HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2017 ; 14(6): 1265–1275. doi:10.1109/TCBB.2016.2586042.

Detecting pairwise interactive effects of continuous random variables for biomarker identification with small sample size

Amin Ahmadi Adl¹, Hye-Seung Lee³, and Xiaoning Qian^{1,2,3}

¹Department of Computer Science and Engineering, University of South Florida, Tampa, Florida, 33620

²Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843

³Health Informatics Institute, Department of Pediatrics, University of South Florida, Tampa, FL 33620

Abstract

Aberrant changes to interactions among cellular components have been conjectured to be potential causes of abnormalities in cellular functions. By systematic analysis of high-throughput -omics data, researchers hope to detect potential associations among measured variables for better biomarker identification and phenotype prediction. In this paper, we focus on the methods to measure pairwise interactive effects among continuous random variables, representing molecular expressions, with respect to a given categorical outcome. Together with a comprehensive review on the existing measures, we further propose new measures that better estimate interactive effects, especially in small sample size scenarios.

We first evaluate the performance of the existing and new methods for both small and large sample sizes based on simulated datasets that shows our proposed methods outperform previous methods in general. The best performing method for small sample size scenarios suggested by simulation experiments is then implemented to estimate interactive effects among genes with respect to the metastasis outcome in two breast cancer studies based on micro-array gene expression datasets. Our results further demonstrate that integrating detected interactive effects together with individual effects can help in finding more accurate biomarkers for breast cancer metastasis, which are indeed involved in important pathways related to cancer metastasis based on gene set enrichment analysis.

Index Terms

Synergy; Interactions; Biomarker Identification

1 Introduction

Recent advancements in biochemical technologies, such as gene expression profiling and next generation sequencing, provide high-throughput measurements of molecular activities at the genome scale [1]. In clinical trials for studying complex diseases such as cancer, we are interested in predicting certain phenotypes of interest including the disease onset or treatment response based on these high-throughput -omics measurements. There have been several important findings by estimating statistical association between individual molecular activities with the phenotype or trait of interest for novel biomarker identification and early prediction of disease [2]–[10]. However, these analyses focusing on the individual or marginal effects may not be sufficient as stated both in Genome Wide Association Studies (GWAS) [4]–[6] and in gene expression analysis [11]–[13], since two variables with no significant individual effects can be highly “synergistic” and informative of the outcome when considered together [14] (Figure 1). For example, in [15], multivariate logic predictive relations among genes were studied and resulted several predictive XOR relationships, similar to what is shown in Figure 1, among which they showed that if both Interleukin18 and MRC1 genes are up-regulated or if both are down-regulated, then EHHADH gene is down-regulated. Identifying such synergistic effects among genes not only helps in identifying more accurate biomarkers for phenotype prediction, but also helps in revealing functional interactions among cellular components that are specifically related to the phenotype of interest [14], [16]. However, due to significantly increased computational complexity as well as sample size requirements when considering interactions, many of the current studies ignore interactive effects among candidate variables when analyzing for potential biomarkers, which may not be sufficient for complete understanding of complex diseases as discussed in [2], [3].

In this paper, we adopt the term **interactive effect** to refer to the statistical association between two random variables with respect to an outcome that is due to pure cooperative or synergistic effect of the two variables when considered together. For example, in statistical analysis of biomedical data, the synergistic effect is usually measured by calculating specific information theoretic measures (detailed in Section 2.1), or by the model coefficients for multiplication terms in regression or classification models as “interaction” (in Section 2.2). In other words, we are interested in measuring how much better we can predict the outcome by considering two variables together in general functional forms compared to the aggregated individual power of the variables.

To take into account such interactive effects among variables for biomarker identification, several methods have been proposed in both GWAS [4]–[6] and other -omics data analyses [11]–[13]. The performance of these methods highly depends on accurate and reliable interactive effect estimation. But it is difficult to achieve especially when the sample size is small compared to the number of variables, which is common when analyzing -omics data. Although there have been recent efforts to estimate interactive effects with different statistical models [4]–[6], [16]–[20], most of these methods either have made simplified data distribution assumptions or are limited to quantized or discrete random variables. When data (for example, micro-array gene expression data) have numerical quantities and are better characterized by continuous random variables, it is essential to avoid ad-hoc quantization of

such numerical quantities to prevent the loss of information, especially when the sample size is limited.

In order to derive more accurate and reliable biomarker identification, but without introducing too many false positives due to over-complicated models and small sample size, we focus on studying pairwise interactive effect estimation, only considering interactions between pairs of involved continuous random variables in this paper. To be specific, we represent data by $\mathbf{X}$, a n -by- p matrix, where p is the number of candidate variables and n is the number of observations or samples. Typically in biomedical data we have: $p \gg n$. Throughout this paper, we investigate methods for measuring the interactive effect between two continuous random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ with respect to the outcome variable $\mathbf{y}$, which, for example, can be binary denoting whether the corresponding samples come from “healthy” or “diseased” subjects.

We first review the previously proposed methods, especially focusing on the ones studying continuous random variables, and then further propose a set of new methods to better detect interactive effects. We have categorized existing methods together with our newly proposed methods into three categories: (1) Information theoretic methods such as “synergy” [14], [16], [18], in which marginal and conditional probability distributions of $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{y}$ are estimated from data to compute the information theoretic synergy as interactive effect. In this category, the reliability of interactive effect estimation depends on how well the involved probability distributions are estimated; (2) Classification based methods [11], [12], [19], in which first a classification model is learned from the data and then the interactive effect is measured by some parameters in the model, such as coefficients in Logistic Regression, relative values of variables in Relative Expression Analysis (REA), or estimated synergistic predictive power of $\mathbf{x}_i$ and $\mathbf{x}_j$ to $\mathbf{y}$ using classification accuracy; and (3) Association based methods, where the interactive effect is measured by the increase in association between $\mathbf{x}_i$ and $\mathbf{x}_j$ after observing the outcome $\mathbf{y}$. In Figure 2, we have laid out a hierarchy of the approaches studied in this paper with both the existing methods and our newly proposed methods.

We evaluate the performance of all the methods based on simulations by comparing their power in detecting underlying interactions among simulated random variables with data of different sample sizes. The simulation results show that our proposed methods outperform many existing methods for interactive effect estimation in general. We then take the best performing method, based on simulated datasets of small sample size, to estimate synergistic interaction among profiled genes with respect to breast cancer metastasis in two public breast cancer gene expression datasets. We show that taking into account interactive effects among genes helps in finding better biomarkers for more accurate prognosis of breast cancer metastasis than traditional biomarker identification based on only individual effects. We further show that the identified biomarkers, by considering interactive effects, contain the genes that are involved in important pathways related to cancer metastasis.

