Journal of proteome research

Article

pubs.acs.org/jpr

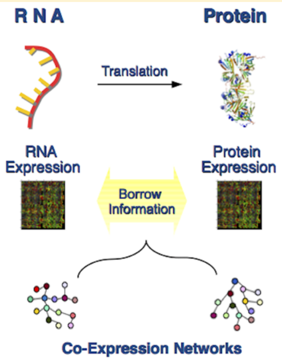
New Method for Joint Network Analysis Reveals Common and Different Coexpression Patterns among Genes and Proteins in Breast Cancer

Francesca Petralia, Won-Min Song, Zhidong Tu,* and Pei Wang*

Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, 770 Lexington Avenue, 14th Floor, NewYork, New York 10065, United States

Supporting Information

ABSTRACT: We focus on characterizing common and different coexpression patterns among RNAs and proteins in breast cancer tumors. To address this problem, we introduce Joint Random Forest (JRF), a novel nonparametric algorithm to simultaneously estimate multiple coexpression networks by effectively borrowing information across protein and gene expression data. The performance of JRF was evaluated through extensive simulation studies using different network topologies and data distribution functions. Advantages of JRF over other algorithms that estimate class-specific networks separately were observed across all simulation settings. JRF also outperformed a competing method based on Gaussian graphic models. We then applied JRF to simultaneously construct gene and protein coexpression networks based on protein and RNAseq data from CPTAC-TCGA breast cancer study. We identified interesting common and differential coexpression patterns among genes and proteins. This information can help to cast light on the potential disease mechanisms of breast cancer.



INTRODUCTION

In a recent breast cancer study conducted by the *Clinical Proteomic Tumor Analysis Consortium (CPTAC)*,[1,2] extensive global- and phospho- protein profiling were obtained for a subset of breast cancer samples that have been extensively characterized in the *The Cancer Genome Atlas (TCGA)*.[3] This is so far the first attempt to study protein activities in breast cancer samples using sophisticated protein experiments on a large scale. Then, by leveraging genomic analytical outputs from TCGA on these samples,[3] we have the unique opportunity to characterize and compare gene coexpression networks and protein coexpression networks based on the same set of samples. This is of great interest because knowledge about the common and different coexpression patterns among RNAs and proteins could improve our understanding of complicated gene regulatory mechanisms in tumor samples and could also facilitate the detection of important disease genes and therapeutical targets.

In the past decade, numerous statistical and machine learning methods have been proposed to construct gene−gene regulatory networks and protein−protein regulatory networks, including coexpression network methods,[4] Bayesian network methods,[5,6] and Gaussian graphical models.[7−10] Review of these and other methods are available in Hecker et al.[11] and Lee et al.;[12] however, all of these proposed models are designed to construct one network at a time. Presumably, we can use RNA expression data and protein expression data to construct two networks separately and then compare, but such an approach is certainly less optimal, as gene expressions and protein expressions in one tumor sample are closely related. On one hand, protein levels are regulated by their RNA levels, so we expect the two coexpression networks share common structures. On the other hand, protein activities are subject to a large amount of post-transcriptional modifications, and thus we also expect to observe unique interaction patterns in each network. This motivates us to perform joint learning of both RNA-seq and proteomic networks, so we are able to borrow information across different data sets and better capture the common correlation structure, which shall then improve the overall accuracy of network estimation.

Various methods have been proposed to jointly estimate different networks.[13−17] Some of them[13,14,17] have been specifically designed to estimate time-varying graphical models in the context of time-series data. Guo et al.[16] and Danaher et al.[15] proposed likelihood-based methods for the joint estimation of multiple related Gaussian graphical models. Their approaches rely on two different regularization schemes penalizing differences across partial correlation structures of different classes. While both papers successfully demonstrate the advantage of jointly modeling multiple networks over estimating individual networks separately, the performance of the methods heavily depends on the Multivariate Gaussian assumptions of gene expression distributions, which may not hold in real biological systems.

Special Issue: Large-Scale Computational Mass Spectrometry and Multi-Omics

Received: October 2, 2015
Published: January 6, 2016

DOI: 10.1021/acs.jproteome.5b00925
J. Proteome Res. 2016, 15, 743−754

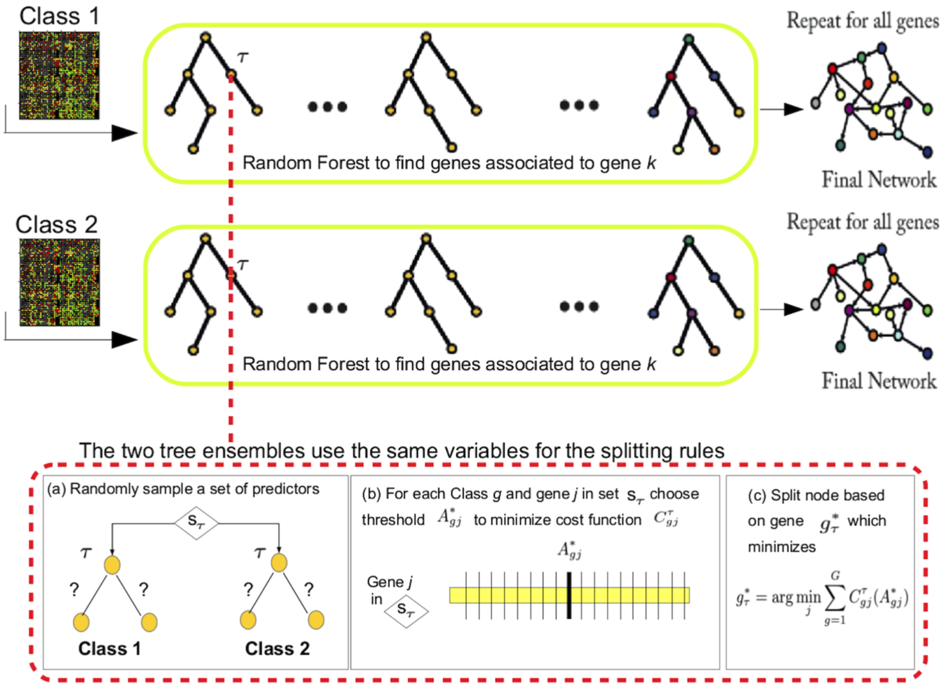

Class 1
Random Forest to find genes associated to gene k
τ
Repeat for all genes
Final Network
Class 2
Random Forest to find genes associated to gene k
τ
Repeat for all genes
Final Network
The two tree ensembles use the same variables for the splitting rules
(a) Randomly sample a set of predictors
S_τ
τ
?
?
Class 1
Class 2
(b) For each Class g and gene j in set S_τ choose threshold A^*_{gj} to minimize cost function C^τ_{gj}
A^*_{gj}
Gene j in S_τ
(c) Split node based on gene g^*_τ which minimizes
$g^*_\tau = \arg\min_j \sum_{g=1}^{G} C^\tau_{gj}(A^*_{gj})$

Figure 1. JRF schematic. For simplicity, let us assume that there are only two classes and that each data contains the same number of samples. For each target gene, g_k, we run random forest for each class and the two tree ensembles are forced to share some structure as follows. First, for each node τ, the same set of genes S_τ is randomly sampled from the entire set of genes (a). Then, the same variable in set S_τ is chosen for the splitting rule. This is achieved by deriving the splitting rule for any gene contained in set S_τ (b) and choosing the best variable to maximize the sum of decrease in node impurity over different classes (c).

Recently, random-forest-based methods have been utilized for the construction of GRN.[18−20] Random forest[21] is a decision-tree-based nonparametric method for building powerful prediction models. It has been extensively utilized for classification and regression problems.[22] Its ensemble structure combined with its nonparametric nature delivers excellent performance even when the sample size is moderate. Recently, Huynh-Thu et al.[18] proposed GENIE3, a random-forest-based method for estimating networks. The key idea is to model the expression of each target gene as a function of the expression of all other genes via random forest. The superior performance of GENIE3 was demonstrated in both DREAM 4 in-silico multifactorial challenge[23] and DREAM 5 network inference challenge.[24] In both challenges, GENIE3 and its extensions scored the best. In Maduranga et al.,[20] GENIE3 was further modified to estimate GRN based on time-series data. Recently, Petralia et al.[19] proposed iRafNet, another random-forest-based method that combines different types of heterogeneous data, such as protein−protein interactions, transcription factor-DNA binding, and gene knock-down to estimate GRN. While these random-forest-based methods greatly advance the field of GRN construction, none of them is designed to estimate simultaneously multiple related networks.

