Subnetwork State Functions Define Dysregulated Subnetworks in Cancer

Salim A. Chowdhury[1], Rod K. Nibbe[2,4]
Mark R. Chance[3,4], Mehmet Koyutürk[1,4]

(1) Dept. of Electrical Engineering & Computer Science, (2) Dept. of Pharmacology
(3) Dept. of Physiology & Biophysics, (4) Center for Proteomics & Bioinformatics
Case Western Reserve University, Cleveland, OH 44106, USA.
E-mail: {sxc426,rkn6,mrc16,mxk331}@case.edu

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 greedy 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 simple greedy algorithms. Comprehensive cross-classification experiments show that subnetworks identified by CRANE significantly outperform those identified by greedy algorithms in predicting metastasis of colorectal cancer (CRC).

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 [1]. However, cellular systems are orchestrated through combinatorial organization of thousands of biomolecules [2]. 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 on identification of multiple markers that together exhibit differential expression with respect to various phenotypes [3, 4].

Network-based approaches to identification of multiple markers. High-throughput protein-protein interaction (PPI) data [5] 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 [6, 7]. This observation is exploited to identify novel genetic markers based on network connectivity [8–10]. For the identification of differentially expressed subnetworks with respect to GAL80 deletion in yeast, Ideker *et al.* [11] propose a method that is based on searching for connected subgraphs with high aggregate significance of individual differential expression. Variations of this method are shown to be effective in identifying multiple genetic markers in prostate cancer [12], melanoma [13], diabetes [14], and others [15–17].

Coordinate/synergistic dysregulation. Network-based approaches are further elaborated to capture coordinate dysregulation of interacting proteins at a sample-specific resolution [18]. Ulitksy *et al.* [19] define dysregulated pathways as subnetworks composed of products of genes that are dysregulated in a large fraction of phenotype samples. Chuang *et al.* [20] 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 greedy 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 [20]. Nibbe *et al.* [21, 22] show that this notion of coordinate dysregulation is also effective in integrating protein and mRNA expression data to identify important subnetworks in colon cancer (CRC). Anastassiou [23] 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 [23], important insights can still be gained through pairwise assessment of synergy [24].

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) greedy algorithms may not be able to adequately capture the coordination between multiple genes that provide weak individual signals. In this paper, 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.*, SVMs).

We describe these algorithmic innovations in detail in Section 2.

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 significantly outperform subnetworks identified by greedy algorithms in terms of accuracy of classification. We also investigate the highly informative subnetworks in detail to assess their potential in highlighting the mechanisms of metastasis in 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 protein-protein interaction (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 paper, 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*.

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 [25] 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 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, $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$

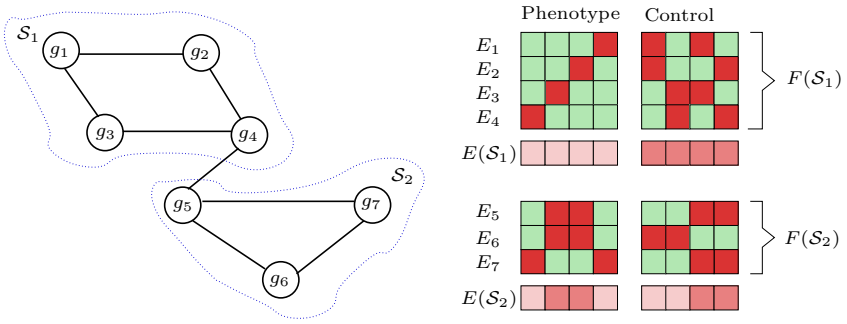


Fig. 1. Additive *vs.* combinatorial coordinate dysregulation. Genes (g) are shown as nodes, interactions between their products are shown as edges. Expression profiles (E) of genes are shown by colormaps. Dark red indicates high expression (`H`), light green indicates low expression (`L`). None of the 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).

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.

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.* [20] 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.

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 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, algorith-

mic 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 [26] to clustering [27] and classification [28]. Ulitsky *et al.* [19] 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 [29]. 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 $\hat{E}_i(j) = \mathtt{H}$ and *low expression* if $\hat{E}_i(j) = \mathtt{L}$. Then, the *combinatorial coordinate dysregulation* of subnetwork $\mathcal{S}$ is defined as

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

where $F_{\mathcal{S}} = \{\hat{E}_1, \hat{E}_2, ..., \hat{E}_m\} \in \{\mathtt{L}, \mathtt{H}\}^m$ 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 [23] 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 [23]. 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 ($I(F_{\mathcal{S}}; C)$), greedy algorithms 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.