The rest of the paper is organized as follows: In the next section, we thoroughly review the existing methods for pairwise interaction detection. We explain our proposed methods for better interactive effect estimation in Section 3. In Section 4, we first test and compare the

performance of all the methods using simulated datasets with a range of different sample sizes. We then take the best performing method under the small sample size scenario and show that it can help us find more accurate and biologically interpretable biomarkers for breast cancer metastasis.

2 Literature review of existing methods

Here we review the existing methods in the literature for measuring interactive effects between two continuous random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ on a categorical outcome $\mathbf{y}$.

2.1 Information theoretic measures

One of the information theoretic measures is based on the definition of synergy $\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y})$ by Anastassiou *et al.* [16], which measures the portion of the information gain from the pair of covariates $(\mathbf{x}_i, \mathbf{x}_j)$ subtracting their individual information gain [21]:

$$\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) = I(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) - I(\mathbf{x}_i; \mathbf{y}) - I(\mathbf{x}_j; \mathbf{y}), \quad (1)$$

where $I(\mathbf{x}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$ is the mutual information measuring the information gain by the amount of reduction in the entropy of $\mathbf{y}$ after knowing the value of either individual or multiple variables. Hence, it is essential to accurately estimate $I(\mathbf{x}; \mathbf{y})$, which can be challenging under the small sample size scenario with $p \gg n$. Further expanding the formula in (1), we can rewrite the synergy based on conditional entropies: $\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) = H(\mathbf{x}_i, \mathbf{x}_j) - H(\mathbf{x}_i, \mathbf{x}_j|\mathbf{y}) + H(\mathbf{x}_i|\mathbf{y}) - H(\mathbf{x}_i) + H(\mathbf{x}_j|\mathbf{y}) - H(\mathbf{x}_j) = H(\mathbf{y}|\mathbf{x}_i) + H(\mathbf{y}|\mathbf{x}_j) - H(\mathbf{y}|\mathbf{x}_i, \mathbf{x}_j) - H(\mathbf{y})$. For quantized random variables $\mathbf{x}$ and $\mathbf{y}$, we can have relatively reliable Maximum Likelihood Estimates (MLEs) of conditional entropies $H(\mathbf{y}|\mathbf{x}) = \sum_{\mu} P(\mathbf{x} = \mu) \sum_{\nu} -P(\mathbf{y} = \nu|\mathbf{x} = \mu) \log(P(\mathbf{y} = \nu|\mathbf{x} = \mu))$ if the sample size is reasonably large [16]–[18], [20]. When $\mathbf{x}$ is continuous and $\mathbf{y}$ is discrete, we can compute conditional entropies

$H(\mathbf{x}|\mathbf{y}) = \sum_{\nu} P(\mathbf{y} = \nu) \int_{-\infty}^{+\infty} [-P(\mathbf{x} = \mu|\mathbf{y} = \nu) \log(P(\mathbf{x} = \mu|\mathbf{y} = \nu))] d\mu$; and $H(\mathbf{y}|\mathbf{x}) = \int_{-\infty}^{+\infty} [P(\mathbf{x} = \mu) \sum_{\nu} -P(\mathbf{y} = \nu|\mathbf{x} = \mu) \log(P(\mathbf{y} = \nu|\mathbf{x} = \mu))] d\mu$. Clearly, in this case, it is more difficult to have reliable conditional entropy estimates, especially with the limited sample size, which is often the case in biomedical research. We now first review the existing methods that estimate conditional entropies for continuous input variables with binary outcome.

2.1.1 Quantization methods—The first intuitive approach for continuous random variables is to quantize the data and estimate the conditional entropies for quantized random variables. Generally, the observed values of continuous random variables can be categorized into several groups for quantization. The obtained quantized variable $\mathbf{x}^c$ will be used to compute the conditional entropy $H(\mathbf{y}|\mathbf{x}^c)$ as an approximate estimate for $H(\mathbf{y}|\mathbf{x})$ and consequently to estimate $\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y})$.

One simple quantization scheme is to take the corresponding sample mean estimates

$m_i = \frac{1}{n} \sum_{\ell=1}^n x_i^\ell$ as the threshold to dichotomize the original measurements as Bernoulli random variables. There are other quantization schema. For example, instead of using mean values, we can quantize by quantiles. Also, to estimate $H(\mathbf{y}|\mathbf{x}_i, \mathbf{x}_j)$, the quantization of the pair of variables can be done simultaneously. For example, one can categorize the observed values of $\mathbf{x}_i$ and $\mathbf{x}_j$ into two groups by using the line $\mathbf{x}_i - \mathbf{x}_j = 0$ as a separating line that categorizes the data points in the $(\mathbf{x}_i, \mathbf{x}_j)$ plane into two groups: The sample points with $\mathbf{x}_i \geq \mathbf{x}_j$ and the points with $\mathbf{x}_i < \mathbf{x}_j$, which indeed is the essence of Relative Expression Analysis (REA) [12].

The main problem with those quantization methods is that the appropriate thresholds or separating criteria for quantization are generally unknown. More complicated clustering methods have been proposed for automatic quantization. For example, Anastassiou *et al.* [14] have proposed an UPGMA hierarchical clustering method to quantize continuous variables. By taking clustering results at different levels of the hierarchy, the average estimates of conditional entropies with different numbers of clusters are used as the robust interactive effect estimates [14]. Other clustering methods (e.g. in [22]) can be used for more complicated quantization of continuous variables. We will test the performance of the following specific methods in this category:

- *SynDichoMean*: dichotomize by sample mean estimates.
- *SynDichoUPGMA*: quantize by the UPGMA clustering [14]. Implementation details can be found in Section 1 of supplementary materials.

2.1.2 Simplified Gaussian assumptions—Instead of quantization, we can also estimate conditional entropies by assuming that input variables follow simple distributions as done in [18]. For example, assuming that $\mathbf{x}$ follows a Gaussian distribution (a multivariate Gaussian distribution if $\mathbf{x}$ denotes multiple variables), we can obtain the corresponding MLEs of the distribution parameters, the expectation and standard deviation in this case. Given the standard deviation σ (or the covariance matrix Σ for multivariate Gaussian), the

entropy $H(\mathbf{x})$ can be directly computed as $\frac{1}{2} \ln(2\pi e \sigma^2)$ (or $\frac{1}{2} \ln|2\pi e \Sigma|$ for multivariate cases). Based on this, we can then estimate the synergy $\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y})$. In our simulation experiments, we will test its performance and refer to this method as *SynGaussX*.

2.2 Classification-based methods

2.2.1 Logistic regression—The most commonly used statistical model for relating input variables $\mathbf{x}$ and a given binary outcome $\mathbf{y}$ is arguably the logistic regression model [23], in which the effect of a variable on the outcome can be measured by its corresponding model coefficient after fitting with the given data. Based on this, the interactive effect of two variables $\mathbf{x}_i$ and $\mathbf{x}_j$ on the outcome $\mathbf{y}$ can be computed similar to individual effect by $-\log(p)$ where p is the p-value of the coefficient β after fitting the following logistic regression model $\log(g/(1-g)) = a_0 + a_1\mathbf{x}_i + a_2\mathbf{x}_j + \beta\mathbf{x}_i\mathbf{x}_j$, in which $g = p(\mathbf{y} = 1|\mathbf{x}_i, \mathbf{x}_j)$ [19]. We later refer to this method as *LogRegCoeff* in our simulation experiments and we note that *LogRegCoeff* is a quadratic method due to the integration of the interaction terms.