We propose JRF (Joint Random Forest), a new algorithm for the simultaneous estimation of protein coexpression and gene coexpression networks. The key idea is to borrow information across different data by forcing the class-specific tree ensembles to use the same genes for the splitting rules. In this way, regulatory relationships playing important roles in multiple classes will be detected with better power. After testing the performance of our algorithm on several in silico experiments and comparing JRF with different methods including GENIE3[18] the Joint Graphical Lasso (JGL),[15] we applied JRF for the simultaneous construction of gene coexpression networks and protein coexpression networks. For our analysis, we considered gene expression from TCGA and protein expression from CPTAC data for 62 breast cancer patients. We derived the two networks and identified protein-specific hub genes and gene modules, showing the unique contribution of proteomic data to breast cancer research.

■ MATERIALS AND METHODS

Random Forest for Network Construction

Random forest is a nonparametric algorithm that models a response variable via a collection of decision trees. Specifically, each decision tree is constructed based on a random subset of training samples, and the splitting variable at each node of a decision tree is chosen from a randomly sampled subset of predictors upon maximizing a certain utility function (e.g., decrease in node impurity).

Recently, Huynh-Thu et al.[18] introduced GENIE3, a random-forest-based model for inferring gene regulatory networks (GRNs). In GENIE3, first, for each target gene k, its expression is modeled as a function of the expression of all other genes via random forest. Then, regulatory event ($j \rightarrow k$) is measured by $I_{j\rightarrow k}$, the importance score of gene j in the random forest model of gene k. Specifically, the importance score $I_{j\rightarrow k}$ is defined as the summation of node impurities across all nodes that utilize predictor j for the splitting rule divided by the total number of trees in the random forest model of gene k.

Joint Random Forest

We start with some notations. Denote G as the total number of different classes. For each class $g \in \{1, ..., G\}$, denote the gene expression data for p genes and n_g individuals as $\mathbf{X}^g_{n_g \times p}$, and the ith observation of gene j under class g as x^g_{ij}. An overview of JRF is shown in Figure 1, where, for simplicity, we consider only two classes. The key step of JRF is to construct G random forest models simultaneously using data from G classes when predicting the expression of a target gene k based on the expression of all other genes. We propose to use the same predictor variables for splitting rules in different trees corresponding to different classes. The goal is to borrow information across different classes, so that regulatory relationships can be better detected if there are coherent signals across different classes. Specifically, when we grow G decision trees in parallel for G classes, at one node τ, we decide the splitting variable based on the following procedure:

(1) Randomly sample a set $\mathcal{S}_\tau$ containing N genes from the entire set of genes except gene k.

(2) For each class, the best splitting rule based on each candidate predictor in set $\mathcal{S}_\tau$ is derived. Let $\mathcal{P}_g^\tau$ be the subset of samples allocated to node τ under class g. The splitting rule based on the jth predictor is of the form

$$\mathcal{L}_{gj}^\tau(a_{gj}^*) = \{i \in \mathcal{P}_g^\tau\colon x_{ij}^g > a_{gj}^*\}$$

$$\mathcal{R}_{gj}^\tau(a_{gj}^*) = \{i \in \mathcal{P}_g^\tau\colon x_{ij}^g \leq a_{gj}^*\}$$

where $\mathcal{L}_{gj}^\tau$ and $\mathcal{R}_{gj}^\tau$ are the subsets of samples in class g allocated to the left and right children of node τ with a splitting rule based on the jth predictor. For each predictor in set $\mathcal{S}_\tau$, the best threshold $a_{gj}^*(\tau)$ is derived as follows

$$a_{gj}^*(\tau) = \arg\max_{\omega} C_{gj}^\tau(\omega)$$

where

$$C_{gj}^\tau(\omega) = v[\mathcal{P}_g^\tau] - v[\mathcal{L}_{gj}^\tau(\omega)] - v[\mathcal{R}_{gj}^\tau(\omega)]$$

and $v(\mathcal{P}) = \sum_{x \in \mathcal{P}} (x - \bar{x})^2$ with $\bar{x}$ representing the mean of set $\mathcal{P}$. In other words, C_{gj}^τ is the decrease in node impurity after splitting node τ in the gth tree according to gene j.

(3) Finally, gene g_τ^* is chosen as the splitting variable of node τ for all classes based on the following criteria:

$$g_\tau^* = \arg\max_j \sum_{g=1}^{G} \frac{C_{gj}^\tau(a_{gj}^*)}{n_g}$$

Because JRF utilizes the same splitting variable for corresponding nodes across all classes, predictors associated with gene k in multiple classes are more likely to be chosen for the splitting rules. For each step in the tree construction, (1−3) only apply to classes for which τ is not a final node. As in the original random forest model,[21] each tree grows until either the total number of observations allocated to the final leaves falls below a certain prespecified threshold or the maximum number of possible nodes is reached. It is worth mentioning that when the number of classes is one, JRF reduces to GENIE3, the original random forest model for network inference. It is worth noting that to implement JRF, data sets need to be standardized to mean zero and unit variance. Importance scores depend on the scale of the data and therefore variables (genes) need to be standardized before the random forest models are fitted.

Assessing JRF Performance

On the basis of the previously described procedure, JRF constructs G random forest models simultaneously for each target gene k and returns a ranking of gene−gene interactions based on importance scores. To assess the performance of JRF in predicting the true interactions, receiver operating characteristic curves (ROC) and precision-recall curves can be computed by setting different thresholds on importance scores. In this paper, JRF is compared with JGL,[15] GENIE3-Sep, and GENIE3-Comb on several in silico experiments. GENIE3-Sep is GENIE3 used to estimate networks based on data from G classes separately, while GENIE3-Comb is GENIE3 used to estimate a unique network based on the union of data from all classes. All random-forest-based algorithms (JRF, GENIE3-Comb, and GENIE3-Sep) provide a ranking of regulatory relationships based on importance scores, and ROC curves were computed by setting different thresholds on these scores. Instead, for JGL, ROC curves were constructed by considering different values for the two parameters controlling the level of sparsity (as shown by Figure S1 in the Supporting Information)
.

Another approach to evaluate the performance of JRF is choosing a proper cutoff value for importance scores using permutation techniques. Let $I_{j\to k}^g$ be the importance score associated with the regulatory event ($j \to k$) in the gth class-specific tree ensemble. Specifically, this importance score is defined as $I_{j\to k}^g = \frac{1}{T} \sum_{\tau \in \mathcal{N}_{gj}} C_{gj}^\tau$ where T is the number of trees and $\mathcal{N}_{gj}$ is the set of nodes which utilize gene j for the splitting rule in the tree ensemble used to predict gene k based on the gth class-specific data. Because in this paper we are interested in estimating undirected networks, we derive importance score I_{j-k}^g for every edge (j−k) as the average between importance scores $I_{k\to j}^g$ and $I_{j\to k}^g$, and the final undirected networks can therefore be derived by applying a cutoff on the edge importance scores. To derive a proper cutoff value for importance scores, we utilize the following permutation-based procedure:

(a) For $b \in \{1, \cdots, B\}$, with B being the number of permutations:

(a.1) For any target gene k, we first permute its sample order within each class and fit G random forest models via JRF to predict the expression of gene k based on the expression of all other genes in G classes, respectively.

(a.2) Repeat (a.1) for all genes and compute the final importance scores for each edge in each class, which are denoted as $\{\mathbf{I}_{p\times p}^{g(b)}\}_{g=1}^G$.