Subnetwork state functions informative of phenotype. Let $f_{\mathcal{S}} \in \{\mathtt{H}, \mathtt{L}\}^m$ 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

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

where

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

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. It can be shown that for any $\mathcal{S} \subseteq \mathcal{V}$, $0 \leq J(f_{\mathcal{S}}; C) \leq \max\{-p(c)\log p(c), -(1 - p(c))\log(1 - p(c))\} = j_{\max}(p(c))$ [30], where $p(c)$ denotes the fraction of phenotype samples among all available samples. Therefore, in practice, we allow the user to specify a threshold j^{**} in the range $[0, 1]$ and adjust it as $j^* = j^{**} j_{\max}(p(c))$, to make the scoring criterion interpretable and uniform across all datasets. 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.

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 [31] 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 1. *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:*

$$J(f_{\mathcal{R}}; C) \leq p(f_{\mathcal{S}}) \max_{c \in \{0,1\}} \left\{ p(c|f_{\mathcal{S}}) \log \frac{1}{p(c)} \right\} = J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{4}$$

The proof of this theorem is provided in the supplementary materials [30]. 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 $g_i \in \mathcal{V} \setminus \mathcal{S}$ 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 in the supplementary materials [30].

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 [20]. 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 $J(f_{\mathcal{S}}, C)$ for a specific state function $f_{\mathcal{S}}$ is associated with a potentially high $I(F_{\mathcal{S}}, C)$ for the corresponding subnetwork $\mathcal{S}$. 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 $I(F_{\mathcal{S}}, C)$ and use the top K disjoint (non-overlapping in terms of their gene content) subnetworks with maximum $I(F_{\mathcal{S}}, C)$ 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 ($F_{\mathcal{S}}$) 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

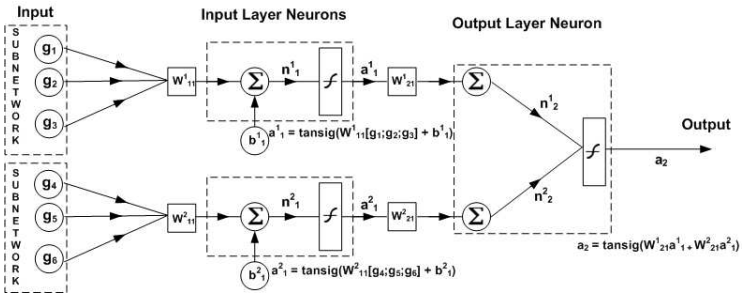


Fig. 2. Neural network model used to utilize subnetworks identified by CRANE for classification. Each subnetwork is represented by an input layer neuron and these neurons are connected to a single output layer neuron.

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 take a voting amongst them to determine the final class label of a particular sample.

3 Results and Discussion

In this section, we evaluate the performance of CRANE in identifying state functions associated with metastasis of colorectal cancer (CRC). We first compare the classification performance of the subnetworks associated with these state functions against single gene markers and subnetworks identified by two greedy algorithms that aim to maximize additive and combinatorial coordinate dysregulation. Then, we inspect the subnetworks that are useful in classification, and discuss the insights these subnetworks can provide into metastatis of CRC.

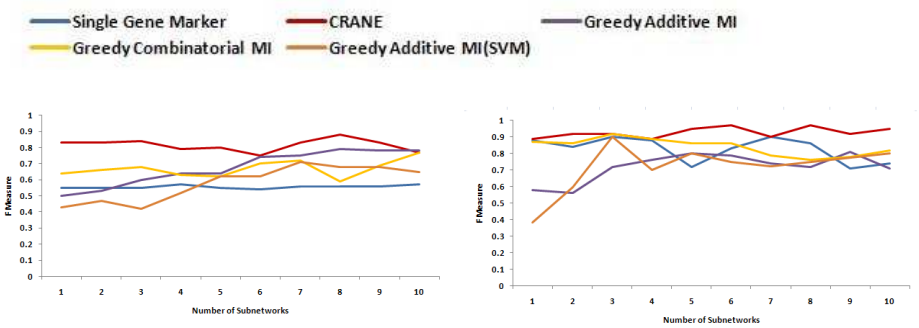
Datasets. In our experiments, we use two CRC related microarray datasets obtained from GEO (Gene Expression Omnibus, `http://www.ncbi.nlm.nih.gov/geo/index.cgi`). These datasets, referenced by their accession number in the GEO database, include the following relevant data:

- *GSE6988* contains expression profiles of 17,104 genes across 29 *vs.* 20 colorectal tumor samples with and without liver metastasis, respectively.
- *GSE3964* contains expression profiles of 5,845 genes across 28 *vs.* 18 colorectal tumor samples with and without liver metastasis, respectively.