2.2.2 Relative expression analysis—This method has been proposed by [12] to find interactive pairs of genes based on their relative expression. Given two input variables $\mathbf{x}_i$ and $\mathbf{x}_j$, the interactive effect associated with the outcome $\mathbf{y}$ can be measured by $IE = |P(\mathbf{x}_i < \mathbf{x}_j | \mathbf{y} = 0) - P(\mathbf{x}_i < \mathbf{x}_j | \mathbf{y} = 1)|$. The estimate for $P(\mathbf{x}_i < \mathbf{x}_j | \mathbf{y} = 0)$ is the frequency of the observations with $\mathbf{x}_i < \mathbf{x}_j$ and $\mathbf{y} = 0$ and similarly $P(\mathbf{x}_i < \mathbf{x}_j | \mathbf{y} = 1)$ is the frequency of observations with $\mathbf{x}_i < \mathbf{x}_j$ and $\mathbf{y} = 1$ by MLE. Looking deep into their method reveals that they have used the linear function $\mathbf{x}_i = \mathbf{x}_j$ to quantize data points in the two-dimensional plane $(\mathbf{x}_i, \mathbf{x}_j)$ and assumed that this line is a proper quantization boundary. Based on this assumption, they expect that for highly interactive pairs, the line $\mathbf{x}_i = \mathbf{x}_j$ better discriminates samples with larger IE. We also evaluate its power of detecting interactions in our experiments and refer to this method as *REA*.

2.2.3 Linear classifiers—For highly interactive pairs of variables, it is intuitive to expect that we can train simple classifiers that possess certain discriminating power. Based on this, a procedure based on permutation tests has been proposed in [11] to test each pair of input variables $\mathbf{x}_i, \mathbf{x}_j$ to find significantly interacting pairs by estimated classification accuracies based on Linear Discriminant Analysis (LDA). They perform two empirical significance tests: the first to rank variable pairs by estimated accuracies and the second to check whether this accuracy is only due to interactive effects of variables by permuting one variable in each pair. Based on these tests, highly ranked pairs were considered significantly interacting pairs. However, the procedure depends on the relative difference of estimated accuracies and does not provide a measure for evaluating interactive effects. Motivated by this procedure, we propose new interactive effect estimates based on the classification performance in the following sections.

3 New interactive effect estimators

3.1 New information theoretic measures

In this section, we first propose two new methods to estimate information theoretic synergy for continuous random variables based on supervised modeling.

3.1.1 Supervised quantization—Previous information theoretic measures are done in an unsupervised fashion in the sense that either quantization or distribution assumptions of $\mathbf{x}$ are made without considering $\mathbf{y}$. Hence, it may lose critical predictive information as the approximation by quantization may destroy the underlying structure of data. We explore new supervised quantization for better estimates of interactive effects by taking into account the information given in $\mathbf{y}$. Specifically, we propose new supervised quantization approaches based on existing classification methods at different levels of complexity, in which we search for the best possible separating boundaries detected by a classification algorithm for quantization. By learning a classifier, all possible separating boundaries based on the type of classifiers, for example, line (linear classifier), hyperbola (quadratic classifier), and Voronoi diagram (KNN classifier), can be explored. The separating boundary that minimizes the classification error is assumed to be the most appropriate boundary for quantizing $\mathbf{x}$. To accomplish this, we first learn a classifier to predict $\mathbf{y}$ from $\mathbf{x}$, and then for each observation $\mathbf{x}^{(b)}, 1 \leq b \leq n$, we predict its corresponding output $\mathbf{y}^{(b)} \in \{0, 1\}$ by the learned classifier,

which is considered as the new quantized value of $\mathbf{x}^{(\ell)}$. Similar to previous information theoretic measures using quantization, we then can estimate the probabilities and compute conditional entropies using quantized variable $\mathbf{x}$.

Note that depending on the classifier used, the separating boundaries may change. For example, simple modeling methods like LDA will find the best possible linear boundaries to categorize the data $\mathbf{x}$ while KNN (K Nearest Neighbor classifiers) can find more complex boundaries with increasing K but may overfit the data. Illustrative examples of how different methods may quantize $(\mathbf{x}_i, \mathbf{x}_j)$ are shown in Figure 3. We expect that the integration of the outcome information and exploring non-linear interactions among variables can capture more general interactive relationships among variables than unsupervised linear ones like the previously discussed methods [11], [12], [14]. In order to test our hypothesis, we evaluate the performance of the following supervised quantization methods:

- *SynDichoLDA*: LDA classifier (MATLAB implementation) [24] is used to find a linear separating boundary to quantize $\mathbf{x}$.
- *SynDichoQDA*: Quadratic Discriminant Analysis (QDA) classifier (MATLAB implementation) [24] is used to find a quadratic separating boundary to quantize $\mathbf{x}$.
- *SynDichoKNN*: KNN classifier is used to find a nonlinear separating boundary to quantize $\mathbf{x}$. In order to avoid overfitting, we implement a method similar to *SynDichoUPGMA* and estimate interactive effects by averaging over multiple KNN classifiers with different K . Implementation details can be found in Section 2 of supplementary materials.

3.1.2 Estimating $P(\mathbf{y}|\mathbf{x})$ by supervised learning—In this section, we propose another new way to estimate conditional entropies $H(\mathbf{y}|\mathbf{x})$ by directly estimating the $P(\mathbf{y}|\mathbf{x})$ without quantization.

