51C, TCF7, BCL2A1, CREB3	0.78
4	BMX, RAD51L3, PIN1, SLAMF1, CKLF, ELA3A, MMP2, N-PAC	0.77
5	SMARCA4, ATP5D, BTBD2, FGFBP1, PPP2R5D, PYCARD, PPP2R3A	0.76

indicative of CRC metastasis. This subnetwork contains TNFSF11, MMP1, BCAN, MMP2, TBSH1, and SPP1 and the state function LLLLLH (in respective order) indicates metastatic phenotype with *J*-value 0.33. The combinatorial dysregulation of this subnetwork is 0.72, while its additive coordinate dysregulation is 0.37, i.e., this is a subnetwork which would likely have escaped detection by the additive algorithm (this subnetwork is not listed in Table 2 since it is not among the top five scoring subnetworks). Using the genes in this subnetwork as a seed, we construct a small subnetwork diagram for the purpose of more closely analyzing the post-translational interactions involving these proteins. This is done using Metacore, a commercial platform that provides curated, highly reliable interactions. From this subnetwork, we remove all genes indicated to be not expressed in human colon by the database, and then selectively prune it in order to clearly focus on a particular set of interactions (Fig. 6). It merits noting that, although Brevican (BCAN) is in subnetwork, it is removed for being non-expressed in the human colon, although evidence from the Gene Expression Omnibus (see accession *GDS2609*) (Hong et al., 2007) casts doubt on this, as does the microarray we use for scoring (*GSE6988*).

As seen on the interaction diagram, SPP1 (osteopontin) and TBSH1 (thrombosponidin 1) interact with a number of the integrin heterodimers to increase their activity (green line). Integrin heterodimers play a major role in mediating cell adhesion and cell motility. SPP1, up-regulated in metastasis (Fig. 6), is a well-studied protein that triggers intracellular signaling cascades upon binding with various integrin heterodimers, promotes cell migration when it binds CD44, and when binding the alpha-5/beta-3 dimer in particular, promotes angiogenesis, which is associated with the metastatic phenotype of many cancers (Markowitz and Bertagnolli, 2009). MMP proteins are involved in the breakdown of ECM, particularly collagen which is the primary substrate at the invasive edge of colorectal tumors (Vishnubhotla et al., 2007). MMP-1 has an inhibitory effect on vitronectin (red line), hence the loss of expression of MMP-1 may ''release the brake'' on vitronectin, which in turn may increase the activity of the alpha-v/beta-5 integrin heterodimer. Likewise, MMP-2 shows an inhibitory interaction with the alpha-5/beta-3 dimer,