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.

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 0.15 and 0.50 for discovery of subnetworks on *GSE3964* and *GSE6988* respectively. 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). 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.

Implementation of other algorithms. We also use two greedy algorithms to identify coordinately dysregulated subnetworks, one of which aims to maximize additive coordinate dysregulation [20], while the other aims to maximize combinatorial coordinate dysregulation. We implement the greedy algorithms to identify a subnetwork associated with each gene in the network by seeding the greedy search process from that gene. The greedy algorithms grow 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 ($I(E_{\mathcal{S}}; C)$ or $I(F_{\mathcal{S}}; C)$) and use the top K disjoint subnetworks to train and test classifiers, for different values of K. The binarization scheme for greedy identification of combinatorially dsregulated subnetworks is identical to that for CRANE. While quantizing $E_{\mathcal{S}}$ to compute $I(E_{\mathcal{S}}; C)$, as suggested in [20], we use $\lfloor \log_2(|\mathcal{U}|) \rfloor + 1$ bins where $|\mathcal{U}|$ denotes the number of samples. Note that, in [20], 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 any of the three algorithms.



Training: GSE6988, Testing: GSE3964 Training: GSE3964, Testing: GSE6988

Fig. 3. Classification performance of subnetworks identified by CRANE in predicting colon cancer metastasis, as compared to those identified by greedy algorithms that aim to maximize combinatorial or additive coordinate dysregulation, as well as single-gene markers. Subnetworks identified by CRANE and greedy combinatorial algorithm are used to train neural networks (NNs), while those identified by the greedy 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 greedy coordinate dysregulation) used in classification, vertical axes show F Measure achieved by the corresponding classifier.

The design of classifiers for combinatorially dsregulated subnetworks identified by the greedy algorithm is also identical to that for subnetworks identified by CRANE. For the subnetworks with additive coordinate dysregulation, we compute the subnetwork activity $E_{\mathcal{S}}$ for each subnetwork, and use these as features to train and test two different classifiers: (i) a support vector machine (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.

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 *GSE3964*. Similarly, we use subnetworks discovered on *GSE3964* to train classifiers using the same dataset and perform testing of these classifiers on *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 performance of the subnetworks identified by CRANE and greedy algorithms is compared in Figure 3. In the figure, for each $1 \leq K \leq 10$, the 'F Measure' is reported for each classifier. F measure is representative of the

overall performance of a classifier, which is calculated as the geometric mean of precision (selectivity) and recall (sensitivity) of predictions. 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. An F measure of 1.0 indicates that the classifier provides perfect precision without sacrificing recall (or vice versa).

As seen in Figure 3, subnetworks identified by CRANE significantly 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 accuracy of 97% and 88% for classification of samples of *GSE6988* and *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 achived by the greedy additive algorithm is 79% and 81% for the classification of *GSE3964* and *GSE6988*, respectively. On the other hand, the greedy algorithm for combinatorial coordinate dysregulation outperforms the greedy additive algorithm on the classification of *GSE6988* and performs poorly compared to CRANE. These results show that, besides the combinatorial formulation of coordinate dysregulation, the search algorithm implemented by CRANE also adds to the power of identified subnetworks in discriminating metastatic and non-metastatic samples.

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 results of these experiments are given in detail in the supplementary document[30].To summarize, we observe that classification performance is quite robust against variation in α ranging from 10% to 50%, while best performance is observed when $\alpha = 25\%$. As expected, classification performance improves by increasing j^{**}.While increasing d improves performances as would be expected, this improvement satirizes 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 grows exponentially with increasing subnetwork size. Finally, as larger b improves classification performance in general by increasing the breadth of the search, we observe no exception to this behavior.

Table 1. Five subnetworks that are associated with the most informative state functions discovered on GSE6988.