Usually, classification methods are used to learn the relationships between input variables $\mathbf{x}$ and the outcome $\mathbf{y}$. After learning the classifier, for any observation $\mathbf{x}^{(\ell)}$, $1 \leq \ell \leq n$, the classifier can predict its correspondingly most probable output $\mathbf{y}^{(\ell)} \in \{0, 1\}$. In addition, many classification methods, such as Naive Bayes, Logistic Regression, LDA, and QDA, can provide estimates of the posterior probabilities $P(\mathbf{y}|\mathbf{x} = \mathbf{x}^{(\ell)})$ based on the learned classification models. For example, in LDA, to estimate the posterior probability, first the probabilities $P(\mathbf{y})$, $P(\mathbf{x})$, and $P(\mathbf{x}|\mathbf{y})$ can be reliably estimated by assuming Binomial and Gaussian distributions, respectively. Then, the posterior probability $P(\mathbf{y}|\mathbf{x})$ can be derived based on the Bayes rule: $P(\mathbf{y}|\mathbf{x}) = P(\mathbf{x}|\mathbf{y})P(\mathbf{y})/P(\mathbf{x})$. In other classification methods, including Decision Trees, Support Vector Machines (SVM), and KNN, the posterior probability $P(\mathbf{y} = \mathbf{y}^{(\ell)}|\mathbf{x} = \mathbf{x}^{(\ell)})$ for a given input $\mathbf{x}^{(\ell)}$ can be considered as a value describing how confident the classifier is on $\mathbf{y}^{(\ell)}$. For example, in KNN, if from the K nearest neighbors of $\mathbf{x}^{(\ell)}$, t of them have class label 1, then we can approximately estimate that the posterior probability

$P(\mathbf{y}=1|\mathbf{x}=\mathbf{x}^{(\ell)}) = \frac{t}{K}$ and $P(\mathbf{y}=0|\mathbf{x}=\mathbf{x}^{(\ell)}) = \frac{K-t}{K}$. With these estimates of $P(\mathbf{y}|\mathbf{x})$, we can always estimate the conditional entropies $H(\mathbf{y}|\mathbf{x})$ using following formula:

$$H(\mathbf{y}|\mathbf{x}) = \int_{-\infty}^{+\infty} P(\mathbf{x}=x) \left[\sum_{y \in \{0,1\}} -P(\mathbf{y}=y|\mathbf{x}=x) \log(P(\mathbf{y}=y|\mathbf{x}=x)) \right] dx. \quad (2)$$

But the problem with the above formula is that we need to estimate $P(\mathbf{x})$, for example, by making a Gaussian assumption or by quantization as discussed earlier. Such assumptions and approximations may lead to inaccurate estimates of the desired conditional entropies. To overcome this problem, the empirical distribution of $\mathbf{x}$ based on the given n observations can

be used: $P(\mathbf{x}=x) = \frac{1}{n} \sum_{\ell=1}^n \delta(x - \mathbf{x}^{(\ell)})$, in which $\delta(\cdot)$ is the Dirac Delta function. Substituting this empirical distribution of $P(\mathbf{x})$ into the equation (2), we get:

$$H(\mathbf{y}|\mathbf{x}) = \int_{-\infty}^{+\infty} \frac{1}{n} \sum_{\ell=1}^n \delta(x - \mathbf{x}^{(\ell)}) \left[\sum_{y \in \{0,1\}} -P(\mathbf{y}=y|\mathbf{x}=x) \log(P(\mathbf{y}=y|\mathbf{x}=x)) \right] dx. \quad (3)$$

Taking the integral into summation, we can derive

$$H(\mathbf{y}|\mathbf{x}) = \frac{1}{n} \sum_{\ell=1}^n \int_{-\infty}^{+\infty} \delta(x - \mathbf{x}^{(\ell)}) \left[\sum_{y \in \{0,1\}} -P(\mathbf{y}=y|\mathbf{x}=x) \log(P(\mathbf{y}=y|\mathbf{x}=x)) \right] dx. \quad (4)$$

Also, based on the definition of the Dirac Delta function, we have

$$\int_{-\infty}^{+\infty} \delta(x - \alpha) f(x) dx = f(\alpha), \text{ which implies that:}$$

$$H(\mathbf{y}|\mathbf{x}) = \frac{1}{n} \sum_{\ell=1}^n \sum_{y \in \{0,1\}} -P(\mathbf{y}=y|\mathbf{x}=\mathbf{x}^{(\ell)}) \log(P(\mathbf{y}=y|\mathbf{x}=\mathbf{x}^{(\ell)})). \quad (5)$$

Based on the above formula, we can efficiently estimate $H(\mathbf{y}|\mathbf{x})$ with different classification methods.

In our simulation experiments, we test the performance of the following specific methods for measuring interactive effects based on the direct estimation of $P(\mathbf{y}|\mathbf{x})$:

- *SynPosteriorLDA*: The posterior probability estimated by LDA (MATLAB implementation) [24].
- *SynPosteriorQDA*: The posterior probability by QDA (MATLAB implementation) [24].
- *SynPosteriorKNN*: The posterior probability estimated by KNN. For implementation details to avoid overfitting by KNN, please refer to Section 2 of the supplementary materials.

3.2 Methods based on classification accuracy

We further introduce new methods for estimating interactive effects based on classification accuracy. For two continuous variables $\mathbf{x}_i$ and $\mathbf{x}_j$ and the outcome $\mathbf{y}$, we measure the interactive effect as

$$IE(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) = \text{Acc}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) - \text{Acc}(\mathbf{x}_i; \mathbf{y}) - \text{Acc}(\mathbf{x}_j; \mathbf{y}), \quad (6)$$

in which $\text{Acc}(\mathbf{x}; \mathbf{y})$ is the training accuracy of a chosen classification method with input $\mathbf{x}$ and output $\mathbf{y}$. In this new measure, similar to the definition of information theoretic synergy we try to capture the interactive effect by subtracting the individual effect of individual variables. In addition, to obtain more reliable accuracy estimates, we use the posterior probability to weigh the correct and incorrect predictions as follows: Consider a vector $\mathbf{p} = [p_1, p_2, \dots, p_n]$ as the posterior probability vector obtained by the classifier for inputs $\mathbf{x} = [\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n]$, where $p_\ell = P(\mathbf{y} = 1 | \mathbf{x} = \mathbf{x}^\ell)$ and $\mathbf{y} = [\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n]$ is the actual outcome vector with $\mathbf{y}^\ell \in \{-1, 1\}$ (instead of $\{0, 1\}$). The accuracy measure used in (6) is defined as

$\text{Acc} = \frac{(\mathbf{p} - 0.5 \times \mathbf{1})\mathbf{y}^T}{n}$, in which $\mathbf{1}$ is an n -dimensional vector with all elements equal to 1. Based on the above formula, $\text{Acc} \in [-0.5, 0.5]$, where -0.5 corresponds to the cases with all inverse predictions, 0.5 for all correct predictions, and 0.0 for random predictions. Consequently, this new measure IE takes a value between -1.5 and 1.5 . We will test the following methods for measuring the interactive effect based on classification accuracy:

- *AccLDA*: The training accuracy obtained by LDA [24] is used to estimate interactive effect.
- *AccQDA*: The training accuracy by QDA [24] is used.
- *AccKNN*: The training accuracy obtained by KNN is used. For implementation details to avoid overfitting by KNN, please refer to Section 2 of supplementary materials.