(b) For each threshold ι, we compute

$$f(\iota) = \frac{\frac{1}{B} \sum_{b=1}^{B} \sum_{j\neq k} \mathbf{1}(I_{j-k}^{g(b)} > \iota)}{\sum_{j\neq k} \mathbf{1}(I_{j-k}^g > \iota)}$$

where $\mathbf{1}(\bullet)$ is the indicator function, equal to one if event $\bullet$ occurs and zero otherwise.

$f(\iota)$ can serve as an approximation of the false discovery rate (FDR).[25] In the following numerical studies, we use $\iota_0 = \min\{\iota\colon f(\iota) \leq 0.001\}$ and declare an edge between j and k in class g if $I_{j-k}^g > \iota_0$.

Computational Complexity

The computational complexity of JRF is $O(pTN\sum_{g=1}^G \log(n_g)n_g)$, where p is the number of genes, T is the number of trees in each forest, N is the number of variables sampled at each node, G is the number of classes, and n_g is the number of samples in each class. This complexity is in the same order as the complexity of applying GENIE3 to estimate G networks for G classes separately. In the contrast, the time complexity of GGM-based approaches, such as JGL, often has the order of

p^3.[15] So when p is large, JRF and GENIE3 enjoy better computational efficiency, which is supported by our numerical investigations (see Supporting Information).

In practice, the number of trees (T) and the number of potential regulators to be sampled at each node (N) are user-specified parameters. T is usually chosen sufficiently large because the tree ensemble provides more accurate results as T increases. In our numerical studies, we reported results for T = 1000. We also evaluated larger T values, such as 2000 and 5000, but observed similar results (data not shown). For N, a common choice is $N = \sqrt{p-1}$ where $(p-1)$ is the number of predictors for each target gene. In general, large value of N results in predictions with high bias; while low value results in predictions with high variance.

We implemented an R package JRF using C routines wrapped with an R interface. The R package is publicly available through CRAN. Codes and a tutorial for JRFs implementation can also be downloaded from http://research.mssm.edu/tulab/software/JRF.html. Similarly to other tree-based algorithms for network construction, JRF can be easily parallelized because the estimation process consists of p-independent subproblems. A comparison between JRF, JGL, GENIE3-Sep, and GENIE3-Comb in terms of computational time is presented in Table S1.

Data

Synthetic Data. For data generation, we first considered two networks containing 500 nodes connected by 498 and 249 edges, respectively. In particular, Network 1 is a network with two disjoint components and Network 2 consists of only the first component of Network 1. In this example, we focused on two network topologies: power law[10,26] and star topology (Figure S2 in Supporting Information). For each network topology, we simulated 20 replicates involving n = 50 samples from a Gaussian graphical model. Network structures and covariance matrices were simulated in the same way as Danaher et al. (2014)[15] (further details can be found in the Supporting Information). Besides this example, we also consider a series of other different simulation settings, including (1) two non-nested networks; (2) five non-nested networks; (3) two nonrelated networks; and (4) data from non-Gaussian distributions. Please see Supporting Information sections 1.3−1.7 for more details. For each data set, variables were standardized to mean zero and unit variance.

Breast Cancer Data. The expressions of 5864 proteins for 62 breast cancer tumors[27] (proteome-ratio-norm-noNA-normal.gct(04-Jun-2014 version)) were downloaded from the CPTAC Data Coordinating Center (http://proteomics.cancer.gov/programs/cptacnetwork) sponsored by the National Cancer Institute. For three samples having technical duplicates, we took the average abundance of each protein across the duplicates. We further standardized protein expression so that each sample had median 0 and median absolute deviation (MAD) 1. Finally, we excluded proteins with interquartile range <70% quantile. The resulting data matrix consisted of 1759 proteins. Level-three RNAseq data of breast tumor samples were obtained from the TCGA Web site (http://tcga-data.nci.nih.gov/tcga/). First, we log-transformed the data and then replaced all missing values with 0. Then, we standardized each sample to have median 0 and MAD 1. We focused on the 62 samples contained in the proteomic data and excluded genes with more than 10% missing values. Then, we selected the top 10% genes with largest interquartile range across samples. The resulting data matrix based on RNAseq data contained 1464 genes. As final subset, we considered 528 genes obtained by overlapping the set of 1464 selected genes and the set of 1759 selected proteins. For each data set, genes and proteins were standardized to mean zero and unit variance.

RESULTS AND DISCUSSION

Synthetic Data

In this section, JRF was compared with JGL,[15] GENIE3-Sep, and GENIE3-Comb on several in-silico experiments. In all implementations, we used T = 1000 and $N = \sqrt{p-1}$. We evaluated the performances of different algorithms based on different network structures, sample sizes, and number of genes.

Network Structure. Figure 2 shows the average ROC curves of each method over 20 replicates for Network 1 and Network 2 separately. JRF outperforms all other methods under all different settings, while GENIE3-Sep is the second best performer. Figure S3 shows the boxplot of the area under the precision-recall curve (AUPR) for JRF, GENIE3-Sep, and GENIE3-Comb. As shown, JRF outperforms competitors also in terms of AUPR. The advantage of JRF over GENIE3-Sep is more evident for Network 2, which contains only shared edges. This is expected because JRF should be able to detect common edges with better power.

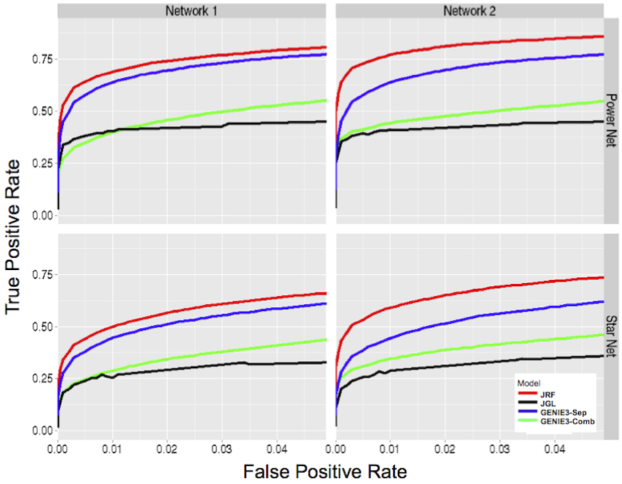


Figure 2. Mean of ROC curves for Network 1 (first column) and Network 2 (second column) over 20 replicates for JRF (red), GENIE3-Comb (green), GENIE3-Sep (blue), and JGL (black). For each replicate, we sampled 50 samples from a Gaussian graphical model based on the power law topology (first row) and star topology (second row).

To further assess the performance of the proposed model, we included more simulation scenarios in the Supporting Information. In particular, in Supporting Information section 1.3, we considered the case of non-nested networks sharing some structure, while, in section 1.4, we considered the case where five different networks were estimated simultaneously. As shown by Figures S4 and S5, JRF outperformed competitors under all of these simulation scenarios. In addition, it would be interesting to evaluate the performance of JRF when class-specific networks do not share any structure. For this purpose, we simulated two different networks with no common edge and applied different methods on the corresponding data (Supporting Information section 1.5). As shown in Figure S6, performance of JRF is very comparable to that of GENIE3-Sep on this simulation example. This suggests that JRF is sufficiently robust and works in the absence of common structure. Moreover, to assess the performance of JRF in the presence of nonlinearities in the data, we further generated data examples using GeneNetWeaver,[28] an open-source software for the generation of in-silico data from a set of ordinary and stochastic differential equations (Supporting Information section 1.6). As shown in the supplemental Figure S7 and S8, JRF again outperformed other algorithms in terms of both area under the ROC curve (AUC) and area under precision-recall curve (AUPR). Finally, Figure S9 shows results for different algorithms in estimating networks with 1000 nodes (further details can be found in Supporting Information section 1.7), in which we can see that JRF outperformed competitors in terms of both AUC and AUPR. It is worth mentioning that the increased dimensionality did not negatively affect results, suggesting that JRF can handle problem with high dimensionality.