Rank	Proteins	Comb. Coor. Dysregulation	Most Significantly Enriched Process	Enrichment p-value
1	JAK2, STAT5A, IL7R, STAT3, IL2RA	0.80	Lymphocyte Proliferation	1×10^{-9}
2	CASP1, LMNA, CTCF, APP, APBA1	0.77	Cell Adhesion	1×10^{-6}
3	TRAF1, CFLAR, NFKB1, FBXW11, NFKBIB	0.56	Inflammation	1×10^{-8}
4	XRCC5, VAV1, ARGHDIA, RAC2, NOS2A	0.55	Inflammation	1×10^{-4}
5	CD9, KIT, BTK, WAS, NCK1	0.48	Cell Adhesion	1×10^{-4}

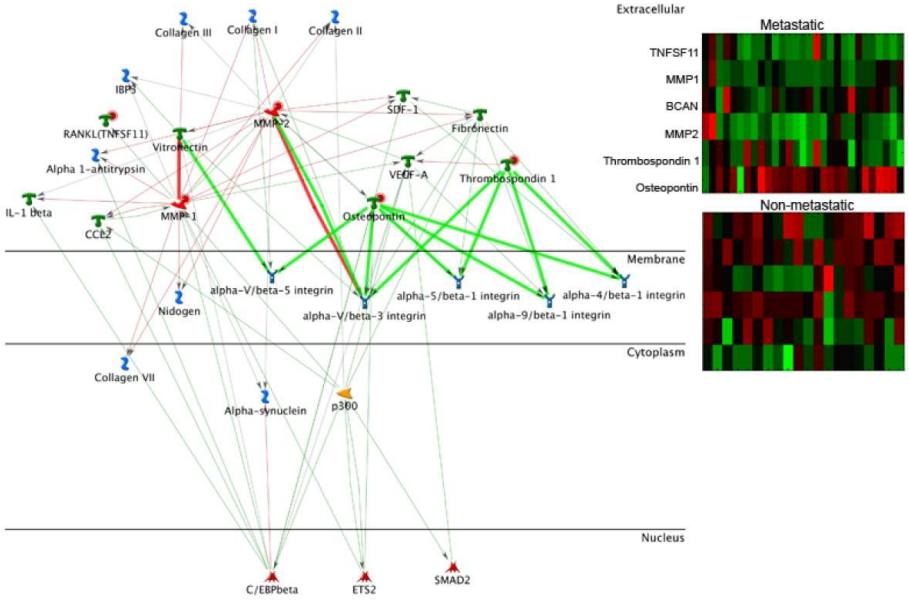


Fig. 4. Hypothesis-driver subnetwork - interaction diagram illustrating key interactions with gene products from a subnetwork identified by CRANE as indicative of CRC metastasis. Shown are the gene products in discovered subnetwork (red circles) and their direct interactions with other proteins. Green lines represent an activating interaction, red lines indicate an inhibitory interaction. Arrows indicate direction of interaction. Inset is the expression pattern of subnetwork proteins at the level of mRNA.

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 [32], the degradation of the extracellular matrix (ECM) at the invasive front of the primary tumor [33], and is associated with chronic inflammation [34]. 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 1).

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 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 greedy method based on additive dysregulation (this subnetwork is not listed in Table 1 since it is not among the top five scoring subnetworks). Using the genes in this sub-

network 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 (Figure 4). 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*) 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 (see inset in Figure 4), 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 [35]. MMP proteins are involved in the breakdown of ECM, particularly collagen which is the primary substrate at the invasive edge of colorectal tumors [36]. 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, 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

This work is supported, in part, by NSF CAREER Award CCF-0953195 and NIH Grant, UL1-RR024989 Supplement, from the National Center for Research Resources (Clinical and Translational Science Awards).

References

1. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A., Lusis, A.J.: An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics **37**(7) (June 2005) 710–717
2. Papin, J.A., Hunter, T., Palsson, B.O., Subramaniam, S.: Reconstruction of cellular signalling networks and analysis of their properties. Nature Reviews Molecular Cell Biology **6**(2) (February 2005) 99–111
3. Ideker, T., Sharan, R.: Protein networks in disease. Genome Res. **18**(4) (April 2008) 644–652
4. Rich, J., Jones, B., Hans, C., Iversen, E., McClendon, R., Rasheed, A., Bigner, D., Dobra, A., Dressman, H., Nevins, J., West, M.: Gene expression profiling and genetic markers in glioblastoma survival. Cancer Research **65** (2005) 4051–4058
5. Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y.V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.P.P., Duewel, H.S., Stewart, I.I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.L.L., Moran, M.F., Morin, G.B., Topaloglou, T., Figeys, D.: Large-scale mapping of human protein-protein interactions by mass spectrometry. Molecular systems biology **3** (2007)
6. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabasi, A.L.: The human disease network. PNAS **104**(21) (May 2007) 8685–8690
7. Rhodes, D.R., Chinnaiyan, A.M.: Integrative analysis of the cancer transcriptome. Nat Genet **37 Suppl** (June 2005)
8. Franke, L., Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., Wijmenga, C.: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet **78**(6) (2006) 1011–1025
9. Karni, S., Soreq, H., Sharan, R.: A network-based method for predicting disease-causing genes. Journal of Computational Biology **16**(2) (2009) 181–189
10. Lage, K., Karlberg, O.E., Størling, Z.M., Páll, Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., Brunak, S.: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology **25**(3) (March 2007) 309–316
11. Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. In: ISMB. (2002) 233–240
12. Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D., Wang, J.: Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. Bioinformatics **23**(16) (2007) 2121–2128
13. Şerban Nacu, Critchley-Thorne, R., Lee, P., Holmes, S.: Gene expression network analysis and applications to immunology. Bioinformatics **23**(7) (2007) 850–858