3.3 Methods based on association analysis

We propose to measure the interactive effect between variables $\mathbf{x}_i$ and $\mathbf{x}_j$ with respect to the outcome by comparing the statistical association between $\mathbf{x}_i$ and $\mathbf{x}_j$ given $\mathbf{y}$, denoted by $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{y})$, and the statistical association between $\mathbf{x}_i$ and $\mathbf{x}_j$ without knowing the value of $\mathbf{y}$, denoted by $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j)$. We expect that for highly interactive pairs, $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{y})$ will be larger than $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j)$ as two variables are more statistically associated after observing the outcome. Based on this, we can measure the interactive effect as the difference between $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{y})$ and $\text{ASC}(\mathbf{x}_i, \mathbf{x}_j)$:

$$IE(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) = \text{ASC}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{y}) - \text{ASC}(\mathbf{x}_i, \mathbf{x}_j). \quad (7)$$

We can use any of the existing measures to test the association or dependence between two variables in the above formula including Pearson correlation or mutual information. Recently, a new dependence measure called Maximal Information Coefficient (MIC) has

been proposed [25]. Unlike the usual association measures which mostly look for specific functional associations (such as linear, exponential, periodic, etc.), MIC can find a wide range of relationships among variables. Using MIC as the dependence measure, we estimate the interactive effect as:

$$IE(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y}) = \text{MIC}(\mathbf{x}_i; \mathbf{x}_j | \mathbf{y}) - \text{MIC}(\mathbf{x}_i; \mathbf{x}_j). \quad (8)$$

We refer to this method for measuring the interactive effect as *MICIE* in our simulation experiments. The implementation details are given in Section 3 of supplementary materials. It is notable to say that if we use the mutual information $I(\mathbf{x}_i; \mathbf{x}_j)$ as the measure for association, the corresponding interactive effect measure in (7) is in fact equal to the synergy $\text{Syn}(\mathbf{x}_i, \mathbf{x}_j; \mathbf{y})$ by straightforward algebraic manipulations.

4 Experiments and Discussion

4.1 Performance evaluation by simulated data

We first evaluate and compare the performance of all the previously presented methods based on simulated datasets.

We simulate random datasets by extending a casecontrol model adopted in [14]. In order to generate datasets with interactive effects, we first randomly assign an outcome variable $\mathbf{y}$ uniformly distributed in $\{0, 1\}$. Based on each randomly assigned outcome value, we generate a sample of input variables as follows: We first randomly generate 50 input variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{50}$. Among these input variables, we randomly select six of them to simulate the individual effects and another six distinct pairs of other variables from all 1, 225 possible variable pairs to simulate significant interactive effects on disease outcome $\mathbf{y}$. Extending the method in [16] to capture more general interacting relationships among variables, we simulate three different types of interaction patterns between two interacting variables $\mathbf{x}_i, \mathbf{x}_j$ and the outcome $\mathbf{y}$: “simple”, “complex”, and “very complex”, which are depicted in Figure 4. These simulated interacting pairs show strong interactive effects with respect to $\mathbf{y}$ and weak individual effects. Among six pairs of interacting variables, we simulate the data by including two pairs with “simple” interactive effects; two “complex” interactive effects, and the remaining two “very complex” interactive effects. If the outcome variable is 1, we randomly assign a value drawn from an equally weighted Mixture-of-Gaussian (MoG) with two Gaussian components with the means $(-0.5, -0.5)$ and $(0.5, 0.5)$ for “simple” interacting pairs $(\mathbf{x}_i, \mathbf{x}_j)$; assign a random value drawn from an equally weighted MoG with five components with the means $(-1, -1)$, $(1, -1)$, $(0, 0)$, $(-1, 1)$, and $(1, 1)$ for “complex” interacting pairs; and finally randomly assign a value drawn from an equally weighted MoG with eight components with the respective means $(-1.5, -1.5)$, $(-1.5, 0.5)$, $(-0.5, -0.5)$, $(-0.5, 1.5)$, $(0.5, -1.5)$, $(0.5, 0.5)$, $(1.5, -0.5)$, and $(1.5, 1.5)$ for “very complex” interacting pairs. Otherwise when $\mathbf{y}^{(b)} = 0$, we assign a random value drawn from an equally weighted MoG with two components with the means $(-0.5, 0.5)$ and $(0.5, -0.5)$ for “simple” interacting pairs; assign a value drawn from an equally weighted MoG with four components with the means $(0, -1)$, $(-1, 0)$, $(1, 0)$, and $(0, 1)$ for “complex” interacting pairs; and assign a value randomly drawn from an equally weighted MoG with eight

components with their means equal to $(-1.5, -0.5)$, $(-1.5, 1.5)$, $(-0.5, -1.5)$, $(-0.5, 0.5)$, $(0.5, -0.5)$, $(0.5, 1.5)$, $(1.5, -1.5)$, and $(1.5, 0.5)$ for “very complex” interacting pairs. Furthermore, for the first three variables with individual effects, we randomly assign a value drawn from a Gaussian distribution with the mean 0.5 when $\mathbf{y}^{(b)} = 1$; otherwise the value is drawn from a Gaussian with the mean -0.5 . For other three variables with individual effects, we assign a random value drawn from a MoG with two components whose respective means are 0.5 and 1.5 when $\mathbf{y}^{(b)} = 1$ and otherwise the value is drawn from a MoG with the means -0.5 and -1.5 . Finally, for the rest of 50 variables, the corresponding values are either drawn from a two-component MoG with the means at -0.5 and 0.5 or from a four-component MoG with their corresponding means equal to -1.5 , -0.5 , 0.5 , and 1.5 regardless of $\mathbf{y}$. We have also introduced 10% noise to the outcome by flipping the simulated outcome with 0.1 probability after generating each sample. Finally, for all the Gaussian distributions in the above simulations, the standard deviation is fixed at 0.5. The above steps for simulating a single sample are summarized in Algorithm 1 in Section 4 of supplementary materials.

From this case-control model, we have simulated 1,000 datasets for each one of the following sample sizes: 20, 40, 60, 80, 100, 120, and 140. With the corresponding datasets, we first estimate interactive effects for all 1,225 pairs of variables based on different methods. Based on the Receiver Operating Characteristic (ROC) analysis of estimated interactions, we evaluate these different interaction detection methods by the AUC (area under ROC curve) values, comparing to the ground truth in simulated datasets. In other words, the variable pairs are first sorted based on their estimated interactive effects. Assuming that the top i pairs are the identified interacting pairs for a given method, true positive and false positive rates are calculated for all i ($i = 1, 2, \dots, 1225$) by comparing the top-ranked pairs with the actual interacting pairs with simulated interactive effects. Then the ROC curves are approximated based on those computed true positive and false positive rates. Finally, AUC values are calculated as the performance of the method. The average AUC values obtained by different methods from different categories are shown in Figure 5 (A–D). Furthermore, to better illustrate how different methods compare to each other, we have put together all the curves (except the obvious low performing REA) in Figure 5(E). In order to better visualize the performance differences under small sample size and large sample size scenarios, we have drawn bar plots with standard deviations for 20 samples and 140 samples in Figure 6(A).

To further verify the consistency of the results, we have performed another set of simulated experiments with more simulated input variables (200) and compared with the previous results as shown in Figure 6(B).