Effect of Sample Size. In this section, we assess the performance of JRF and other algorithms on data sets involving different number of samples. Figure 3 shows boxplots of AUC over 20 replicates resulting from JRF, GENIE3-Sep, and GENIE3-Comb for different sample sizes, that is, n = 50, 100, 200. The AUCs of JRF are significantly larger than that of GENIE3-Sep based on 20 replicates under all sample sizes. As shown in the Supplemental Figure S10, JRF outperforms competitors also in terms of AUPR for different samples sizes. As shown, the smaller the sample size the bigger the improvement of JRF over GENIE3-Sep is. This result is expected because as the sample size increases GENIE3-Sep should have sufficient information to estimate the networks separately.

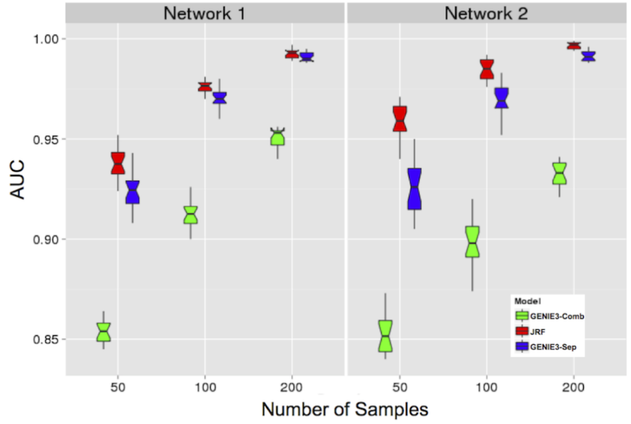


Figure 3. Boxplot of AUC over 20 replicates for JRF (red), GENIE3-Comb (green), and GENIE3-Sep (blue) for different sample sizes, that is, n = 50, 100, and 200. Data: samples simulated from a Gaussian graphical model on a power law topology.

Edge Specificity. In this section, we compare JRF and GENIE3-Sep regarding their abilities to estimate class-specific edges. We considered the 20 replicate data sets (n = 100), which was simulated based on the power law network topology in the previous section. For both JRF and GENIE3-Sep, edges were declared using the permutation procedure illustrated above with B = 500 and an FDR threshold of 0.001. Table 1a shows the performance of JRF and GENIE3-Sep in estimating Network 1 and Network 2. For each network, we reported the minimum and the maximum value of true positive rate (TPR), false-positive rate (FPR), and false discovery rate (FDR) across 20 replicates. In particular, these quantities are defined as $TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, and $FDR = FP/(FP + TP)$, where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives. For both networks, JRF achieved better power (TPR) than GENIE3-Sep, while the resulting FPRs were quite comparable. Table 1b shows a comparison of the two algorithms regarding the detection of class-specific (differential) edges. Again, we computed the rate of true differential edges and false differential edges and showed the minimum and the maximum value of these quantities over 20 replicates. While the TPRs for detecting differential edges were similar for the two methods, the FPRs of GENIE3-Sep were significantly larger than that of JRF. In addition, JRF leads to substantially lower FDRs for detecting differential edges than GENIE3-Sep. This again suggests the merit of joint learning in JRF.

Coexpression RNA and Protein Networks in Breast Cancer

We applied JRF to construct coexpression networks based on the CPTAC global proteomics data set and TCGA RNAseq data of 62 breast cancer samples.

Networks via JRF. For simplicity, we will refer to the coexpression networks constructed by JRF from global proteomics data and RNAseq data as protein-network and RNA-network. At FDR cutoff of 0.001, we detected 687 common edges shared by both networks, 502 edges unique to protein-network, and 382 edges unique to RNA-network. The topologies of the two networks are shown in Figure 4, while the full list of interactions can be found in Additional File 1 in the Supporting Information. Interestingly, a few network modules in protein-network have a high percentage of edges unique to protein-network. No such structures is observed in RNA-network. Figure 5 highlights genes whose degrees (i.e., total

Table 1. (a) *TPR*, *FPR*, and *FDR* of Network 1 and Network 2 and (b) *TPR*, *FPR* and *FDR* of Differential Edges (Diff.Net)[a]

			TPR		*FPR*		*FDR*	
	model	network	min	max	min	max	min	max
(a)	JRF	Net 1	0.64	0.69	4e-4	7e-4	0.13	0.20
		Net 2	0.68	0.78	2e-4	6e-4	0.15	0.26
	GENIE3-Sep	Net 1	0.62	0.66	4e-4	6e-4	0.12	0.20
		Net 2	0.57	0.67	2e-4	6e-4	0.11	0.27
(b)	JRF	Diff.Net	0.56	0.65	1e-4	2e-4	0.09	0.18
	GENIE3-Sep	Diff.Net	0.59	0.67	3e-4	4e-4	0.19	0.29

[a] For each quantity (TPR, FPR, and FDR) we show the minimum value (Min) and the maximum value (Max) across 20 replicates. Note that the true discovery rate (TDR) is defined as $TDR = (1 - FDR)$.