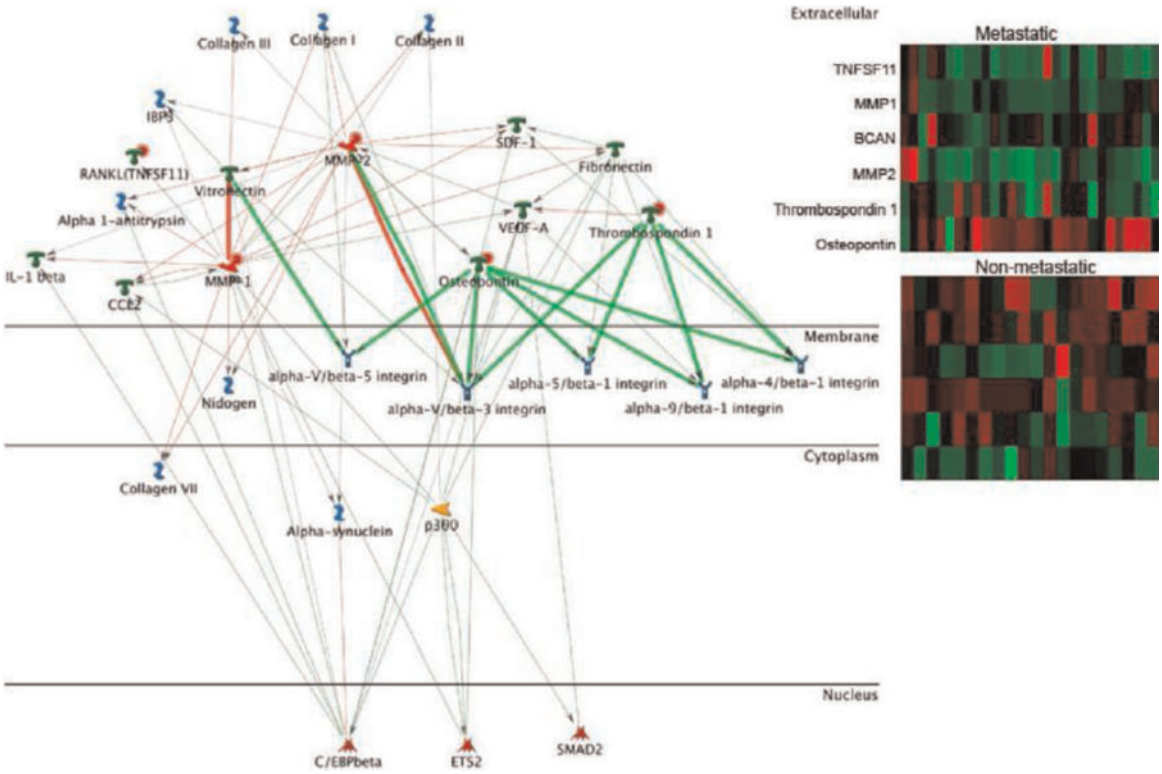

FIG. 6. Hypothesis-driver subnetwork: interaction diagram illustrating key interactions with gene products from a subnetwork identified by CRANE as indicative of CRC metastasis. Shown are the gene products in discovered subnetwork (red circles) and their direct interactions with other proteins. Green lines represent an activating interaction; red lines indicate an inhibitory interaction. Arrows indicate direction of interaction. (**Inset**) Expression pattern of subnetwork proteins at the level of mRNA.

which may counteract to some extent the activating potential of SPP1, suggesting that a loss of MMP-2 may exacerbate the metastatic phenotype. Taken together, these interactions suggest a number of perturbation experiments, perhaps by pharmacological inhibition or siRNA interference of the integrin dimmers or MMP proteins, to evaluate the role of these interactions, individually or synergistically, in maintaining the metastatic phenotype. Note also that, alpha-v/beta-5 integrin does not exhibit significant differential expression at the mRNA-level, suggesting that the state function identified by CRANE may be a signature of its post-translational dysregulation in metastatic cells.

4. CONCLUSION

We present a novel framework for network-based analysis of coordinate dysregulation in complex phenotypes. Experimental results on metastasis of colorectal cancer show that the proposed framework can achieve almost perfect performance when discovered subnetworks are used as features for classification. These results are highly promising in that the state functions that are found to be informative of metastasis can also be useful in modeling the mechanisms of metastasis in cancer. Detailed investigation of the state functions and the interactions between proteins that together compose state functions might therefore lead to development of novel hypotheses, which in turn may be useful for development of theurapetic intervention strategies for late stages of cancer.

ACKNOWLEDGMENTS

We would like to thank Vishal Patel, Jill Barnholtz-Sloan, Xiaowei Guan, and Gurkan Bebek, of Case Western Reserve University for many useful discussions. This work was supported, in part, by the NSF National Science Foundation (CAREER Award CCF-0953195) and the National Institutes of Health (Grants UL1-RR024989 from the National Center for Research Resources, Clinical and Translational Science Awards; P30-CA043703 from the Case Western Reserve University Cancer Center Proteomics Core; and T32-GM008803 from the NIGMS, Institutional National Research Service Award).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Akutsu, T., and Miyano, S. 2001. Selecting informative genes for cancer classification using gene expression data. *Proc. IEEE-EURASIP Workshop Nonlinear Signal Image Processing* 3–6.

Akutsu, T., Miyano, S., and Kuhara, S. 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 17–28.

Anastassiou, D. 2007. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* 3, 83.

Cabusora, L., Sutton, E., Fulmer, A., et al. 2005. Differential network expression during drug and stress response. *Bioinformatics* 21, 2898–2905.

Chowdhury, S.A., and Koyutürk, M. 2010. Identification of coordinately dysregulated subnetworks in complex phenotypes. *Pac. Symp. Biocomput.* 133–144.

Chuang, H.-Y., Lee, E., Liu, Y.-T., et al. 2007. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140.

Nacu, Ş., Critchley-Thorne, R., Lee, P., et al. 2007. Gene expression network analysis and applications to immunology. *Bioinformatics* 23, 850–858.

Ewing, R.M., Chu, P., Elisma, F., et al. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3, 89.