As the results show, the unsupervised clustering based quantization, *SynDichoUPGMA*, is able to outperform the simple mean quantization, *SynDichoMean*, when we have enough samples. In contrast, our supervised quantization method based on KNN (*SynDichoKNN*) performs better in both small and large sample size scenarios. We note that we have averaged over a wide range of K to avoid overfitting for our KNN-based methods, which helps improve the performance for the small sample size scenario.

By comparing the corresponding quantization methods with the methods based on posterior probability estimates, it is clear that further modeling the posterior probability helps improve the detection accuracy. The results also show that our proposed methods, *SynPosteriorQDA* and *SynPosteriorKNN*, outperform all other information theoretic based methods, which shows that modeling the posterior probabilities can help further improve the performance in cases with a range of different sample sizes. Finally, as the results clearly show, our proposed measurements based on classification accuracy, *AccQDA* and *AccKNN*, are the best performing methods for small sample sizes and large sample sizes, respectively. Moreover, when we have a limited number of samples (20), *AccQDA* achieves higher AUC values than *AccKNN*, probably due to its controlled model complexity compared to *AccKNN*.

If we classify the methods based on their complexity into three classes: linear, quadratic, and complex non-linear modelings, we can see that our new classification accuracy based measures always outperform the other methods in each class, which suggests the effectiveness of this new set of measures regardless of the complexity of the modeling. Based on the same classification, we can see that complex non-linear modeling methods mostly perform better for larger sample sizes while simpler quadratic modelings are better for smaller sample sizes as they have a lower chance of overfitting the data. Furthermore, as the results clearly show, all the methods based on linear models *SynDichoLDA*, *SynPosteriorLDA*, and *AccLDA* fail to capture interactive effects when the given datasets contain more general interacting relationships. Nevertheless, our proposed *AccLDA* is the best performing method among all the linear modeling methods. The performance of the *REA* method is the worst among all the methods that we have tested as it uses a fixed linear separating boundary.

We have also compared the methods based on the running time that each method takes to calculate all the interactive effects in a set of simulated datasets. Please refer to Section 8 in supplementary materials for more details.

The methods discussed in this paper analyze each variable pair respectively to estimate their potential statistical association with the outcome. Due to multiple hypothesis testing, these methods may lead to large false positives. Another way to investigate the problem is by penalized feature selection [10], [26]. For example, we can test interactive effects by simultaneously modeling all the variables in the system using the logistic regression with L_1 penalty (LASSO logistic regression) [27]. To study interactive effects by LASSO, $p(p-1)/2$ pairwise interaction terms ($x_i x_j$) in addition to p original variables can be included in the logistic regression model. As the number of terms in the model grows quadratically with respect to the number of variables, it can quickly become computationally and statistically prohibitive, especially when we have only a limited number of samples in many biomedical applications. In order to make sure that the studied pairwise methods have reasonable performance with small sample sizes compared to LASSO-based methods, we also evaluate the performance by the LASSO model and compare it with all the previous methods based on the same simulated datasets with 50 input variables. In addition to 50 original variables in the LASSO model, we also add 1,225 interaction terms to the model. We fit the LASSO model with the L_1 -norm regularization coefficient selected by a standard 10-fold

crossvalidation procedure (MATLAB implementation [28]). The absolute values of the fitted coefficients are ranked for calculating the AUC based on the ground truth in the simulated datasets. Similar to the other methods, the performance of LASSO has been evaluated for different sample sizes using the same 1,000 simulated datasets. The results are shown in Figure 5(E) together with all the other methods. As the results show, the performance of LASSO is worse than all the other non-linear methods when we have relatively small numbers of samples. Based on these results, our proposed pairwise association estimates can provide effective screening for feature selection with the consideration of potential interactions among features; following which more thorough modeling can be implemented with the reduced model complexity.

4.2 Biomarker identification with interactive effects

One application of identifying interactive effects among variables is to help find more accurate biomarkers for phenotype prediction. To overcome the high complexity of exhaustive search, traditional biomarker identification methods mostly focus on individual effects [10]. However, those methods usually miss the potential biomarkers with weak individual effects while their interactions may possess significant predictive power on the outcome. Several methods have been proposed to take into account the interactive effects among variables [11], [13], [29], [30]. Among those methods, Adl *et al.* [13] proposed a novel network-based feature ranking method that takes into account both interactive and individual effects to find accurate biomarkers for prognosis of type 1 diabetes. In this framework, individual and interactive effects from input variables are integrated in a “synergy network,” where nodes represent involved input variables and edges correspond to potential interactions among variables. The node weights are assigned based on the estimated individual effects of their corresponding variables and the edges are weighted by the estimated interactive effects. Based on this synergy network, a spectral feature ranking method that considers both individual and interactive effects has been developed to help in finding better biomarkers for phenotype prediction. In the presence of high interactive effects, the network-based ranking may rank highly the genes with high interactive effects even if they have no individual effects on the outcome. When using non-linear classifiers such as QDA and SVM, one can take advantage of those variables with high interactive effects to build more powerful classifiers.

Two micro-array datasets for Breast Cancer metastasis are used in this study. The first one is referred to as “USA” dataset [9] and contains the gene expression profiles of 286 samples, in which 107 eventually developed metastasis. The other dataset, referred to as “Netherlands” dataset [8], contains 295 samples, in which 79 eventually developed metastasis. There are a total of 6,168 genes whose expression were measured in both datasets. We only use these genes as input variables for both datasets. For the genes that are measured by more than one probe in a dataset, we use the average of the corresponding expression values.

In this set of experiments, we first adopt the network-based ranking in [13] to identify biomarkers for predicting breast cancer metastasis based on the “USA” dataset. We used *AccQDA* for measuring the interactive effects due to the small sample size in the “USA”

dataset ($p = 6,168$ and $n = 286$). We also estimate the individual effect of each variable $\mathbf{x}_i$ similarly by $\text{Acc}(\mathbf{x}_i; \mathbf{y})$.

We evaluate the performance using a standard 10-fold cross-validation evaluation method. Based on nine folds of the data (training set), we construct the “synergy network” and rank the variables by the network-based ranking method [13]. Then the top $\mathbf{T}$ variables in the ranking list are used to learn an SVM classifier (with complexity parameter $C = 100$ and RBF kernel with $\sigma = 1$) based on the same nine folds of training data. The performance of the classifier is then evaluated based on the remaining one fold (test set) by AUC. This procedure is repeated ten times for each test set. The average AUC is then reported as the estimated performance of the network-based biomarker identification. In order to show that the high performance obtained is due to integrating interactive effects, we further use the same 10-fold cross validation procedure described above to measure the performance of individual-based ranking, in which we rank the features only based on their individual effects. The results for both network-based ranking and individual-based ranking for $\mathbf{T} = 1, 5, 10, 15, \dots, 50$ are shown in Figure 7(A).