DOI: 10.1021/acs.jproteome.5b00925

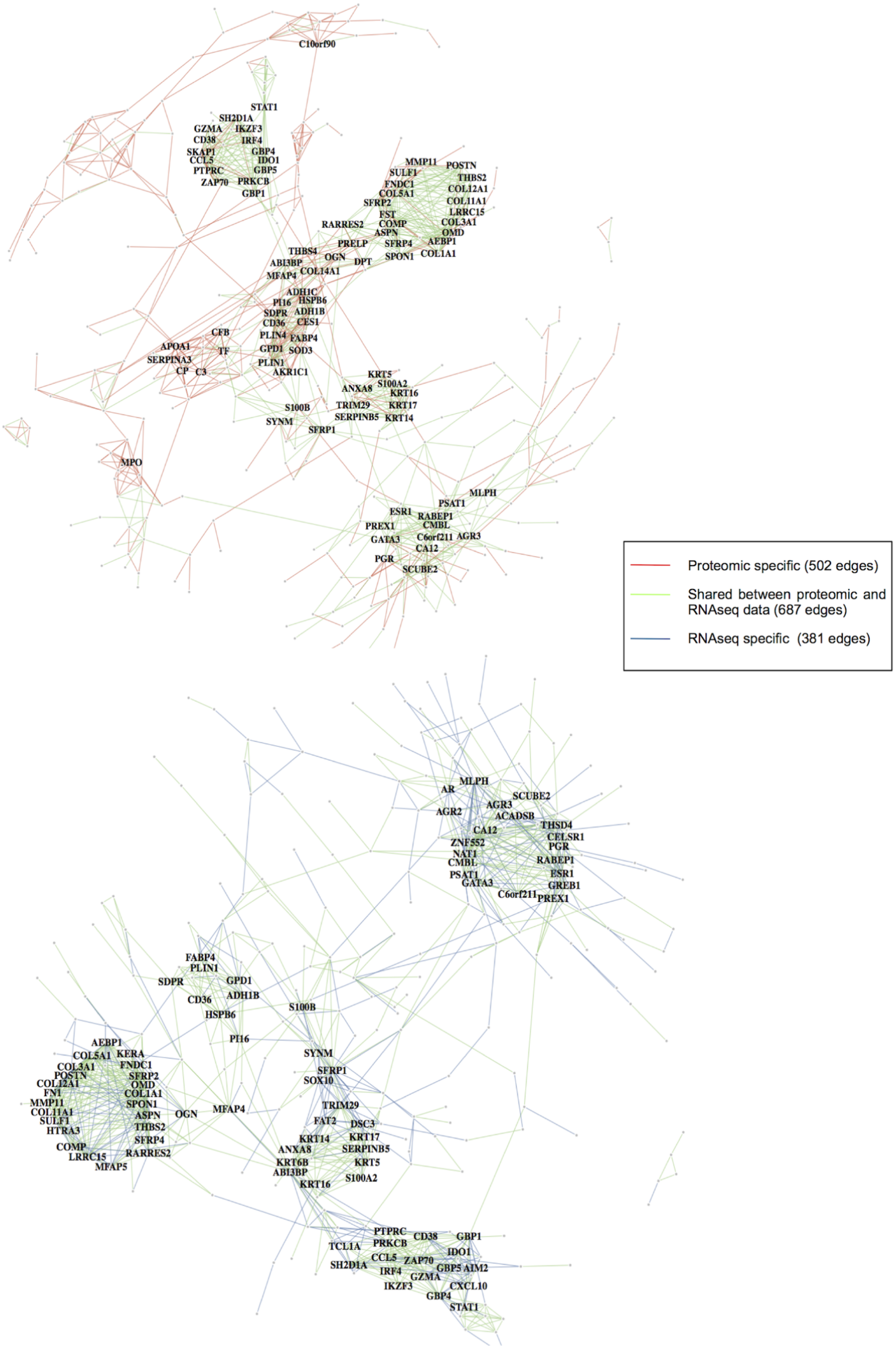


Figure 4. Network based on proteomic data (a) and RNAseq data (b) resulting from JRF. Green edges (687) are shared between proteomic and RNAseq data, red edges (502) are unique to proteomic data, while blue edges are unique to RNAseq data (382). Only names of genes with at least ten connecting edges are shown in the plot. For both networks, the full list of interactions can be found in additional file 1.

number of edges connecting to one gene) in protein-network are much higher than that in RNA-network. Three out of the top five genes with more protein-unique edges, C3, CFB, and MPO, are related to immune response functions. It is well known that immune response plays a major role in the patient response to chemotherapy treatments.[29] Some of these genes have been shown to be predictive of survival outcomes of breast cancer patients. For example, C3 and CP are predictive of chemoresistance in breast cancer patients,[30] and high activity of MPO genotypes can enhance efficacy of chemotherapy for early-stage breast cancer.[31] Other remarkable proteins are MMP8 and MMP9. Genes belonging to the MMP family have

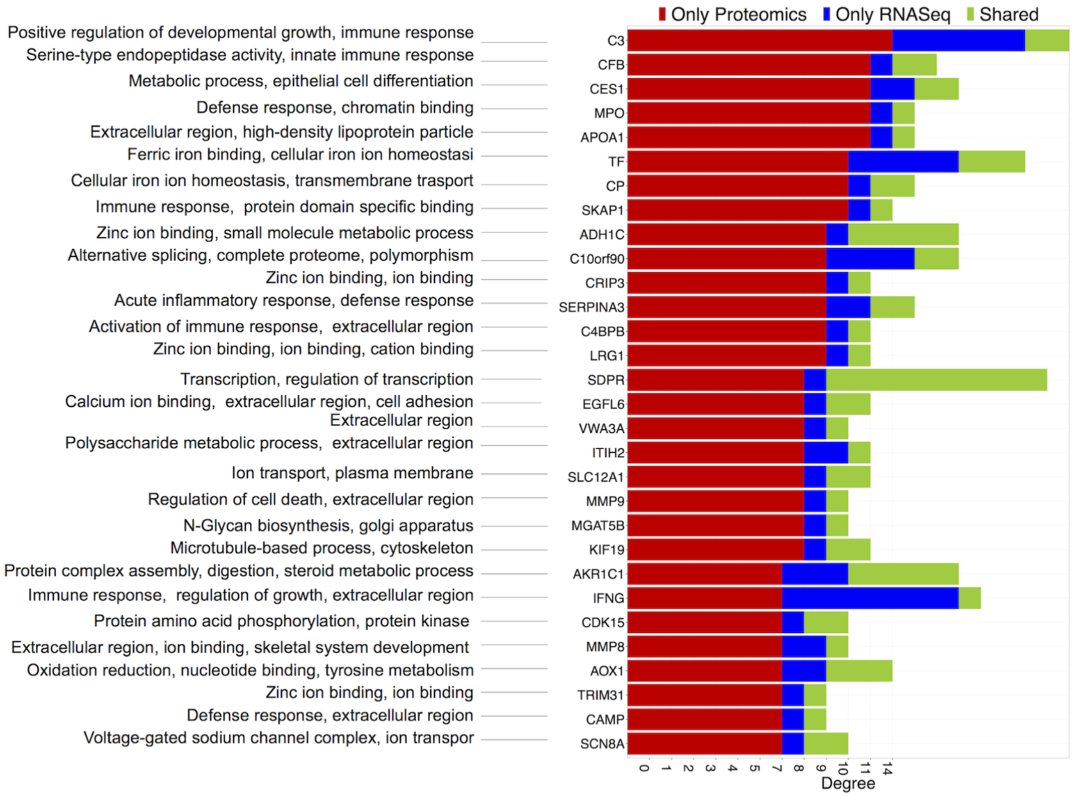


Figure 5. Number of connecting edges for top connected genes in the network based on proteomic data. For each gene, the green bar corresponds to the number of connecting edges shared between proteomic and RNAseq data; the red bar indicates the number of Protein-specific edges; while the blue bar indicates the number of RNA-specific edges. For each gene, we list its biological functions.

been linked to cancer progression for their ability to degrade the extracellular matrix and may be implicated in the formation of metastasis.[32,33]

We then focused on network modules unique to protein-network. We derived network modules based on edge betweenness using function "edge.betweenness.community" available in the R package igraph.[34] In Figure 6, we show the protein-network with gene modules enriched of at least one GO category. The full list of genes contained in each module can be found in additional file 1. Figure 6 shows, for each module, the most enriched GO category and the corresponding Benjamini adjusted p-value.[35,36] Two interesting protein-specific modules (with more protein-specific edges than shared edges) are C1 and C2. C1 is enriched of "fibroblast growth factor receptor activity", as shown in Figure 7, genes are more tightly correlated in the proteomic data compared with the RNAseq data. Various studies have investigated the role of the FGFR pathway as a predictive marker for breast cancer.[37,38] As shown in Figure 7, FGFR2 and FGFR3 show a correlation close to one. Recently, Cerliani et al.[39] reported a strong correlation between FGFR2 and FGFR3 based on protein expression mentioning that there is no previous evidence of correlation between these two proteins for breast cancer patients. In addition, the interaction between FGFR2 and FGFR3 is contained in the STRING database[40,41] (see Table S2). Another interesting protein-specific module is C2. This module is enriched of "extracellular region". Similarly to module C1, genes in module C2 are higher correlated in the proteomic data compared with the RNAseq data (Figure 7). This cluster contains genes such as APOB, C3, ORM1, and CP that have been recently shown to be predictive of chemoresistance in breast cancer patients.[30] As shown by Figure 7, protein expressions of these genes are highly correlated. Another important gene in this module with a high number of connecting edges is APOA1 (the fifth highest connected node according to Figure 5). This gene has been recently identified as a potential target for disease progression using whole-genome trancriptomic and whole proteomic.[42] As shown in Table S2, 11 of the relationships in Figure 7 are contained in the STRING database. These results suggest that proteomic data can provide complementary information to genomics data and help to reveal important biological mechanisms.