14. Liu, M., Liberzon, A., Kong, S.W., Lai, W.R., Park, P.J., Kohane, I.S., Kasif, S.: Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genetics **3**(6) (June 2007) e96+
15. Cabusora, L., Sutton, E., Fulmer, A., Forst, C.V.: Differential network expression during drug and stress response. Bioinformatics **21**(12) (2005) 2898–2905
16. Patil, K.R., Nielsen, J.: Uncovering transcriptional regulation of metabolism by using metabolic network topology. PNAS **102**(8) (February 2005) 2685–2689
17. Scott, M.S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D.Y., Hallett, M.: Identifying regulatory subnetworks for a set of genes. Mol Cell Prot (2005) 683–692
18. Chowdhury, S.A., Koyutürk, M.: Identification of coordinately dysregulated subnetworks in complex phenotypes. In: PSB. (2010) 133–144
19. Ulitsky, I., Karp, R.M., Shamir, R.: Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In: RECOMB. (2008) 347–359
20. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., Ideker, T.: Network-based classification of breast cancer metastasis. Mol Syst Biol **3** (October 2007)
21. Nibbe, R.K., Ewing, R., Myeroff, L., Markowitz, M., Chance, M.: Discovery and scoring of protein interaction sub-networks discriminative of late stage human colon cancer. Mol Cell Prot **9**(4) (2009) 827–845
22. Nibbe, R.K., Koyutürk, M., Chance, M.R.: An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol **6**(1) (2010) e1000639+
23. Anastassiou, D.: Computational analysis of the synergy among multiple interacting genes. Mol Syst Biol **3**(83) (2007)
24. Watkinson, J., Wang, X., Zheng, T., Anastassiou, D.: Identification of gene interactions associated with disease from gene expression data using synergy networks. BMC Systems Biology **2**(1) (2008)
25. Quackenbush, J.: Microarray data normalization and transformation. Nat Genet **32 Suppl** (December 2002) 496–501
26. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the boolean network model. Pacific Symposium on Biocomputing. (1999) 17–28
27. Koyutürk, M., Szpankowski, W., Grama, A.: Biclustering gene-feature matrices for statistically significant dense patterns. In: IEEE Computational Systems Bioinformatics Conference (CSB'04). (2004) 480–484
28. Akutsu, T., Miyano, S.: Selecting informative genes for cancer classification using gene expression data. In: Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing. (2001) 3–6
29. Shmulevich, I., Zhang, W.: Binary analysis and optimization-based normalization of gene expression data. Bioinformatics **18**(4) (2002) 555–565
30. Chowdhury, S.A., Nibbe, R.K., Chance, M.R., Koyutürk, M.: Supplement to "Subnetwork state functions define dysregulated subnetworks in cancer". (`http://vorlon.case.edu/~mxk331/crane/recomb2010_supplement.pdf`)
31. Smyth, P., Goodman, R.M.: An information theoretic approach to rule induction from databases. IEEE Trans. on Knowl. and Data Eng. **4**(4) (1992) 301–316
32. Paschos, K., Canovas, D., Bird, N.: The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. Cell Signal **21(5)** (May 2009) 665–674
33. Zucker, S., Vacirca, J.: Role of matrix metalloproteinases (mmps) in colorectal cancer. Cancer Metastasis Rev. **23(1-2)** (Jan-Jun 2004) 101–117
34. McConnell, B., Yang, V.: The role of inflammation in the pathogenesis of colorectal cancer. Curr Colorectal Cancer Rep. **5(2))** (April 2009) 69–74

35. Markowitz, S., Bertagnolli, M.: Molecular origins of cancer: Molecular basis of colorectal cancer. N Engl J Med **361(25)** (Dec 2009) 2449–2460
36. Vishnubhotla, R., Sun, S., Huq, J., Bulic, M., Ramesh, A.: Rock-ii mediates colon cancer invasion via regulation of mmp-2 and mmp-13 at the site of invadopodia as revealed by multiphoton imaging. Laboratory Investigation **87** (2007) 1149–1158