As shown in the figure, the performance of the network-based ranking, which takes into account interactive effects, is substantially better than individual-based ranking. We performed a two-sample t-test based on the two sets of AUC values (one from network-based ranking and the other from individual-based ranking). The p-values obtained for all $\mathbf{T}$ s ($\mathbf{T} \in \{1, 5, 10, 15, \dots, 50\}$) are 0.22, 0.19, 0.04, 0.09, 0.001, 0.005, 0.005, 0.05, 0.03, 0.06, and 0.15, respectively. The trend of p-values further verifies that network-based ranking outperforms the individual-based ranking quite significantly in the range of 15 to 30 top genes, which demonstrates the importance of interactive effects in identifying more accurate biomarkers and better phenotype prediction. Also, as we expect, the first few genes in network-based ranking may not show any improvement over individual-based ranking, which is due to the fact that there might not be significant interactive effects within those first few genes; but the performance of network-based ranking improves when we include more genes that have high interactive effects with the first few genes. As we perform biomarker identification 10 times during our cross-validation procedure we have ten possibly different sets of biomarkers based on different training sets, for each of the top $\mathbf{T}$ genes. To have a reliable single set of biomarkers for a given $\mathbf{T}$, we select the biomarkers that at least appear in two of the ten biomarker sets and report them as the final biomarker set. In order to make sure that those identified biomarkers are predictive of breast cancer metastasis, we test their performance on the independent “Netherlands” dataset [7], [8]. We train a SVM classifier (with the same complexity parameter $C = 100$ and RBF kernel with $\sigma = 1$) for the “Netherlands” dataset using the identified biomarkers from the “USA” dataset for each $\mathbf{T}$. To evaluate metastasis prediction performance, we perform 100 repeated 10-fold cross-validation tests. The results for the final biomarker sets identified by both individual-based and network-based ranking for all $\mathbf{T}$ s ($\mathbf{T} \in \{1, 5, 10, 15, \dots, 50\}$) are shown in Figure 7(B). As one can see, the network-based ranking not only performs better than individual-based ranking within the dataset, but also performs substantially better than individual-based ranking across the datasets, which again verifies that considering interactions among biomarkers is crucial in identifying more accurate biomarkers, which also perform stably

across datasets. We have also performed the two-sample t-test and the p-values obtained for $T = 1$ and $T = 3$ are 0.24 and 0.02 respectively, where the p-values for all the other T s are less than $1e - 5$. This further shows that the improvement obtained by considering the interactive effects among variables is statistically significant, even when the biomarkers are tested using an independent dataset.

We have further provided the final set of biomarkers for $T = 20$ in Table 1 of supplementary materials for both the network-based and individual-based rankings. We have performed gene set enrichment analysis using GSEA software [31], in which we have compared the final set of biomarkers with the curated KEGG pathways, BioCarta pathways, and the REACTOME gene set collection. The results by network-based ranking are provided in Table 1. Interestingly, the final biomarker set obtained by the network-based method has statistically significant overlap with cell junction and communication pathways and the pathways related to cell motility, which have been shown to be related to cancer metastasis [32]–[35]. In contrast, when we perform the same enrichment analysis using the biomarkers identified by the individual-based ranking method (please refer to Section 7 in supplementary materials), none of those important pathways is significantly enriched. This clearly shows that the network-based analysis, by considering the interactive effects among genes, has the potential to identify genes that may be involved in biological functions strongly related to certain disease development. As the results show, considering interactive effects can help both in providing biomarkers with higher predictive power and in revealing the underlying cellular mechanisms for disease development.

4.3 Statistical significance

All the measurements discussed in this paper are based on the statistics estimated from data. Knowing the distribution of the corresponding null hypothesis for different statistics is very important to avoid false discovery by making sure that the variable pairs with high interactive effects can be shown to have statistically significant enough interaction. For discrete random variables, [18] studied the mutual information and derived a distribution for its corresponding null hypothesis. Also, [17] studied the statistical significance for interactions among quantized variables and provided the null hypothesis distribution. However, as they discussed, the performance depends on the sample size. For all the other methods, the significance analysis has always been performed by empirically estimating the null hypothesis distribution using permutations as done in [11], [14], which is computationally demanding. These permutation methods can be similarly applied to all of our proposed methods as well.

5 Conclusions

In this study, we have comprehensively reviewed the methods provided for measuring pairwise interactive effects of continuous random variables in biomedical applications. To further improve the performance, we have proposed new methods by exploring non-linear and/or supervised modeling. We have shown that we can effectively adopt the existing supervised modeling to improve the performance of information theoretic based methods by either using them to perform the supervised quantization of continuous variables or by

directly using posterior probability estimates for interactive effect estimation. We have tested all the methods with different levels of complexity and for different sample sizes. The results show that more complex non-linear modeling like KNN are more accurate in detecting complex interactive effects. However, they require a reasonably large sample size to avoid overfitting. For small sample size scenarios, quadratic modeling with the controlled model complexity, such as QDA, can perform better and more robustly. The results also show that our classification accuracy based methods always outperform the other methods at the same level of model complexity. Furthermore, our results based on the breast cancer datasets show that considering interactive effects is actually useful in identifying more accurate biomarkers, which are biologically interpretable as they have significant intersections with important pathways related to breast cancer metastasis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project was supported in part by Award R21DK092845 from the National Institute Of Diabetes and Digestive and Kidney Diseases, National Institutes of Health.

References

1. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. Knowledge and Data Engineering, IEEE Transactions on. 2004; 16(11):1370–1386.
2. Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: Limits of models displaying no main effect. The American Journal of Human Genetics. 2002; 70(2):461–471. [PubMed: 11791213]
3. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human heredity. 2003; 56(1–3):73–82. [PubMed: 14614241]
4. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003; 19(3):376–382. [PubMed: 12584123]
5. Gayan J, Gonzalez-Perez A, Bermudo F, Saez M, Royo J, Quintas A, Galan J, Moron F, Ramirez-Lorca R, Real L, Ruiz A. A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics. 2008; 9(1):360. [PubMed: 18667089]
6. Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics. 2007; 23(1):71–76. [PubMed: 17092990]
7. Su J, Yoon B-J, Dougherty ER. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. PLoS ONE. 2009; 4:e8161. 12. [PubMed: 19997592]
8. van 't Veer LJ, Hongyue D, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530. [PubMed: 11823860]
9. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, van Gelder MEM, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005; 365(9460):671–679.
10. Saey Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19):2507–2517. [PubMed: 17720704]