Comparison with GENIE3-Sep. Because we are interested in assessing the utility of a joint learning, JRF was compared with the standard random-forest algorithm that constructs the two networks separately. For both methodologies, we derived undirected networks using the same FDR cutoff (0.001). The protein-network resulting from GENIE3-Sep can be found in Figure S11 in the Supporting Information. As shown, 57% of the edges are shared across protein- and RNA-networks under JRF, while only 26% are shared under GENIE3-Sep. This finding can be explained by the fact that a joint learning can facilitate the detection of common edges.

a. Validation Using GO Terms. In this section, we validate our findings using GO categories. In particular, we consider 596 GO terms containing fewer than 200 genes (the list of GO terms can be found in Tables S3−S6 in the Supporting Information). Figure 8a shows the number of edges contained in at least one GO term for different FDR cutoff for both protein-networks. Figure 8b shows the enrichment p-value for different FDR cutoff. The enrichment p-value was calculated as follows. First, we constructed a network based on GO terms

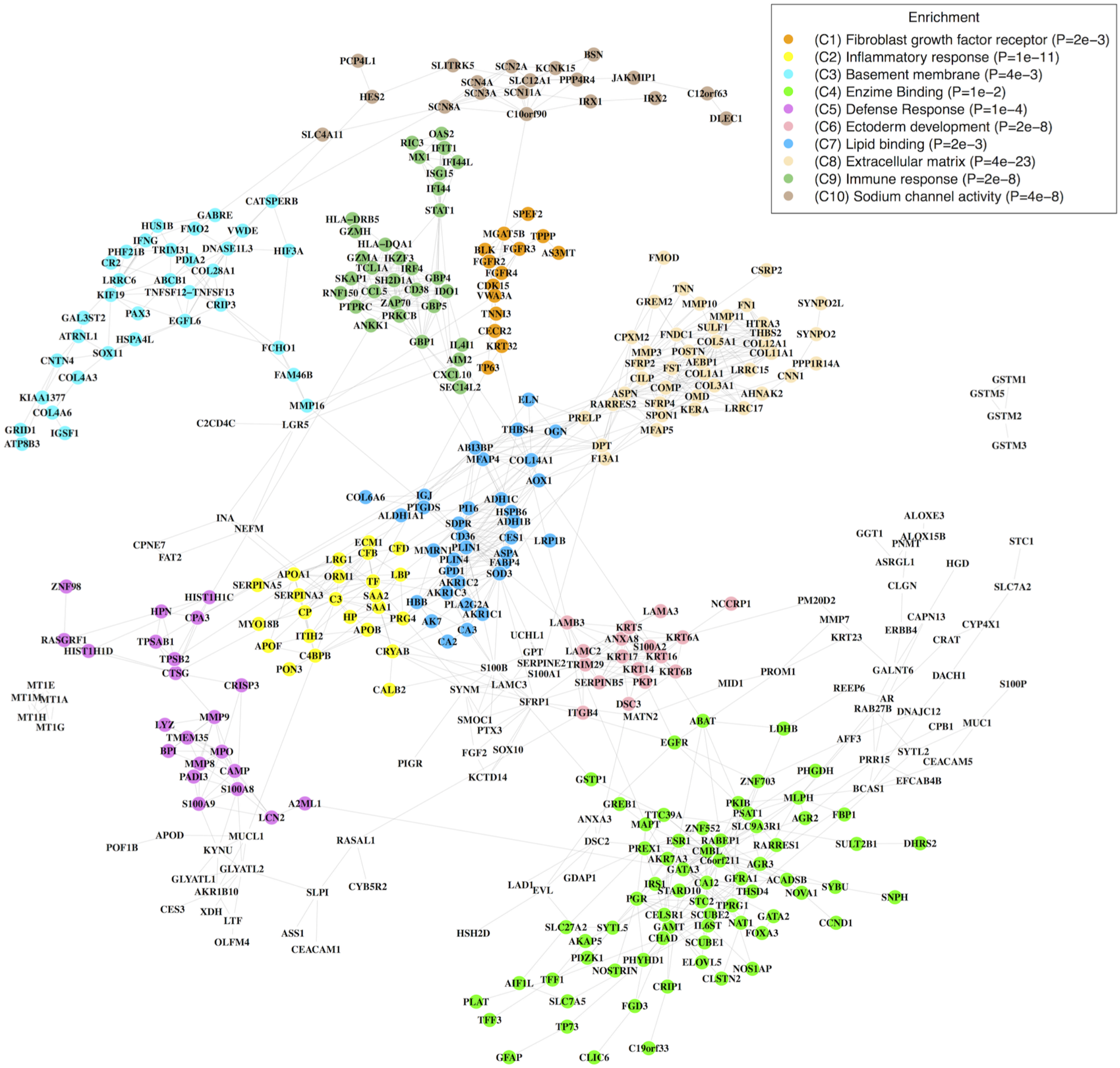


Figure 6. Network based on proteomic data resulting from JRF with corresponding gene modules. Modules were detected using function "edge.betweenness.community" available in the R package igraph. For each module, we show the most enriched GO category with corresponding Benjamini adjusted p-value (P). In particular, only modules reporting a significant enriched Benjamini adjusted p-value ($P \leq 0.01$) are highlighted.

where an edge $(a - b)$ was drawn between every pair of genes (a, b) sharing at least one GO term. Then, a contingency table was constructed considering variables "network based on GO terms" and "network based on proteomic data" with categories "number of edges contained" and "number of edges non-contained", and a Fisher's exact test was performed to derive enrichment p-values. Figure 8a,b shows that JRF results in a network more overlapping with GO terms than the standard random-forest algorithm GENIE3-Sep.

b. Validation Using Esr1 and Gata3 Knockdown Signatures. To further validate our findings, we overlapped both networks with Esr1 and Gata3 knock-down signatures. We identified siRNA knock-down signatures of key transcription factors (TFs) of breast carcinoma in MCF7 cells1 by accessing gene expression data available under GSE31912 in Gene Expression Omnibus[43] (further details about these signatures can be found in Section 2 of Supporting Information). For each network, a neighborhood was defined by defining a threshold k for the number of edges connecting a particular gene to the target gene of interest (either Esr1 or Gata3). For each network in Figure 8, we considered different values of k and counted the number of knockout signatures contained in the neighborhood. Figure 8c,d shows the total number of signatures contained in k-size neighborhoods for Esr1 and Gata3, respectively. As shown, also in this case, JRF results in a more enriched network. Figure S12 contains the same comparison for RNA networks.

Discussion

In this paper, we developed JRF, a random-forest-based algorithm for the simultaneous construction of gene coex-