Franke, L., Bakel, H., Fokkens, L., et al. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* 78, 1011–1025.

Goh, K.-I., Cusick, M.E., Valle, D., et al. 2007. The human disease network. *Proc. Natl. Acad. Sci. USA*, 104, 8685–8690.

Graudens, E., Boulanger, V., Mollard, C., et al. 2006. Deciphering cellular states of innate tumor drug responses. *Genome Biol.* 3, R19.

Guo, Z., Li, Y., Gong, X., et al. 2007. Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. *Bioinformatics* 23, 2121–2128.
Hong, Y., Ho, K.S., Eu, K.W., et al. 2007. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin. Cancer Res.* 13, 1107–1114.
Ideker, T., and Sharan, R. 2008. Protein networks in disease. *Genome Res.* 18, 644–652.
Ideker, T., Ozier, O., Schwikowski, B., et al. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Proc. ISMB* 233–240.
Karni, S., Soreq, H., and Sharan, R. 2009. A network-based method for predicting disease-causing genes. *J. Comput. Biol.* 16, 181–189.
Ki, D.H., Jeung, H.-C., Park, C.H., et al. 2007. Whole genome analysis for liver metastasis gene signatures in colorectal cancer. *Int. J. Cancer* 121, 2005–2012.
Koyutürk, M., Szpankowski, W., and Grama, A. 2004. Biclustering gene-feature matrices for statistically significant dense patterns. *Proc. IEEE Comput. Syst. Bioinformatics Conf. (CSB'04)* 480–484.
Lage, K., Karlberg, O.E., Størling, Z.M., et al. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316.
Liu, M., Liberzon, A., Kong, S.W., et al. 2007. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* 3, e96+.
Markowitz, S., and Bertagnolli, M. 2009. Molecular origins of cancer: molecular basis of colorectal cancer. *N. Engl. J. Med.* 361, 2449–2460.
McConnell, B., and Yang, V. 2009. The role of inflammation in the pathogenesis of colorectal cancer. *Curr. Colorectal Cancer Rep.* 5, 69–74.
Nibbe, R.K., Ewing, R., Myeroff, L., et al. 2009. Discovery and scoring of protein interaction sub-networks discriminative of late stage human colon cancer. *Mol. Cell Prot*. 9, 827–845.
Nibbe, R.K., Koyutürk, M., and Chance, M.R. 2010. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput. Biol.* 6, e1000639+.
Papin, J.A., Hunter, T., Palsson, B.O., et al. 2005. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6, 99–111.
Paschos, K., Canovas, D., and Bird, N. 2009. The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cell Signal* 21, 665–674.
Patil, K.R., and Nielsen, J. 2005. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. USA* 102, 2685–2689.
Quackenbush, J. 2002. Microarray data normalization and transformation. *Nat. Genet.* 32, Suppl, 496–501.
Rhodes, D.R., and Chinnaiyan, A.M. 2005. Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37 Suppl, S31–S37.
Rich, J., Jones, B., Hans, C., et al. 2005. Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res.* 65, 4051–4058.
Schadt, E.E. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717.
Scott, M.S., Perkins, T., Bunnell, S., et al. 2005. Identifying regulatory subnetworks for a set of genes. *Mol. Cell Prot.* 4, 683–692.
Shmulevich, I., and Zhang, W. 2002. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 18, 555–565.
Smyth, P., and Goodman, R.M. 1992. An information theoretic approach to rule induction from databases. *IEEE Trans. Knowl. Data Eng.* 4, 301–316.
Ulitsky, I., Karp, R.M., and Shamir, R. 2008. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. *Proc. RECOMB 2008* 347–359.
Vishnubhotla, R., Sun, S., Huq, J., et al. 2007. Rock-ii mediates colon cancer invasion via regulation of mmp-2 and mmp-13 at the site of invadopodia as revealed by multiphoton imaging. *Lab. Invest.* 87, 1149–1158.
Watkinson, J., Wang, X., Zheng, T., et al. 2008. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst. Biol.* 2, 10.
Zucker, S., and Vacirca, J. 2004. Role of matrix metalloproteinases (MMPS) in colorectal cancer. *Cancer Metastasis Rev.* 23, 101–117.

Address correspondence to:
Dr. Mehmet Koyutürk
Department of Electrical Engineering and Computer Science
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106

E-mail: mxk331@case.edu