11. Hoon K, Watkinson J, Anastassiou D. Biomarker discovery using statistically significant gene sets. *Journal Of Computational Biology*. 2011; 18(10):1–10. [PubMed: 21210728]
12. Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technology in Cancer Research and Treatment*. 2010; 9(2):149–159. [PubMed: 20218737]
13. Adl AA, Qian X, Xu P, Vehik K, Krischer JP. Feature ranking based on synergy networks to identify prognostic markers in DPT-1. *EURASIP Journal on Bioinformatics and Systems Biology*. 2013; 2013(1):1–9. [PubMed: 23311536]
14. 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*. 2008; 2(1):1–16. [PubMed: 18171472]
15. Pal R, Datta A, Fornace AJ, Bittner ML, Dougherty ER. Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS. *Bioinformatics*. 2005; 21(8):1542–1549. [PubMed: 15598836]
16. Anastassiou D. Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*. 2007; (1)
17. Jakulin, A., Bratko, I. Testing the significance of attribute interactions. *Proceedings of the twenty-first international conference on Machine learning; ICML '04*; 2004. p. 52
18. Wu X, Jin L, Xiong M. Mutual information for testing gene-environment interaction. *PLoS ONE*. 2009; 4(2)
19. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*. 2009; (6):392–404.
20. Varadan V, Miller DM, Anastassiou D. Computational inference of the molecular logic for synaptic connectivity in *C. elegans*. *Bioinformatics*. 2006; 22(14):e497–e506. [PubMed: 16873513]
21. Cover, TM., Thomas, JA. *Elements of information theory*. New York, NY, USA: Wiley-Interscience; 1991.
22. Coates A, Lee H, Ng AY. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS 14*. 2011
23. Armitage, P., Berry, G., Matthews, JNS. *Statistical Methods in Medical Research*. 4th Edition. 4. Blackwell Science Ltd; 2002.
24. Krzanowski, W. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press; 1988.
25. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. *Science*. 2011; 334(6062):1518–1524. [PubMed: 22174245]
26. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*. 2008; 9(5):392–403. [PubMed: 18562478]
27. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*. 1994; 58:267–288.
28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. 2009
29. Sajjadi S, Adl A, Zeng B, Qian X. Finding the most discriminating sets of biomarkers by maximum weighted clique. *Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics*. 2011; 1500
30. Sajjadi S, Qian X, Zeng B, Ahmadi Adl A. Network-based methods to identify highly discriminating subsets of biomarkers. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*. 2014; 11(6):1029–1037.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(43):15545–15550. [PubMed: 16199517]
32. Kapoor P, Saunders MM, Li Z, Zhou Z, Sheaffer N, Kunze EL, Samant RS, Welch DR, Donahue HJ. Breast cancer metastatic potential: Correlation with increased heterotypic gap junctional

- intercellular communication between breast cancer cells and osteoblastic cells. *International Journal of Cancer*. 2004; 111(5):693–697. [PubMed: 15252837]
33. Saunders MM, Seraj MJ, Li Z, Zhou Z, Winter CR, Welch DR, Donahue HJ. Breast cancer metastatic potential correlates with a breakdown in homospecific and heterospecific gap junctional intercellular communication. *Cancer Research*. 2001; 61(5):1765–1767. [PubMed: 11280719]
 34. Rosenthal DT, Iyer H, Escudero S, Bao L, Wu Z, Ventura AC, Kleer CG, Arruda EM, Garikipati K, Merajver SD. p38 γ promotes breast cancer cell motility and metastasis through regulation of RhoC GTPase, cytoskeletal architecture, and a novel leading edge behavior. *Cancer Research*. 2011; 71(20):6338–6349. [PubMed: 21862636]
 35. Nicolson GL, Dulski KM, Trosko JE. Loss of intercellular junctional communication correlates with metastatic potential in mammary adenocarcinoma cells. *Proceedings of the National Academy of Sciences*. 1988; 85(2):473–476.

Biographies

Amin Ahmadi Adl Amin Ahmadi Adl is a doctoral student in Computer Science at the University of South Florida. His research interests include Machine Learning and Combinatorial Graph Algorithms, and their applications in Computational Network Biology.

Hye-Seung Lee Hye-Seung Lee is an associate professor in the Health Informatics Institute, Department of Pediatrics and also a joint associate professor in the Department of Epidemiology & Biostatistics, University of South Florida, Tampa, FL. Her current research interests include clinical trial, longitudinal data analysis and statistical genetics.

Xiaoning Qian Xiaoning Qian (S'01–M'07) received the Ph.D. degree in Electrical Engineering from Yale University, New Haven, CT, in 2005. Currently, he is an assistant professor with the Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX. He is affiliated with the Center for Bioinformatics & Genomic Systems Engineering (CBGSE) and the Center for Translational Environmental Health Research (CTEHR) at Texas A&M. He also is a courtesy assistant professor in the Department of Computer Science & Engineering and the Department of Pediatrics at the University of South Florida, Tampa FL. He was with the Bioinformatics Training Program at Texas A&M University, sponsored by the National Cancer Institute (NCI). His current research interests include computational network biology, genomic signal processing, and biomedical image analysis.

Fig. 1.
An illustration of highly interactive effect between two variables x_1 and x_2 with respect to the outcome y . The variables x_1 and x_2 are not able to discriminate the outcome y individually.

Fig. 2. All the methods studied in this paper with their relationships illustrated in a tree graph.

Fig. 3.

Illustration for how different quantization methods work: **A:** Linear separating boundary. **B:** Quadratic separating boundary. **C:** UPGMA clustering based separating boundary. **D:** KNN based separating boundary.

Fig. 4.

Scatter plots illustrating the relationships between two interacting variables associated with the outcome y in our simulated datasets (Exaggerated) coded by different shapes: **A:** Simple interactive effect. **B:** Complex interactive effect. **C:** Very complex interactive effect.

Fig. 5. Simulation dataset results: **A:** Existing information theoretic methods. **B:** Existing classification based methods. **C:** Our proposed information theoretic methods. **D:** Our proposed classification based and association based methods. **E:** Comparison between all methods.

Fig. 6. Simulation results with 20 and 140 samples: **A:** The simulation results based on 50 variables**B:** The simulation results based on 200 variables.

Fig. 7.

The AUC values for predicting breast cancer metastasis based on network-based ranking and individual-based ranking: **A:** Results based on the “USA” dataset. **B:** Results based on the “Netherlands” dataset (An independent dataset) using the biomarkers identified from “USA” dataset.

TABLE 1

Gene set enrichment analysis for identified biomarkers by network-based ranking.

pathway	gene symbols										p-value
	CFL1	GSN	ACTR2	CLDN11	CLDN5	TESK1	UQCRC2	ATP5J	COX15		
BIOCARTA_RHO_PATHWAY ^a	✓	✓	✓							1.98e-06	
REACTOME_CELL_JUNCTION_ORGANIZATION				✓	✓	✓				2.96e-05	
REACTOME_CELL_CELL_COMMUNICATION				✓	✓	✓				1.07e-04	
KEGG_OXIDATIVE_PHOSPHORYLATION							✓	✓	✓	1.51e-04	
REACTOME_TIGHT_JUNCTION_INTERACTIONS				✓	✓					2.26e-04	

^aRho cell motility signaling pathway