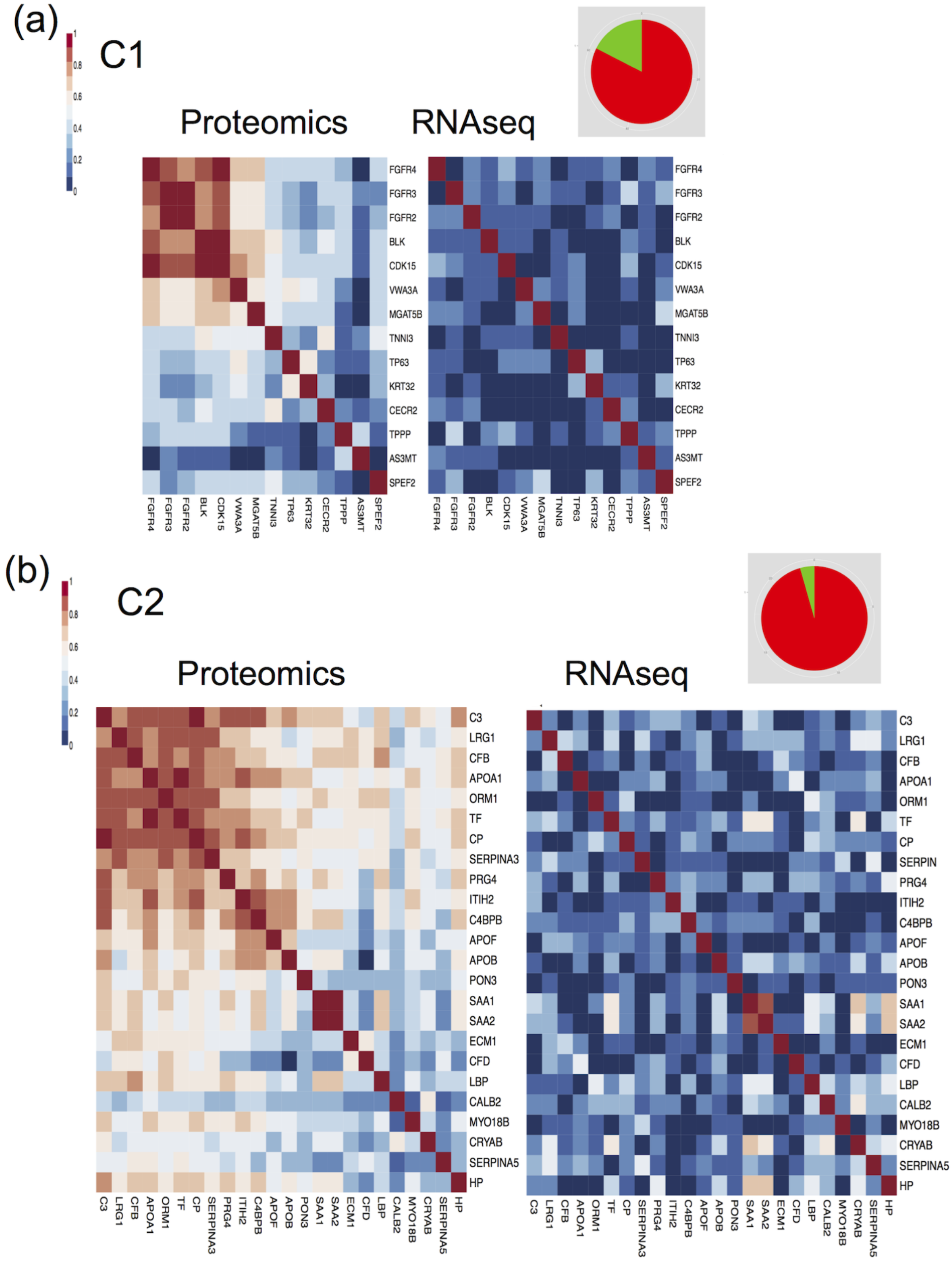


Figure 7. Heatmap of the absolute value of correlation between genes based on proteomic and RNAseq data for module C1 (a) and C2 (b) in Figure 6. For each module, we include a pie plot showing the number of protein-specific interactions (red) and interactions shared across proteomic and RNAseq data (green).

pression and protein coexpression networks. JRF is designed to borrow information across different expression data by selecting the same set of predictors (genes) as splitting variables in the random forest model corresponding to each data. In this way, JRF is able to detect common relationships (edges) with better power and detect differential edges-specific to individual data with fewer false positives.

Because in this study we were mainly interested in assessing the unique contribution of proteomic data to breast cancer research, we focused our attention on protein-specific hub genes and modules. We identified two interesting protein-specific modules containing potential targets for breast cancer treatment. In addition, we compared our algorithm to the original random-forest algorithm, which constructs the two networks separately, and showed that our algorithm leads to

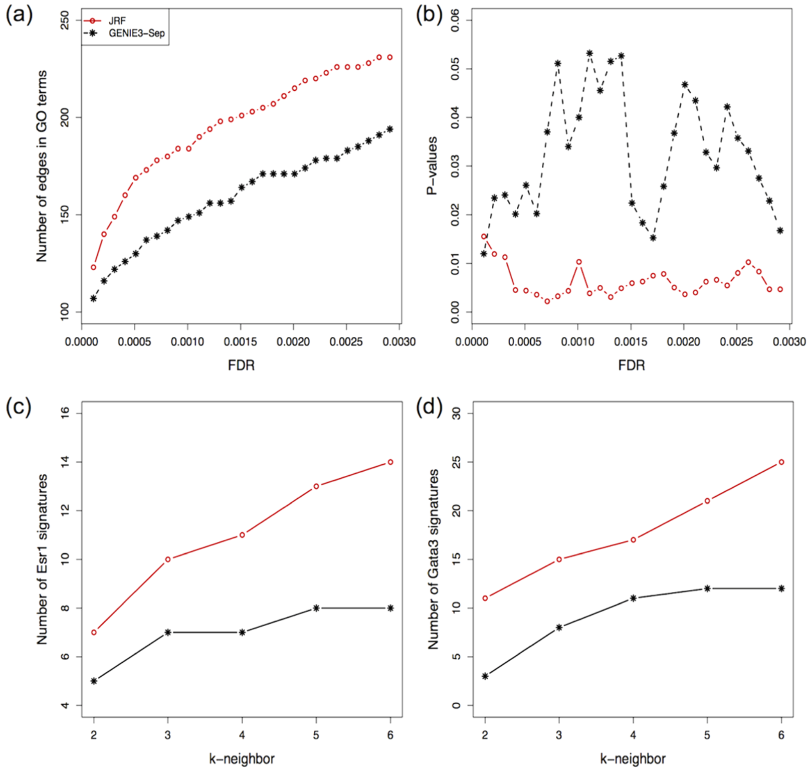


Figure 8. We validate protein-networks in Figure 4 using GO terms and knockout signatures. The first row contains network validation based on GO terms. For our analysis, we considered 596 GO terms containing fewer than 200 genes. (a) Number of edges contained in at least one GO term for different FDR cutoff. (b) *p* value of enrichment for different FDR cutoff. The second row contains validation based on Esr1 (c) and Gata3 (d) signatures. For a particular neighborhood size of either Esr1 or Gata3, we show the total number of knockout signatures contained under JRF (○−) and GENIE3-Sep (★−).

networks more overlapping with available knockout signatures and existing GO terms.

The current version of JRF requires the feature (protein) space to be the same across different classes. Extension of JRF for handling data involving different sets of variables remains as future work. This problem arises in the protein domain because protein abundance can be measured for different phosphorylation-sites that map to a unique protein. As future work, we will design a model able to jointly estimate networks from phosphorylation-site abundance and RNAseq data, simultaneously.

Besides estimating gene-regulatory network and protein-regulatory network, as shown in the paper, JRF can be used in many other applications. For example, we can apply JRF to estimate GRNs for the three breast cancer subcategories (luminal, basal, and her2), GRN for different tissues, and GRN for cancer and normal tissues. Another interesting direction would be assessing the association between miRNAs-genes and miRNAs-proteins simultaneously. miRNAs are molecules that control the growth and proliferation of a cell. It is well known that the downregulation of some miRNAs may play an important role in the progression of cancer.[44] Therefore, a better understanding of the interactions between miRNA, mRNA, and protein expression is crucial to cast light on the potential disease mechanisms of cancer. In this context, JRF can be easily utilized to jointly detect which miRNAs regulate genes and proteins.

Another advantage of JRF relies on its computational efficiency. JRF can be easily parallelized because the computation can be divided into *p*-independent subproblems. On the contrary, for many existing methods such as Bayesian networks and GGM-based algorithms, it is difficult to parallelize their computation. For this reason, JRF could be more preferred when handling large data sets.

The merit of JRF depends on the assumption that networks of different classes share some common structures. To assess the performance of JRF in the absence of common structure, we conducted numerical studies to investigate such cases and concluded that JRF is sufficiently flexible to guarantee good performance even when nonrelated networks are considered.

In JRF, importance scores derived from random forest models are used to rank edge strengths. We propose a permutation-based procedure to derive proper cutoffs on importance scores for selecting confident edges. Specifically, we introduce $f(\imath)$ as an approximation of the false discovery rate (FDR) of the selected edge set based on cutoff $\imath$; however, in the numerical studies, we observe that $f(\imath)$ tends to underestimate FDR. Thus, a conservative threshold on $f(\imath)$ is recommended in practice. New methods to obtain more accurate estimate of FDR for JRF warrant future research.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00925.

The PDF file includes more results from both synthetic data and breast cancer data and is divided into two sections: "In-silico experiments" and "Coexpression RNA and protein networks in breast cancer". Section "In-silico experiments" contains nine subsections: "JGL parameter specification", "Network topology", "Estimation of non-nested networks", "Estimation of five networks", "Estimation of non-related networks', "Gaussian model vs GeneNetWeaver", "Network dimension", "Effect of sample size", and "Computational time". Section "Coexpression RNA and protein networks in breast cancer" includes four subsections: "String database", "Protein and RNAseq networks from GENIE3-Sep", "GO categories", and "Knockout signatures". The XLS file includes the list of gene−gene interactions for both Protein and RNAseq data derived via JRF, the list of genes contained in each of the clusters shown in Figure 6, and the list of interactions obtained via GENIE3-Sep. (ZIP)

AUTHOR INFORMATION

Corresponding Authors

*Z.T.: Tel: +1 (212) 659-8508. E-mail: zhidong.tu@mssm.edu.
*P.W.: Tel: +1 (212) 731-7052. E-mail: pei.wang@mssm.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. F.P. and P.W. are partially supported by National Institutes of Health grants (SUB-U24CA160034 and SUB-R01GM108711). Z.T. received financial support from Berg Pharma as a consultant and support from National Institutes of Health (U01AG046170,U01HG008451) and Leducq Foundation Transatlantic Networks of Excellence Program grant. P.W. is also supported by National Institutes of Health grants (R01GM082802, sub-P01CA53996). We also acknowledge the expert assistance of Steven A. Carl, Karl R. Clauser, Chenwei Lin, D. R. Mani, Philipp Mertins, Amanda Paulovich, Xianlong Wang, and Ping Yan.

REFERENCES

(1) Paulovich, A. G.; Billheimer, D.; Ham, A. J.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H.; et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **2010**, *9* (2), 242−254.

(2) Ellis, M.; Gillette, M.; Carr, S.; Paulovich, A.; Smith, R.; Rodland, K.; Townsend, R.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Liebler, D. CPTAC, Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **2013**, *3* (10), 1108−1112.

(3) The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490* (7418), 61−70.

(4) Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 17 DOI: 10.2202/1544-6115.1128.

(5) Zhu, J.; Zhang, B.; Smith, E. N.; Drees, B.; Brem, R. B.; Kruglyak, L.; Bumgarner, R. E.; Schadt, E. E. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **2008**, *40*, 854−861.

(6) Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **2000**, *7*, 601−620.

(7) Schäfer, J.; Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 32 DOI: 10.2202/1544-6115.1175.

(8) Yuan, M.; Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* **2007**, *94*, 19−35.

(9) Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432−441.

(10) Peng, J.; Wang, P.; Zhou, N.; Zhu, J. Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.* **2009**, *104*, 735−746.

(11) Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: data integration in dynamic models—a review. *BioSystems* **2009**, *96*, 86−103.

(12) Lee, W.-P.; Tzou, W.-S. Computational methods for discovering gene networks from expression data. *Briefings Bioinf.* **2009**, *10*, 408−423.

(13) Zhou, S.; Lafferty, J.; Wasserman, L. Time varying undirected graphs. *Machine Learning* **2010**, *80*, 295−319.

(14) Song, L.; Kolar, M.; Xing, E. P. KELLER: estimating time-varying interactions between genes. *Bioinformatics* **2009**, *25*, i128−i136.

(15) Danaher, P.; Wang, P.; Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2014**, *76*, 373−397.

(16) Guo, J.; Levina, E.; Michailidis, G.; Zhu, J. Joint estimation of multiple graphical models. *Biometrika* **2011**, *98*, asq060.

(17) Ahmed, A.; Xing, E. P. Recovering time-varying networks of dependencies in social and biological studies. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 11878−11883.

(18) Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **2010**, *5*, e12776.

(19) Petralia, F.; Wang, P.; Yang, J.; Tu, Z. Integrative random forest for gene regulatory network inference. *Bioinformatics* **2015**, *31*, i197−i205.

(20) Maduranga, D.; Zheng, J.; Mundra, P. A.; Rajapakse, J. C. *Pattern Recognition in Bioinformatics*; Springer, 2013; Chapter 2, pp 13−22.

(21) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.

(22) Yang, P.; Hwa Yang, Y.; Zhou, B. B.; Zomaya, A. Y. A review of ensemble methods in bioinformatics. *Curr. Bioinf.* **2010**, *5*, 296−308.

(23) Greenfield, A.; Madar, A.; Ostrer, H.; Bonneau, R. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One* **2010**, *5*, e13397−e13397.

(24) Marbach, D.; Costello, J. C.; Küffner, R.; Vega, N. M.; Prill, R. J.; Camacho, D. M.; Allison, K. R.; Kellis, M.; Collins, J. J.; Stolovitzky, G.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796−804.

(25) Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 5116−5121.

(26) Chen, H.; Sharp, B. M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinf.* **2004**, *5*, 147.

(27) Mertins, P.; et al. Proteogenomic analysis of human breast cancer connects genetic alterations to phosphorylation networks. *Eur. J. Cancer* **2014**, *50*, S10.

(28) Schaffter, T.; Marbach, D.; Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **2011**, *27*, 2263−2270.

(29) Andre, F.; Dieci, M. V.; Dubsky, P.; Sotiriou, C.; Curigliano, G.; Denkert, C.; Loi, S. Molecular pathways: involvement of immune pathways in the therapeutic response and outcome in breast cancer. *Clin. Cancer Res.* **2013**, *19*, 28−33.

(30) Hyung, S.-W.; Lee, M. Y.; Yu, J.-H.; Shin, B.; Jung, H.-J.; Park, J.-M.; Han, W.; Lee, K.-M.; Moon, H.-G.; Zhang, H.; et al. A serum protein profile predictive of the resistance to neoadjuvant chemotherapy in advanced breast cancers. *Mol. Cell. Proteomics* **2011**, *10*, M111.011023.

(31) Ambrosone, C. B.; Barlow, W. E.; Reynolds, W.; Livingston, R. B.; Yeh, I.-T.; Choi, J.-Y.; Davis, W.; Rae, J. M.; Tang, L.; Hutchins, L. R.; et al. Myeloperoxidase genotypes and enhanced efficacy of chemotherapy for early-stage breast cancer in SWOG-8897. *J. Clin. Oncol.* **2009**, *27*, 4973−4979.

(32) Thirkettle, S.; Decock, J.; Arnold, H.; Pennington, C. J.; Jaworski, D. M.; Edwards, D. R. Matrix metalloproteinase 8 (collagenase 2) induces the expression of interleukins 6 and 8 in breast cancer cells. *J. Biol. Chem.* **2013**, *288*, 16282−16294.

(33) Zhao, S.; Han, J.; Zheng, L.; Yang, Z.; Zhao, L.; Lv, Y. MicroRNA-203 Regulates Growth and Metastasis of Breast Cancer. *Cell. Physiol. Biochem.* **2015**, *37*, 35−42.

(34) Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **2006**, *1695*, 1−9.

(35) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4*, 44−57.

(36) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1−13.

(37) André, F.; Cortés, J. Rationale for targeting fibroblast growth factor receptor signaling in breast cancer. *Breast Cancer Res. Treat.* **2015**, *150*, 1−8.

(38) Sun, S.; Jiang, Y.; Zhang, G.; Song, H.; Zhang, X.; Zhang, Y.; Liang, X.; Sun, Q.; Pang, D. Increased expression of fibroblastic growth factor receptor 2 is correlated with poor prognosis in patients with breast cancer. *J. Surg. Oncol.* **2012**, *105*, 773−779.

(39) Cerliani, J. P.; Vanzulli, S. I.; Piñero, C. P.; Bottino, M. C.; Sahores, A.; Nuñez, M.; Varchetta, R.; Martins, R.; Zeitlin, E.; Hewitt, S. M.; et al. Associated expressions of FGFR-2 and FGFR-3: from mouse mammary gland physiology to human breast cancer. *Breast Cancer Res. Treat.* **2012**, *133*, 997−1008.

(40) Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **2011**, *39*, D561−D568.

(41) Parker, B. C.; Engels, M.; Annala, M.; Zhang, W. Emergence of FGFR family gene fusions as therapeutic targets in a wide spectrum of solid tumours. *Journal of pathology* **2014**, *232*, 4−15.

(42) Cine, N.; Baykal, A. T.; Sunnetci, D.; Canturk, Z.; Serhatli, M.; Savli, H. Identification of ApoA1, HPX and POTEE genes by omic analysis in breast cancer. *Oncol. Rep.* **2014**, *32*, 1078−1086.

(43) Barrett, T.; Edgar, R. [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. *Methods Enzymol.* **2006**, *411*, 352−369.

(44) Ryan, B. M.; Robles, A. I.; Harris, C. C. Genetic variation in microRNA networks: the implications for cancer research. *Nat. Rev. Cancer* **2010**, *10*, 389−402.